Data-Driven Operations in Changing Environments

by

Ruihao Zhu
B.Eng., Shanghai Jiao Tong University (2015)
B.Eng., University of Michigan (2015)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2021

Author . . . . . .
Department of Aeronautics and Astronautics
June 4, 2021

Certified by . . . . . .
David Simchi-Levi
Professor of Engineering Systems
Thesis Supervisor

Certified by . . . . . .
Eytan Modiano
Professor, Department of Aeronautics and Astronautics
Thesis Committee Chair

Certified by . . . . . .
Hamsa Bastani
Assistant Professor of Wharton School, University of Pennsylvania
Thesis Committee Member

Accepted by . . . . . .
Zoltan Spakovszky
Professor, Department of Aeronautics and Astronautics
Chair, Graduate Program Committee

Data-Driven Operations in Changing Environments

by

Ruihao Zhu

Submitted to the Department of Aeronautics and Astronautics
on June 4, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Rapid development of data science technologies have enabled data-driven algorithms for many important operational problems. Existing data-driven solutions often requires the operational environments being stationary. However, recent examples have shown that the operational environments can change dynamically. It is thus imperative to design data-driven algorithms that is capable of working in time-varying environments.

We first introduce data-driven decision-making algorithms that achieve state-of-the-art *dynamic regret* bounds for non-stationary bandit and reinforcement learning settings. These settings capture applications such as advertisement allocation, dynamic pricing, and inventory control in changing environments. Our main contribution is a general algorithmic recipe for a wide variety of non-stationary bandit and reinforcement learning problems without any knowledge about the environments in advance.

Next, 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 prior that is shared across products. We then propose a meta dynamic pricing algorithm that learns this prior online while solving a Thompson sampling pricing experiments for each product.

Finally, motivated by our collaboration with AB InBev, a consumer packaged goods (CPG) company, we consider the problem of forecasting sales under the coronavirus disease 2019 (COVID-19) pandemic. Our approach combines online learning and pandemic modeling to develop a data-driven *online non-parametric regression* method. Numerical experiments show that our method is capable of reducing the forecasting error in terms of WMAPE (*i.e.*, weighted mean absolute percentage error) and MSE (*i.e.*, mean squared error) by more than 50% for AB InBev.

Thesis Supervisor: David Simchi-Levi
Title: Professor of Engineering Systems

Thesis Supervisor: Eytan Modiano
Title: Professor, Department of Aeronautics and Astronautics

Thesis Supervisor: Hamsa Bastani
Title: Assistant Professor of Wharton School, University of Pennsylvania

Acknowledgments

This thesis provides a partial summary of my amazing and special 5-year journey (despite the unusual COVID-19 pandemic) at MIT where I have the pleasure of being a member of an encouraging and incredible community.

First and foremost, I want to thank my advisor, David Simchi-Levi, for his guidance, care, passion, and support. I have been working with David since 2018 and David has been guiding my research at both strategic and detailed levels. Through these, I have gradually learned to appreciate the beauty and joy of research. Moreover, he has also given me many invaluable suggestions beyond research. I learned from David every day and I can hardly imagine achieving anything without the help, advice, and encouragement from him. In short, David is not only my great advisor, but also a role model.

I would like to thank other members of my thesis committee: Eytan Modiano, Hamsa Bastani, Nikos Trichakis, and Negin Golrezaei. I was fortunate to have Eytan as my former advisor during my master study and I have learned a lot from him. I have enjoyed working with Hamsa and Nikos on many exciting research problems. I also love to work with Negin to teach analytics to students. All of them have given valuable advice to me in various stages.

I would also like to thank Omar Besbes, Hamsa Bastani, Negin Golrezaei, Michelle Wu, and Wang Chi Cheung for their help throughout my academic job search.

I am indebted to my home departments, AeroAstro and IDSS, for the supportive and collaborative atmosphere. I would like to thank Janet Kerrigan, Beth Marois, Kim Strampel, and other staff members for these. I would also like to thank Alexander Rakhlin and Daniela Rus for their support and advice.

I had the privilege to work with a number of fantastic co-authors, which has resulted in this current thesis. Chapters 2 and 3 are joint work with Wang Chi Cheung. Chapter 4 is a joint work with Hamsa Bastani. Finally, Chapter 5 is joint work with Rui Sun and Michelle Wu. I found our collaborations a joy.

I want to acknowledge the support from industry partners. In particular, Chapter 5 is a direct result from a collaboration with AB InBev. This would not be possible without

the help from several people at AB InBev, includingTina Gui, Ivo Montenegro and a few others. I would also like to thank my summer intern mentors at Google and Amazon, they are Branislav Kveton, Craig Boutilier, Dean Foster, Dhruv Madeka, and Abhishek Gupta.

I am lucky to have made friends with many current and former members of the MIT Data Science Lab: Michelle Wu, Wang Chi Cheung, Hanzhang Qin, Rui Sun, Jinzhi Bu, Li Wang, Peter Zhang, Hanwei Li, Yunzong Xu, Xiaoyue Gong (special thanks for sharing her desk with me for a year), Yiqun Hu, Feng Zhu, Chonghua Wang, Sabrina Zhai, Kirby Ledvina, Louis Chen, Ali Shameli, Menglong Li, Zhenzhen Yan, Jinglong Zhao, He Wang, Yehua Wei, Arzum Akkas, Will Ma, Kris Johnson, Elaheh Fata, Milashini Nambiar, Clark Pixton, Zachary Owen. I enjoyed many of the meetings and discussions with you all.

I want to thank my friends from different places throughout my PhD study: Weishun Zhong, Jun Yin, Zhi Xu, Weike Sun, Alan Malek, Chelsea Qiu, Minghao Qiu, Yuchen Wang, Jason Liang, Dylan Foster, Qingkai Liang, Igor Kadota, Wenbo Tao, Anurag Rai, Jianan Zhang, Thomas Stahlbuhk, Rajat Talak, Hyang-Won Lee, Abhishek Sinha, Bai Liu, Xinzhe Fu, Vishrant Tripathi Chin-Chia Hsu, Chulhee Yun, Yan Jin, Qi Yang, Yi Sun, Yuan Yuan, Yiou Wang, Han Su, Jiachen Lin, Renyuan Xu, Meng Qi, Zeyu Zheng, Mingxi Zhu, Wanning Chen, Wenjia Ba, Zhengli Wang, Danqi Luo, Shixin Wang, Haotian Song, Zhengyuan Zhou, and many others. Thank you all for sharing many happy moments with me, including hanging out, hunting for delicious foods, playing basketball, etc.

I also want to thank my pre-MIT friends: Yuanhao Zhai, Huajun Zhang, Huaiyu Li, Shaokun Li, Shuo Han, Xucheng Zhu, Yuan Huang, Peizhen Jiang, Ruoxing Lei, Kunming Wu, Ziyuan Lin, Xian Mo, Wenzhao Li, Chao Qin, and many more for tons of happy memories.

Last but not least, my deepest and most sincere gratitude go to my parents, my girl-friend, Kexin Zhang, and all my other family members. This journey would not be possible without the care and support from all of you. Thank you all so much for your unselfish love and everything.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in data science technologies have enabled big-data analytics for operations management. Currently, most existing works in this field critically assume the operational environments remain unchanged throughout. However, real-world operational environments are often time-varying and dynamically evolving. In this thesis, we consider three different scenarios of operations management in changing environments.

In Chapters 2 and 3, we first consider non-stationary sequential decision-making, reflecting the fact that the environment where the decision-maker operates is often dynamically changing. We develop bandit optimization and reinforcement learning algorithms for various sequential decision-making problems, and apply some of our developed methods to the context of online recommendation, dynamic pricing, and inventory control.

In Chapter 4, motivated by the fact that companies sequentially launch new products. We ask the question: when a company is making decisions, such as pricing decisions, for a new product, should it always start from scratch? Or could it leverage experience gained from past products? We develop learning algorithms that can not only learn within a single product but also learn across products to accelerate decision-making.

In Chapter 5, motivated by our collaboration with AB InBev, a large CPG company that is facing a dramatically changing demand environment due to the COVID-19 pandemic, we develop a novel online non-parametric regression method to help the company to adjust its demand forecast.

1.1 Non-Stationary Bandit Optimization

Consider the following general decision-making framework: a decision-maker (DM) interacts with a *multi-armed bandit* (MAB) system by picking actions one at a time sequentially. Upon selecting an action, she instantly receives a reward drawn randomly from a probability distribution tied to this action. The goal of the DM is to maximize her cumulative rewards. However, she faces the following challenges:

- *Uncertainty:* the reward distribution of each action is initially not known to the DM. She has to estimate the underlying reward distributions via interacting with the environment.

- *Non-Stationarity:* the reward distributions can evolve over time.

- *Partial/Bandit Feedback:* the DM can only observe the random reward of the selected action each time, while the rewards of the unchosen actions are not observed.

Many applications naturally fall into this MAB framework. For instance, assuming linear models for the reward distributions, we can cast the problems of advertisement allocation [127, 62], dynamic pricing [116, 40, 115, 27], and traffic network routing [88, 123] into the above decision-making skeleton.

- **Advertisement Allocation:** An online platform allocates advertisements (ads) to a sequence of users. For each arriving user, the platform has to deliver an ad to her, and only observes each user's response to her displayed ad. The platform has full access to the features of the ads and the users. Following [127, 62], we could assume that a user's click behavior towards an ad, or simply the click through rate (CTR) of this ad by a particular user, follows a probability distribution governed by a common, but initially unknown response function of the features. The platform's goal is to maximize the total number of clicks. However, the unknown response function can change over time. For instance, if it is around the time that Apple releases a new iPhone, one can expect that the popularity of an Apple's ad grows.

- **Dynamic Pricing:** A seller decides the (personalized) price dynamically [116, 115, 40, 27] for each of the incoming customers with the hope to maximize sales profit. Beginning with an unknown demand function, the DM only observes the purchase decision of a customer under the posted price, but not any other price. In addition, the customers' reaction towards the same price can vary across time due to the product reviews, the emergence of competitive products, etc.

- **Traffic Network Routing:** A navigation service provider has to iteratively offer route planning services to drivers from an origin to a destination through a traffic network with initially unknown random delay on each road. For each driver, the provider could only see the delays of the roads traversed by this driver, but not the other roads'. Moreover, the delay distributions could change over time as the roads are also shared by other traffics (*i.e.*, those not using this navigation service). The provider wants to minimize the cumulative delays throughout the course of vehicle routing.

Evidently, the DM faces a trilemma among exploration, exploitation as well as adaptation to changes. On one hand, the DM wishes to exploit, and to select the action with the best historical performances to earn as much reward as possible. On the other hand, she wants to explore other actions to get a more accurate estimation of the reward distributions. The changing environment makes the exploration-exploitation trade-off even more delicate. Indeed, past observations could become obsolete due to the changes in the environment, and the DM needs to explore for changes and refrain from exploiting possibly outdated observations.

We focus on resolving this trilemma in various MAB problems. Traditionally, most MAB problems are studied in the stochastic [21] and adversarial [18] environments. In the former, the uncertain model is static, and each feedback is corrupted by a mean zero random noise. The DM aims at estimating the latent static environment using historical data and converging to the optimum, which is achieved by a static strategy that selects a single action throughout. In the latter, the model is not only uncertain, but also dynamically changed by an adversary. While the DM strives to hedge against the changes, it is generally

impossible to achieve the optimum. Hence, existing research also focuses on competing favorably in comparison to a static strategy.

Unfortunately, strategies for the stochastic environments can quickly deteriorate under non-stationarity as historical data might “expire"; while the permission of a confronting adversary in the adversarial settings could be too pessimistic. Starting from [36, 37], a stream of research works (see Section 1.1.1) focuses on MAB problems in a *drifting* environment, which is a hybrid of a stochastic and an adversarial environment. Although the environment can be dynamically and adversarially changed, the total changes (quantified by a suitable metric) in a T-round problem is upper bounded by B_T $(= \Theta(T^\rho)$ for some $\rho \in (0,1))$, the *variation budget* [36, 37], and the feedback is corrupted by a mean zero random noise. The aim is to minimize the *dynamic regret* [36], which is the optimality gap compared to the sequence of (possibly dynamically changing) optimal decisions, by simultaneously estimating the current environment and hedging against future changes every round. The framework of [36, 37] enable us to compete against the so-called *dynamic comparator*. Most of the existing works for non-stationary bandits have focused on the the relatively ideal case in which B_T is known. In practice, however, B_T is often not available ahead as it is a quantity that requires knowledge of future information. Though some efforts have been made towards this direction [113, 134], how to design algorithms with low dynamic regret when B_T is unknown remains largely as a challenging problem.

1.1.1 Related Works

Stationary and Adversarial Bandits MAB problems with stochastic and adversarial environments are extensively studied, as surveyed in [46, 125]. To model inter-dependence relationships among different arms, models for linear bandits in stochastic environments have been studied. In [19, 68, 160, 62, 3], UCB type algorithms for stochastic linear bandits were studied, and the authors of [3] provided the tightest regret analysis for algorithms of this kind. The authors of [162, 9, 4] also proposed Thompson sampling algorithms for this setting to bypass the high computational complexity of the UCB type algorithms..

Bandits in Drifting Environments Departing from purely stochastic or adversarial settings, Besbes et al. [36, 37] laid down the foundation of bandit in drifting environments, and considered the K-armed bandit setting. They achieved the tight dynamic regret bound $\tilde{O}((KB_T)^{1/3}T^{2/3})$ by restarting the EXP3 algorithm [18] periodically when B_T is known. Wei et al. [174] provided refined regret bounds based on empirical variance estimation, assuming the knowledge of B_T. Wei and Srivastava [177] analyzed the sliding window upper confidence bound algorithm for the K-armed MAB with known B_T setting. Subsequently, Karnin and Anava [113] considered the setting without knowing B_T and $K = 2$, and achieved a dynamic regret bound of $\widetilde{O}(B_T^{9/50}T^{41/50} + T^{77/100})$ with a change point detection type technique. In a recent work, Luo et al. [134] generalized this change point detection type technique to the K-armed contextual bandits in drifting environments, and in particular demonstrated an improved bound $\widetilde{O}(KB_T^{1/5}T^{4/5})$ for the K-armed bandit problem in drifting environments when B_T is not known. Keskin and Zeevi [115] considered a dynamic pricing problem in a drifting environment with 2-dimensional linear demands. Assuming a known variation budget B_T, they proved an $\Omega(B_T^{1/3}T^{2/3})$ dynamic regret lower bound and proposed a matching algorithm by properly discounting historical observations (this includes sliding-window estimation as a special case). When B_T is not known, their algorithm achieves $\tilde{O}(B_T T^{2/3})$ dynamic regret bound. There also exist some heuristic approaches for this (or similar) setting [96, 150]. Finally, various online problems with full information feedback and drifting environments were studied in the literature [61, 37, 107].

	Known B_T	Unknown B_T
[37]	$\widetilde{O}\left(B_T^{1/3}T^{2/3}\right)$	$\widetilde{O}\left(B_T T^{2/3}\right)$
[113]	$\widetilde{O}\left(B_T^{9/50}T^{41/50} + T^{77/100}\right)$	$\widetilde{O}\left(B_T^{9/50}T^{41/50} + T^{77/100}\right)$
[134]	$\widetilde{O}\left(B_T^{1/3}T^{2/3}\right)$	$\widetilde{O}\left(B_T^{1/5}T^{4/5}\right)$
The current thesis	$\widetilde{O}\left(B_T^{1/3}T^{2/3}\right)$	$\widetilde{O}\left(B_T^{1/3}T^{2/3} + T^{3/4}\right)$

Table 1.1: Comparisons between our results and prior works. Here, the dynamic regret bounds only show dependence on B_T and T. $\widetilde{O}(\cdot)$ denotes the function growth, and omits the logarithmic factors.

Bandits in Piecewise Stationary/Switching Environments Apart from drifting environments, numerous research works consider the *piecewise stationary/switching environment*, where the time horizon is partitioned into at most S intervals, and the optimal action(s) can switch from one to another across different intervals. The partition is not known to the DM. Algorithms are designed for various bandit settings, assuming a known S [18, 91, 13, 133, 134, 50], or without knowing S [113, 134]. Notably, the Sliding Window-UCB and the "forgetting principle" was first proposed by Garivier and Moulines [91], while it is only analyzed under K-armed switching environments. But we also have to emphasize that the S is a looser measure of non-stationarity in the sense that every tiny change in the environment could be counted towards the total number of switches. In other words, even if there are a total of T switches, the total variation budget B_T could still be far less than T. Hence, the drifting environment serves as a better proxy for non-stationarity.

Further Contrasts to Existing Works The main idea underpinning our Bandit-over-Bandit framework is to use a learning algorithm to tune the underlying learning algorithm's parameters. While this shares similar spirit to several existing works, such as the heuristic meta bandit [101], the heuristic envelop policy [38], as well as algorithms for bandit corralling (see [7, 134] and references therein), our design is different in the sense that rather than simultaneously maintaining multiple copies of the `SW-UCB` algorithm (similar to [101, 7, 134, 38]), we treat the problem of selecting window size for the `SW-UCB` algorithm as another independent adversarial bandit learning instance. To achieve this, we divide the time horizon into epochs, and force the `SW-UCB` algorithm to restart at the beginning of each epoch. This critical difference allow us to establish (nearly) optimal dynamic regret bound of the `BOB` algorithm while prior works cannot.

Follow-Up Works The results presented in [134] were further improved to the optimal $\widetilde{O}(K^{1/3}B_T^{1/3}T^{2/3})$ dynamic regret bound in [57], but it is unclear how to generalize the techniques in [57] beyond the K-armed bandit setting. In [41, 23], the authors presented optimal learning algorithms for the switching setting without knowing the number of switches. The design of parameter-free online learning algorithms are also considered in other online

learning scenarios, such as bandit convex optimization [188], bandit non-convex optimization [159], and reinforcement learning [60].

1.1.2 Summary of Main Contributions for Chapter 2

We design and analyze a novel algorithmic framework for bandit problems in drifting environments. We begin by demonstrating our results via the lens of the linear model class. However, we emphasize the choice of linear model is by no mean a restriction, and indeed, we demonstrate the generality of our framework to a variety of bandit learning models. Our main contributions can be summarized as follows.

- When the variation budget B_T is known, we characterize the lower bound of dynamic regret, and develop a tuned Sliding Window Upper-Confidence-Bound (`SW-UCB`) algorithm with matching dynamic regret upper bound up to logarithmic factors.
- When B_T is unknown, we propose a novel Bandit-over-Bandit (`BOB`) framework that tunes the window size of the `SW-UCB` algorithm adaptively. When the amount of non-stationarity is moderate to large, the `BOB` algorithm recovers the optimal dynamic regret bound; otherwise, it obtains a dynamic regret bound with best dependence on T compared to prior literature.
- Our algorithm design and analysis shed light on the fine balance between exploration, exploitation and adaptation to changes in dynamic learning environments. We rigorously incorporate the "forgetting principle" [91] into the Optimism-in-Face-of-Uncertainty principle [21, 3], by demonstrating that the DM can enjoy an optimal dynamic regret bound if she keeps disposing of sufficiently old observations. We provide the precise rate of disposal, and rigorously show its convergence to optimality.

1.2 Non-Stationary Reinforcement Learning

Note that in bandit optimization, we assume that the environment's change is unrelated to the DM's action. However, this assumption can be violated in many important operational

problems.

To remove this assumption, we consider a general sequential decision-making framework, where a DM interacts with an initially unknown environment iteratively. At each time step, the DM first observes the current state of the environment, and then chooses an available action. After that, she receives an instantaneous random reward, and the environment transitions to the next state. The reward follows a reward distribution, and the subsequent state follows a state transition distribution. Both distributions depend (solely) on the current state and action. Hence, the environment can be fully characterized by a discrete time MDP. The DM aims to design a policy that maximizes her cumulative rewards, while facing the following challenges:

- **Endogenous Dynamics:** The rewards and state transitions (and hence, future rewards) are influenced by the DM's policy.
- **Non-Stationarity:** The reward and state transition distributions vary (independently of the DM's policy) across time steps.
- **Uncertainty:** Both the reward and state transition distributions are initially unknown to the DM.
- **Bandit/Partial Feedback:** The DM can only observe the reward and state transition resulted by the current state and the action she picks in each time step.

It turns out that many applications can be captured by this framework:

Example 1 (**Inventory Control**). *In inventory control with lost-sales, zero-lead time, and possibly fixed cost [106, 34, 186, 56], a seller repeatedly observes her current stock level (i.e., state) and decides the quantity to order (i.e., action). The ordered quantity then arrives instantaneously. Afterwards, a demand sampled from an initially unknown demand distribution is realized and the seller observes the censored demand. Finally, the seller pays the ordering cost, fixed cost, and lost-sales/holding cost. The goal of the seller is to minimize the cumulative cost. Here, due to the emergence of competing products or other supply chain disruption events, the demand distributions can be time varying.*

Example 2 (**Real-Time Bidding in Ads Auctions**). *Advertisers repeatedly competes for ad display impressions via real-time online auctions [93, 49, 83, 25, 94, 98]. Each advertiser begins with a budget. Upon the arrival of a user, an impression is generated, and the advertisers submit bids (i.e., action) for it subject to her remaining budget. The winning advertiser acquires the impression to display her ad to the user, and observes the user click or no-click behavior (i.e., reward). For each slot won, the advertiser has to make the payment (determined by the auction mechanism) using her remaining budget, and the budget is periodically refilled (i.e., state transition). Each advertiser wants to maximize the number of clicks on her advertisement subject to her own (continuously evolving) budget constraint. Nevertheless, the competitiveness of each auction exhibits non-stationarity as the participating advertisers and the arriving users are different from time to time. Moreover, the popularity of an ad can change due to endogenous reasons. For instance, displaying the same ad too frequently in a short period of time might reduce its freshness, and results in a tentatively low number of clicks (i.e., we can use both the remaining budget and the number of times that the ad is shown within a given window size to model the state of the MDP).*

Motivated by these applications, we consider RL in non-stationary MDPs where both the reward and state transition distributions can change over time, but the total changes (quantified by suitable metrics) are upper bounded by the respective variation budgets [36, 38]. Designing algorithms for RL in non-stationary MDPs can be extremely challenging. This is because, under non-stationarity, historical data samples might incorrectly indicate that state transition rarely happens.

Example 3 (**Perils of Drift**). *Under non-stationarity, the DM can be interacting with different MDPs over time. Consider two different 2-state-2-action MDPs, p^1 and p^2 (as shown in Fig 1-1): under p^1, the blue action always transitions to state 2 and the green action always transitions to state 1; while p^2 is exactly to the opposite. Then it is possible that whenever the DM chooses the green action in state 1 (or 2), the underlying MDP is p^1 (p^2, respectively); when she chooses the blue action in state 1 (or 2), the underlying MDP is p^2 (p^1, respectively). Moreover, due to the bandit feedback, she cannot observe the entire p^1 and p^2. Hence, the collected data would draw the wrong conclusion that neither action*

can result in state transition! We formalize this in Section 3.3.3.

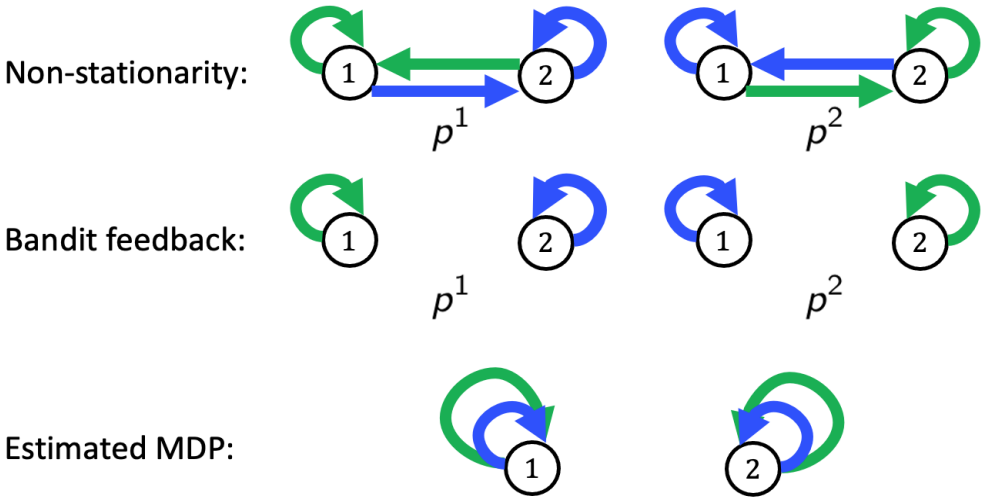


Figure 1-1: Under non-stationary, historical data can falsely indicate that state transition rarely happens.

This challenge is uniquely associated with RL in non-stationary MDPs and does not show up (or can be easily bypassed) in other related but different non-stationary sequential decision-making settings. For example, in non-stationary multi-armed bandits (MAB) [92, 36, 58], there is only one state and the DM does not need to worry about transitions between states; while in RL in piecewise stationary MDPs [108], the DM can leverage the fact that the underlying MDP only changes once in a while to bypass this difficulty.

1.2.1 Summary of Main Contributions for Chapter 3

We set forth to address the above challenges. Assuming that, during the T time steps, the total variations of the reward and state transition distributions are bounded (under suitable metrics) by the variation budgets B_r (> 0) and B_p (> 0), respectively, we design and analyze novel algorithms for RL in non-stationary MDPs. Let $D_{\max}$, S, and A be respectively the (unknown a priori) maximum diameter (measures the hardness of state transition, formally defined in Section 3.1), number of states, and number of actions in the MDP. Our main contributions are:

- We formally delineate the challenge of RL in non-stationary MDPs: existing algorithms for non-stationary sequential decision-makings [108, 92] typically follows the

Optimism in the Face of Uncertainty (OFU) + forgetting principles as follows:

1. estimate unknown parameters with most recently observed data, and construct the *tightest* possible confidence regions accordingly;
2. optimistically search for the most favorable model within the confidence regions and computes the optimistic policy, which is the optimal policy w.r.t. this most favorable model;
3. execute this optimistic policy.

In RL, step 2 is achieved by the Extended Value Iteration (EVI) [108] and the loss of using this optimistic policy is proportional to the least diameter induced by any MDP in the confidence regions. However, we demonstrate in Section 3.3.3 that in RL in non-stationary MDPs, it is possible that the diameter induced by any MDP in the confidence regions constructed according to step 1 can grow wildly, and may thus result in unfavorable dynamic regret bound.

- We develop the Sliding Window UCRL2 with Confidence Widening (`SWUCRL2-CW`) algorithm. When the variation budgets are known, we prove it attains a

$$\tilde{O}\left(D_{\max}(B_r+B_p)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$$

dynamic regret bound. In a nutshell, the confidence widening technique injects extra optimism in the learning algorithm and ensures that either the diameter induced by some MDP in the confidence region is bounded by $D_{\max}$ or a non-negligible amount of variation budget is consumed.

- We propose the Bandit-over-Reinforcement Learning (`BORL`) algorithm that tunes the `SWUCRL2-CW` algorithm adaptively, and retains the same $\tilde{O}\left(D_{\max}(B_r+B_p)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$ dynamic regret bound without knowing the variation budgets.

- As a complement, we show that if for any pair of initial state and target state, there always exists an action such that the probability of transiting from the initial state to the target state by taking this action is lower bounded uniformly over the entire time

horizon, the DM can attain low dynamic regret without widening the confidence regions. We demonstrate that in the context of single item inventory control with fixed cost [183], a mild condition on the demand distribution is sufficient for this extra assumption to hold.

1.3 Meta Dynamic Pricing

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.*, [143]). 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 [168, 164]. 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 [120, 39]. Such algorithms have been shown to be especially useful for products that exhibit relatively short life cycles [78], stringent inventory constraints [179], strong competitive effects [82], or the ability to offer personalized coupons/pricing [185, 26]. 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 [148, 63, 109]. 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 [168, 10, 162, 77].

The decision-maker typically runs a separate pricing experiment (*i.e.*, dynamic pricing algorithm) for each product (or for a set of simultaneously-offered products). However, this approach can waste valuable samples re-discovering information that could have been learned from previously-offered related 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. When offering *multiple* products simultaneously, one must additionally learn cross-product price elasticities in the demand function (to model substitution effects), which may also exhibit patterns that can be learned from substitution patterns of related products in historical data. For example, substitution effects may be stronger between more similar products, or among more price-sensitive customers like students.

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 (or per set of products, if offering multiple products simultaneously), thereby improving profits.

In Chapter 4, we propose an approach to learn 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 [55, 105, 132, 163]. 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 [10], 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 [162]. 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. To this end, we introduce a novel "prior alignment" technique to analyze the regret of Thompson sampling with a mis-specified prior, which may be of independent interest.

Using our alignment technique, we show surprisingly that *despite* prior mis-specification, greedy updating of the prior is sufficient to learn effectively across pricing experiments when the prior covariance is known. However, when the prior has an unknown covariance matrix, it is beneficial to widen the estimated prior covariance 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).

1.3.1 Related Works

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 [124, 20, 69, 161, 35, 112, 42]. This literature primarily proposes learning algorithms for a single experiment, while our focus is on meta-learning across experiments. Meta-learning can take the form of constructing an empirical Bayesian prior [149, 14], or leveraging low-dimensional structure between problems [30]. We take an empirical Bayesian approach to sequential decision-making. While there has been some prior work on meta-learning in bandits [102, 136, 173, 165] and more generally in reinforcement learning [80, 81, 180], these papers only provide heuristics for learning exploration strategies given a fixed set of past problem instances. They do not prove any theoretical guarantees on the performance or regret of the meta-learning algorithm. To the best of our knowledge, our work 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 [78, 179], among online retailers that constantly monitor competitor prices and adjust their own prices in response [82], or when prices can be personalized based on customer-specific price elasticities, *e.g.*, through personalized coupons [185]. Several papers have designed near-optimal dynamic pricing algorithms for pricing a product by balancing the resulting exploration-exploitation trade-off [120, 39, 15, 74, 99, 44, 71, 117]. Recently, this literature has shifted focus to pricing

policies that dynamically optimize the offered price with respect to exogenous features [148, 63, 109] and customer-specific price elasticity [26]. We adopt the linear demand model proposed by [26], which allows for feature-dependent heterogeneous price elasticities.

When sellers offer multiple products simultaneously, one may wish to perform price experiments *jointly* on a set of products to capture substitution effects or overlapping inventory constraints [117, 8, 77]. However, in these papers, price experimentation is still performed independently on the current set of products, and any learned parameter knowledge is not shared across future sets of products to inform future demand learning. In contrast, we propose a meta dynamic pricing algorithm that learns the distribution of unknown parameters of the demand function across products. While we focus largely on the single-product setting for ease of exposition, we show how our algorithm and theoretical results carry over straightforwardly for multi-product settings with substitution effects; in fact, transfer learning from historical data may be even more valuable in these settings since the number of parameters (*e.g.*, cross-product elasticities) to learn is much larger.

Our learning strategy is based on Thompson sampling, which is widely considered to be state-of-the-art for balancing the exploration-exploitation tradeoff [168]. Several papers have studied the sensitivity of Thompson sampling to prior misspecification. For example, [105] 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. [132] 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 [10] and [47] 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 [55]. For instance, [31] empirically find through simulations that the conservative

prior-independent Thompson sampling is significantly outperformed by vanilla Thompson sampling *even* when the prior is misspecified.[1] 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 [163]. 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 when the prior covariance is unknown, 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.

1.3.2 Summary of Main Contributions for Chapter 4

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 balance 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*). `Meta-DP++` additionally accounts for uncertainty in the estimated prior by conservatively widening the prior as a function of its estimation error.

3. *Theory:* Unlike standard approaches, our algorithm can leverage shared structure across products to achieve regret that scales sublinearly in the number of products

[1] We provide some theoretical support for this finding, since we show that limited prior mis-specification does not affect the rate of convergence (*e.g.*, when the prior covariance is known but the mean is unknown).

N. We prove upper bounds $\tilde{O}(d^2\sqrt{NT}+d^3\sqrt{T})$ and $\tilde{O}(\min\{d^2NT^{\frac{1}{2}},\ d^4N^{\frac{1}{2}}T^{\frac{3}{2}}\}) = \tilde{O}(d^3(NT)^{\frac{5}{6}})$ on the meta regret of `Meta-DP` and `Meta-DP++` respectively. In both cases, our meta-learning approach matches the performance of prior-independent algorithms for small N, and outperforms them in experiment-rich experiments (*i.e.*, when $N=\tilde{\Omega}(d)$ and $N=\tilde{\Omega}(d^4T^2)$ respectively). A key ingredient of our analysis is a "prior alignment" proof technique that may be of general interest for analyzing the regret of mis-specified Thompson Sampling instances.

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).

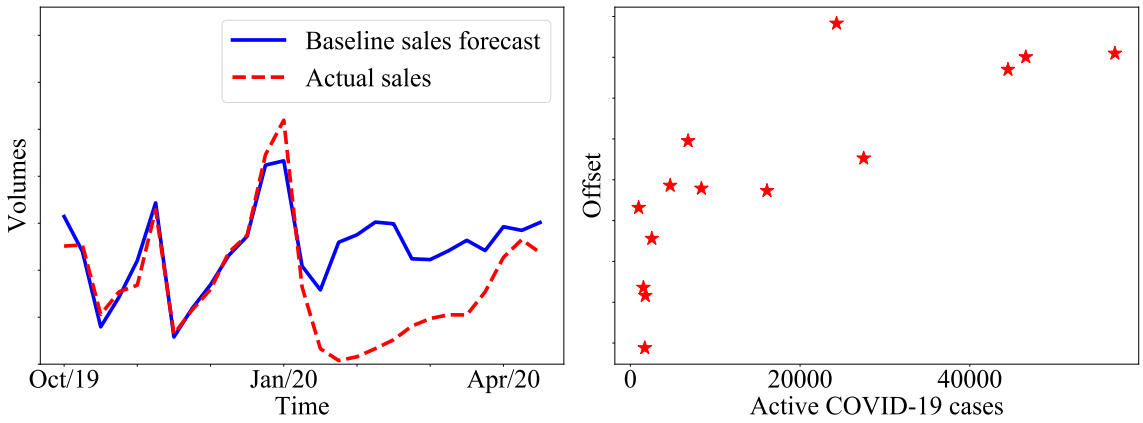
1.4 Calibrating Sales Forecast in a Pandemic

This work is motivated by our collaboration with Anheuser-Busch InBev (AB InBev), a multi-national drink and brewing company in the consumer packaged goods (CPG) sector. In 2019, AB InBev's annual sales were 52.3 billion USD as stated in its annual report [1]. According to [45], the company was expected to have a 28% market share of global volume beer sales in 2017, which makes it the largest beer company worldwide.

To improve operational efficiency, AB InBev maintains a baseline sales forecast for its products in each geographical region of interest. The baseline sales forecast is important for many operational decisions such as inventory management, promotion campaigns, and financial planning. The baseline sales forecast is trained by an offline statistical learning algorithm with historical sales data as well as social & economic data, and it can accommodate different update frequencies (*e.g.*, weekly or monthly) to make predictions for different business applications.

As an example, Fig. 1-2(a) displays the weekly baseline forecast sales volumes and the actual sales volumes in a certain geographical region before and after the beginning of the COVID-19 pandemic [178]. Evidently, the baseline sales forecast enjoys a high prediction accuracy and is capable of adapting to the seasonality patterns of the actual sales volumes during normal times. However, since the emergence of the COVID-19 pandemic,

the accuracy of the forecast plummeted drastically. This has thus placed a hurdle for AB InBev since it relies heavily on the baseline sales forecast to make operational decisions. Hence, it is of great importance for the company to incorporate the impact of the pandemic into its sales forecast.



(a) Weekly forecast demand vs. weekly real demand (b) Offsets vs. active COVID-19 case numbers

Figure 1-2

Given the incredible performance of the baseline sales forecast in normal times, we decided to follow an *add-on* procedure that iterates between the following two steps over time:

1. Predict the *offset*, *i.e.*, the difference between baseline sales forecast and actual sales volumes, caused by the COVID-19 pandemic.

2. Generate a *calibrated sales forecast* by subtracting the predicted offset from the baseline sales forecast, *i.e.*, Calibrated sales forecast $=$ Baseline sales forecast $-$ Predicted offset.

Since the baseline sales forecast is available through AB InBev's offline learning algorithm, we focus on forecasting the offset, *i.e.*, step 1. As suggested by [142], during this pandemic, customer behavior is driven almost solely by the dynamics of the pandemic. Therefore, for each region, we propose to use its number of active COVID-19 cases, *i.e.*, a proxy of the dynamics of the COVID-19 pandemic, to predict the offset. We thus visualize the relationship between the weekly averaged active COVID-19 case numbers and the weekly

averaged offsets during the pandemic in Fig. 1-2(b). From the plot, we observe that a higher number of active COVID-19 cases is more likely to result in larger offset. Note that the future active COVID-19 case numbers can be simulated by existing pandemic modeling techniques, such as the Susceptible-Infectious-Removed (SIR) epidemic model [114]. These observations immediately suggest a *greedy least squares* (GLS) policy that sequentially fits (*e.g.*, via least squares method) the best *isotonic* (non-decreasing) function [24, 156, 184] over all historical (active COVID-19 cases, offset) observations, and predict the future offsets by applying the outputs of the SIR epidemic model to the fitted isotonic function.

Unfortunately, the GLS policy suffers from some critical limitations. The GLS policy implicitly assumes that there exists an isotonic function f^* such that the statistical relationship between the active COVID-19 cases and the offset is governed by f^*, *i.e.*,

$$\text{offset} = f^*(\text{active COVID-19 cases}) + \textbf{independent zero-mean noise term}. \tag{1.1}$$

This is a commonly made assumption in existing non-parametric regression literature. Nonetheless, this is a strong assumption which makes the GLS policy susceptible. For example, the noise terms can be correlated over time (*i.e.*, temporal correlations) and could also exhibit correlations across different geographical locations (*i.e.*, spatial correlations):

- **Temporal Correlations:** suppose we use the noise terms to model the customer's (possibly random) vigilance about COVID-19 pandemic. Then, it is possible that when the first COVID-19 case occurs or the rapidly increasing active COVID-19 cases would lead to temporarily higher vigilance; while a decreasing active cases and the deployment of vaccination would lead to temporarily lower vigilance. One can also expect that as time elapses, customers might gradually adapt to the COVID-19 pandemic with better understanding of the pandemic.

- **Spatial Correlations:** since the changing sales volumes in many regions are caused by the same key drivers (*e.g.*, the COVID-19 pandemic, governments' mitigation policy), the customers' behavior can be very closed to each other in different geographical regions. That means, the relationship between active COVID-19 cases and

offset might share similar patterns across different geographical regions.

We show in Proposition 33 of Section 5.1.4 that if the noise terms are allowed to be non-independent (*i.e.*, there exist temporal correlations), the performance of the GLS policy can be ungrounded even if

1. The noise terms have almost zero-mean;
2. $f^*(\cdot)$ can fit the (active COVID-19 cases, offset) observations as well as if the observations are generated according to Eq. (1.1).

Even worse, the GLS policy might continue to perform poorly if the noise terms of all the geographical regions are similar to the ones in Proposition 33 due to spatial correlations.

In fact, this is a common caveat with classical statistical theory. In the classical statistical theory of prediction, data is assumed to be a realization of a stationary stochastic process and the effectiveness of a forecasting policy is provided in an expectation sense, which means the forecasting policy can perform poorly w.r.t. certain realization of the data as long as its performance on other realizations could compensate this deficit. However, in our problem, due to different kinds of correlations, we might repeatedly encounter the unfavorable data sequence and the performance guarantee of classical statistical theory based forecasting policy no longer holds. With these, we ask the following question:

Can we design sales forecast policy that does not rely on any statistical assumptions?

1.4.1 Modeling Approach: Combining Non-Parametric Regression, Game Theory, and Pandemic Modeling

To address the above question, we consider the *competitive online non-parametric regression* setup: in each time step, a decision-maker (DM) predicts the label (*i.e.*, offset in sales volumes) of a covariate (*i.e.*, current active COVID-19 case numbers) given past (covariate, label) observations. Each covariate is generated by a simulatable (*e.g.*, via the SIR epidemic model) generative process. We are looking for a computationally-efficient algorithm that minimizes *regret*, which is the difference between the squared ℓ_2-norm associated with

labels generated by the algorithm and labels generated by an adversary and the squared ℓ_2-norm associated with labels generated by the best isotonic (non-decreasing) function in hindsight, *i.e.*, *oracle*, and the adversarial labels. In this setup, the adversary seeks to maximize the regret.

Our notion of regret follows [152] and the connections between learning and repeated games [53]. Specifically, we make no statistical assumption (*e.g.*, eq. (1.1)) on the label's generative process, but in order to maximize the regret, the adversary would try to maximize our loss while minizing the loss of the best isotonic function/oracle. That is, we encode our prior belief that the family of isotonic functions is expected to perform well into the definition of oracle without enforcing eq. (1.1). This ensures our model and the forthcoming solution are going to provide performance guarantee for all possible (covariate, label) observations.

1.4.2 Related Works

[152, 89] have studied the online non-parametric regression problem with a general benchmark. However, their algorithms are often computationally inefficient for our problem (see Section 1.1 of [121]).

When the benchmark is an isotonic function and all the covariates are known beforehand, *i.e.*, the fixed-design setting, [121] first shown that many existing online learning algorithms, *e.g.*, online gradient descent [193], exponential gradient descent [119], and follow the leader [103], could only ensure sub-optimal regret upper bounds, and they further developed a computationally-efficient exponential weights algorithm [131] to attain the minimax-optimal regret bound $\Theta(T^{1/3})$. [121] has also demonstrated that the DM has to suffer $\Theta(T)$ regret if the covariates are chosen by an adversary without revealing them at the beginning. The fixed-design setting corresponds to knowing exactly the active COVID-19 case numbers over the entire time horizon in AB InBev's case, and it is thus impractical. Later on, [122] studied the case where the (unknown *a priori*) covariates arrives in random permutation order, and they shown that the regret bound of this problem is also $\tilde{O}(T^{1/3})$. Unfortunately, this covariate generative process is still too restrictive for our application.

Isotonic functions have found numerous applications in statistics (see [156] and refer-

ences therein). More recently, [100, 27, 70] have used (known *a priori*) isotonic functions to model non-linearity in customer behaviors, *i.e.*, purchasing probability, in the context of dynamic pricing. Isotonic functions have also been used to model the reward distributions in bandit optimization [79, 58, 191, 190].

Big-data anlalytics and machine learning techniques have recently emerged as a popular tool to perform sales prediction and related problems [76, 52]. For example, [66] investigated the benefits of information sharing in improving sales forecast accuracy. [28] proposed to solve the sales forecast and order optimization problem with singlestep machine-learning algorithms. [67] conducted empirical studies on how inventory availability information can affect sales. [129] also considered the joint sales prediction, product shipping and allocation problem using data-driven approaches. Our work contributes to this line of works by demonstrating how one can combine online learning and pandemic modeling to perform sales forecast calibration.

1.4.3 Summary of Main Contributions for Chapter 5

We develop a data-driven online non-parametric regression method that sequentially takes the (past and simulated future) active COVID-19 case numbers as input, and then outputs the level of calibration of AB InBev's baseline sales forecast. Specifically, for a T time steps online non-parametric regression problem, our contributions can be summarized as follows:

- **Rate-optimal learning policy:** We develop a computationally-efficient Simulating Exponential Weights (`SEW`) policy that applies the simulated future covariates as an extra (in addition to past covariate, label observations) input to the exponential weights algorithm [131]. Different than conventional online learning algorithms [53, 46], which make predictions based only on historical observations, the `SEW` policy additionally makes use of the simulated future covariates, and this makes its regret analysis challenging. We bypass this difficulty by exploiting the generative process of the covariates and the relaxation framework (see Section D.2 of the appendix for more details) from sequential complexity theory of online learning [152], and show

that the SEW policy achieves a regret bound of order $\tilde{O}(T^{1/3})$. Comparing this regret upper bound with the regret lower bound established in Theorem 5 of [121], we know that it is minimax-optimal.

- **Numerical experiments:** We evaluate the SEW policy on both synthetic and AB InBev's datasets. With the synthetic datasets, we compare the performance of the SEW policy against the exponential weights (EW) algorithm for fixed-design online isotonic regression [121] and the online linear regression (OLR) algorithm [170]. The results show that by using the SEW policy, the cost of not knowing the covariates in advance is small, and the prediction accuracy is significantly higher than that of the OLR algorithm in various cases. With the AB InBev's dataset, we compare the performance of the SEW policy calibrated sales forecast with two benchmark methods—the OLR calibrated sales forecast and the baseline sales forecast. We study both the weekly and the monthly update scenarios. The results demonstrate that our method outperforms the competing methods by more than 20% in terms of WMAPE (*i.e.*, weighted mean absolute percentage error) and MSE (*i.e.*, mean squared error) by more than 50% in the monthly forecast (AB InBev's main focus) and 15% in the weekly forecast.

Chapter 2

Non-Stationary Bandit Optimization

2.1 Problem Formulation for Drifting Linear Bandits

We start by introducing the notations to be used and the model formulation. From the current section to Section 2.5, we focus on the drifting linear bandit problem, which serves to illustrate our algorithmic framework. After that, we provide generalizations to a wide variety of bandit problems in drifting environments in Section 2.6.

2.1.1 Notation

Throughout this chapter, all vectors are column vectors, unless specified otherwise. We define $[n]$ to be the set $\{1,2,\ldots,n\}$ for any positive integer n. We denote $\langle x,y\rangle = x^\top y$ as the inner product between $x,y \in \mathbb{R}^d$. For $p \in [1,\infty]$, we use $\|\boldsymbol{x}\|_p$ to denote the p-norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$. For a positive definite matrix $A \in \mathbb{R}^{d\times d}$, we use $\|\boldsymbol{x}\|_A$ to denote $\sqrt{\boldsymbol{x}^\top A\boldsymbol{x}}$ of a vector $\boldsymbol{x} \in \mathbb{R}^d$. We denote $x \vee y$ and $x \wedge y$ as the maximum and minimum between $x,y \in \mathbb{R}$, respectively. We adopt the asymptotic notations $O(\cdot), \Omega(\cdot)$, and $\Theta(\cdot)$ [65]. When logarithmic factors are omitted, we use $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$, $\widetilde{\Theta}(\cdot)$, respectively. With some abuse, these notations are used when we try to avoid the clutter of writing out constants explicitly.

2.1.2 Learning Protocol

In each round $t \in [T]$, a decision set $D_t \subseteq \mathbb{R}^d$ is presented to the DM. Then, the DM chooses an action $X_t \in D_t$. Afterwards, the reward

$$Y_t = \langle X_t, \theta_t \rangle + \eta_t$$

is revealed to the DM as a whole. We allow D_t to be chosen by an *oblivious adversary*, who chooses the decision sets $\{D_t\}_{t=1}^T$ before the protocol starts [54]. The parameter vector $\theta_t \in \mathbb{R}^d$ is an unknown d-dimensional vector, and η_t is a random noise drawn i.i.d. from an unknown sub-Gaussian distribution [154] with variance proxy R. By definition, this means $\mathbf{E}[\eta_t] = 0$, and $\forall \lambda \in \mathbb{R}$ we have $\mathbf{E}[\exp(\lambda \eta_t)] \leq \exp(\lambda^2 R^2/2)$. Following the convention of the existing linear bandit literature [3, 9], we assume there are positive constants L and S, such that $\|X\|_2 \leq L$ for all $X \in D_t$ and all $t \in [T]$, and $\|\theta_t\|_2 \leq S$ holds for all $t \in [T]$. In addition, the instance is normalized so that $|\langle X, \theta_t \rangle| \leq 1$ for all $X \in D_t$ and $t \in [T]$. The constants L, S are known to the DM.

We consider the *drifting environment* [36], where θ_t can change over different t, with the constraint that the sum of the Euclidean distances between consecutive θ_t's is bounded from above by the variation budget $B_T = \Theta(T^\rho)$ for some $\rho \in (0,1)$, *i.e.*,

$$\sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|_2 \leq B_T. \tag{2.1}$$

We allow θ_t's to be chosen by an oblivious adversary. It is worth pointing out that the concepts of a drift environment and variation budget were originally introduced in [37] and [36, 38] for the full information setting and the partial/bandit feedback setting, respectively.

We define $\mathscr{H}_t = \{D_s, X_s, Y_s\}_{s=1}^{t-1} \cup \{D_t\}$ as the available history information at round $t \in [T]$. The DM's goal is to design a non-anticipatory policy π, which only uses the information $\mathscr{H}_t$ in each round t, to maximize the cumulative reward. Equivalently, the goal is to minimize the *dynamic regret*, which is the worst case cumulative regret against the optimal policy π^*, that has full knowledge of θ_t's. Denoting $x_t^* = \arg\max_{x \in D_t} \langle x, \theta_t \rangle$, the

dynamic regret of a non-anticipatory policy π is mathematically expressed as

$$\mathscr{R}_T(\pi) = \mathbf{E}\left[\text{Regret}_T(\pi)\right] = \mathbf{E}\left[\sum_{t=1}^{T} \langle x_t^* - X_t, \theta_t \rangle\right],$$

where the expectation is taken with respect to the randomness of X_t and $\mathscr{H}_t$ as well as the (possible) randomness of the policy.

Remark 1. *A related non-stationary environment is the piecewise stationary environment [91], which allows θ_t's to change at most S times throughout the time horizon. However, as discussed in Section 1.1.1, this can be a looser measure of non-stationarity as a very tiny change in the environment is still counted towards the total number of switches. That is to say, even if there are a total of T switches, the total variation could still be far less than T.*

2.2 Lower Bound for Drifting Linear Bandits

We first provide a lower bound on the the dynamic regret for the linear model.

Theorem 1. *In the drifting linear bandit setting, for any $T \geq d$ and $B_T \in [dT^{-1/2}, 8d^{-2}T]$, there exists decision sets $\{D_t\}_{t=1}^T$ and reward vectors $\{\theta_t\}_{t=1}^T$, such that for all $t \in [T]$ and all $x \in D_t$, we have $\|x\| \leq 1$, $\|\theta_t\| \leq 1$, and $\|\langle x, \theta_t \rangle\| \leq 1$, and the dynamic regret for any non-anticipatory policy π satisfies $\mathscr{R}_T(\pi) = \Omega\left(d^{2/3} B_T^{1/3} T^{2/3}\right)$.*

Poof Sketch. The complete proof is presented in Section A.1 of the appendix. The construction of the lower bound instance is similar to the approach of [36]. The nature divides the whole time horizon into $\lceil T/H \rceil$ blocks of equal length $H = \lceil (dT)^{2/3} B_T^{-2/3} \rceil$ $(\leq T)$ rounds, and the last block can possibly have less than H rounds. In each block, the nature initiates a new stationary linear bandit instance with parameter vectors from the set $\{\pm\sqrt{d/4H}\}^d$. We set up the instance so that the parameter vector of a block cannot be learned using the observations from the previous blocks. Consequently, every online policy must incur a regret of $\Omega(d\sqrt{H})$ in each block, by applying the regret lower bound for stationary linear bandits (for example, see [125]) on each block. Since there are at least $\lfloor T/H \rfloor$ blocks, the total dynamic regret is $\Omega(dT/\sqrt{H}) = \Omega(d^{2/3} B_T^{1/3} T^{2/3})$. □

2.3 Sliding Window Regularized Least Squares Estimator

As a preliminary, we introduce the sliding window regularized least squares estimator (SW-RLSE), which is the key tool in estimating the unknown parameters $\{\theta_t\}_{t=1}^T$ online. The SW-RLSE generalizes the sliding window sample estimator proposed by [91] for the K-armed bandits in piecewise stationary environments. In addition, our SW-RLSE can be constructed for any sequence of arm pulls, which is different from [115], who require each arm (in their setting a posted price) to be pulled equally often. Despite the underlying non-stationarity in our model, we show that the estimation error of our SW-RLSE scales gracefully with the variation of θ_t's across time.

To motivate SW-RLSE, consider a round t, where the DM aims to estimate θ_t based on the historical observation $\{(X_s, Y_s)\}_{s=1}^{t-1}$. The design of SW-RLSE is based on the *forgetting principle* [91], which argues the following: the DM could estimate θ_t using only $\{(X_s, Y_s)\}_{s=1\vee(t-w)}^{t-1}$, the observation history during the time window $(1\vee(t-w))$ to $(t-1)$, instead of all prior observations. The rationale is that, under non-stationarity, the observations far in the past are obsolete, and they are not as informative for regressing θ_t. The principle crucially hinges on w, which is a positive integer called the window size. Intuitively, when the variation across $\theta_1, \ldots, \theta_T$ increases, the window size w should be smaller, since the past observations become obsolete at a faster rate. We treat w as a fixed parameter in this section, and then shine lights on choosing w in subsequent sections.

The SW-RLSE $\hat{\theta}_t$ is the optimal solution to the following ridge regression problem with regularization parameter $\lambda > 0$:

$$\min_{\theta:\theta\in\mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1\vee(t-w)}^{t-1} (X_s^\top \theta - Y_s)^2.$$

Define matrix $V_{t-1} := \lambda I + \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top$. The SW-RLSE $\hat{\theta}_t$ can be explicitly expressed

as

$$\hat{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1} X_s Y_s\right) = V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \theta_s + V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s. \tag{2.2}$$

Next, we demonstrate the accuracy of the SW-RLSE. Denoting

$$\beta := R\sqrt{d\ln\left(\frac{1+wL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S, \tag{2.3}$$

we provide an error bound on estimating the latent reward, *i.e.*, the confidence radius, of any action $x \in D_t$ in a round t, under the following regularity assumption made in [75] over the decision sets D_t's.

Assumption 1. *There exists an orthonormal basis $\Psi = (\psi_1, \ldots, \psi_d)$ such that for any $t \in [T]$ and any $X \in D_t$, there exists a number $z \in \mathbb{R}$ and an $i \in [d]$ such that $X = z \cdot \psi_i$.*

Remark 2. *One can easily verify that this assumption holds in the multi-armed bandits case. Of course, this assumption allows for more general models than the multi-armed bandits setting as it still allows each of the time-varying D_t's to have arbitrarily large number of actions.*

In what follows, we analyze the linear bandit setting under Assumption 1. We also discuss how to remove this assumption in Remark 4 of the forthcoming Section 2.5.

Theorem 2. *For any $t \in [T]$ and any $\delta \in [0,1]$, we have with probability at least $1-\delta$,*

$$\left|x^\top(\hat{\theta}_t - \theta_t)\right| \le L\sum_{s=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta\,\|x\|_{V_{t-1}^{-1}}$$

holds for all $x \in D_t$.

Proof Sketch. The complete proof is in Section A.2 of the appendix. Note that

$$\hat{\theta}_t - \theta_t = V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t) + V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s - \lambda\,\theta_t\right),$$

we first upper bound the first term as $\left\|V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}X_sX_s^\top(\theta_s-\theta_t)\right\|_2 \le \sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_2$, and then adopts Theorem 2 from [3] for the second term, *i.e.*, with probability at least $1-\delta$, $\left\|\sum_{s=1\vee(t-w)}^{t-1}\eta_sX_s-\lambda\theta_t\right\|_{V_{t-1}^{-1}} \le \beta$. Therefore, fixed any $\delta\in[0,1]$, we have that for any $t\in[T]$ and any $x\in D_t$,

$$\begin{aligned}\left|x^\top(\hat{\theta}_t-\theta_t)\right| &= \left|x^\top\left(V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}X_sX_s^\top(\theta_s-\theta_t)\right)+x^\top V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1}\eta_sX_s-\lambda\theta_t\right)\right| \\ &\le \|x\|_2\cdot\left\|V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}X_sX_s^\top(\theta_s-\theta_t)\right\|_2+\|x\|_{V_{t-1}^{-1}}\left\|\sum_{s=1\vee(t-w)}^{t-1}\eta_sX_s-\lambda\theta_t\right\|_{V_{t-1}^{-1}} \qquad (2.4)\\ &\le L\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_2+\beta\|x\|_{V_{t-1}^{-1}},\end{aligned}$$

where we have applied the triangle inequality and the Cauchy-Schwarz inequality successively in inequality (2.4). □

2.4 Sliding Window-Upper Confidence Bound (SW-UCB) Algorithm: An Optimal Strategy

In this section, we describe the Sliding Window Upper Confidence Bound (SW-UCB) algorithm for the linear model. When the variation budget B_T is known, we show that SW-UCB algorithm with a tuned window size achieves a dynamic regret bound which is optimal up to a multiplicative logarithmic factor. When the variation budget B_T is unknown, we show that SW-UCB algorithm can still be implemented with a suitably chosen window size so that the regret dependency on T is optimal, akin to that of [115].

2.4.1 Design Intuition

In the stochastic environment where the reward function is stationary, the well known UCB algorithm follows the principle of optimism in face of uncertainty [21, 3]. Under this principle, the DM selects an action that maximizes the UCB, which is the value of "mean

plus confidence radius" [21]. To follow the principle, we first construct an UCB on the latent mean reward $\langle x, \theta_t \rangle$ for each $x \in D_t$ in each round $t \in [T]$. By Theorem 2, the UCB of $x \in D_t$ in each round $t \in [T]$ is $\langle x, \hat{\theta}_t \rangle + L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\| + \beta \|x\|_{V_{t-1}^{-1}}$. We then choose the action X_t with the highest UCB, *i.e.*,

$$X_t = \arg\max_{x \in D_t} \left\{ \langle x, \hat{\theta}_t \rangle + L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\| + \beta \|x\|_{V_{t-1}^{-1}} \right\} = \arg\max_{x \in D_t} \left\{ \langle x, \hat{\theta}_t \rangle + \beta \|x\|_{V_{t-1}^{-1}} \right\}. \tag{2.5}$$

Upon selecting X_t, we have

$$\langle x_t^*, \hat{\theta}_t \rangle + L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta \|x_t^*\|_{V_{t-1}^{-1}} \leq \langle X_t, \hat{\theta}_t \rangle + L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta \|X_t\|_{V_{t-1}^{-1}} \tag{2.6}$$

by virtue of the UCB action selection rule. From Theorem 2, we further have with probability at least $1 - \delta$,

$$\langle x_t^*, \theta_t \rangle \leq \langle x_t^*, \hat{\theta}_t \rangle + L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta \|x_t^*\|_{V_{t-1}^{-1}} \tag{2.7}$$

and

$$\langle X_t, \hat{\theta}_t \rangle + L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta \|X_t\|_{V_{t-1}^{-1}} \leq \langle X_t, \theta_t \rangle + 2L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + 2\beta \|X_t\|_{V_{t-1}^{-1}}. \tag{2.8}$$

Combining inequalities (2.6), (2.7), and (2.8), we establish the following high probability upper bound for the expected per round regret, *i.e.*, with probability $1 - \delta$,

$$\langle x_t^* - X_t, \theta_t \rangle \leq 2L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + 2\beta \|X_t\|_{V_{t-1}^{-1}}. \tag{2.9}$$

The regret upper bound of the SW-UCB algorithm (to be formalized in Theorem 3) is thus

$$2\sum_{t\in[T]} L \sum_{s=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta \|X_t\|_{V_{t-1}^{-1}} = \widetilde{O}\left(wB_T + \frac{dT}{\sqrt{w}}\right). \tag{2.10}$$

2.4.2 Design Details

In this section, we describe the details of the SW-UCB algorithm following the discussions above. The SW-UCB algorithm selects a regularization parameter λ (> 0), and initializes $V_0 = \lambda I$. In each round t, the SW-UCB algorithm first computes the estimate $\hat{\theta}_t$ for θ_t according to eq. (2.2), and then finds the action X_t with largest UCB by solving the optimization problem (2.5). Afterwards, the corresponding reward Y_t is observed. The pseudo-code of the SW-UCB algorithm is shown in Algorithm 1.

Algorithm 1 SW-UCB algorithm for drifting linear bandits

1: **Input:** Sliding window size w, dimension d, variance proxy of the noise terms R, upper bound of all the actions' Euclidean norms L, upper bound of all the θ_t's Euclidean norms S, and regularization constant λ.
2: **Initialization:** $V_0 \leftarrow \lambda I$.
3: **for** $t = 1, \ldots, T$ **do**
4: Update $\hat{\theta}_t \leftarrow V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1} X_s Y_s\right)$.
5: $X_t \leftarrow \arg\max_{x\in D_t}\left\{x^\top \hat{\theta}_t + \beta \|x\|_{V_{t-1}^{-1}}\right\}$, where β is defined in (2.3).
6: Observe $Y_t = \langle X_t, \theta_t\rangle + \eta_t$.
7: Update $V_t \leftarrow \lambda I + \sum_{s=1\vee(t-w+1)}^{t} X_s X_s^\top$.
8: **end for**

2.4.3 Dynamic Regret Analysis

We are now ready to formally state a dynamic regret upper bound of the SW-UCB algorithm for drifting linear bandits.

Theorem 3. *For the drifting linear bandit setting, the dynamic regret of the SW-UCB algorithm is upper bounded as* $\mathscr{R}_T(\text{SW-UCB algorithm}) = \widetilde{O}(wB_T + dT/\sqrt{w})$. *When* B_T *is known, by*

taking $w = \Theta\left((dT)^{2/3} B_T^{-2/3}\right)$, *the dynamic regret of the* `SW-UCB` *algorithm is*

$$\mathscr{R}_T\left(\texttt{SW-UCB algorithm}\right) = \widetilde{O}\left(d^{\frac{2}{3}} B_T^{\frac{1}{3}} T^{\frac{2}{3}}\right).$$

When B_T *is unknown, by taking* $w = \Theta\left((dT)^{2/3}\right)$, *the dynamic regret of the* `SW-UCB` *algorithm is*

$$\mathscr{R}_T\left(\texttt{SW-UCB algorithm}\right) = \widetilde{O}\left(d^{\frac{2}{3}} B_T T^{\frac{2}{3}}\right).$$

Poof Sketch. The complete proof is in Section A.3 of the appendix. The proof involves upper bounding the two terms on the L.H.S. of eq. (2.10). The first term can be upper bounded by a intuitive telescoping sum. For the second term, we first remark a similar quantity is analyzed in [3], which involves a matrix telescoping technique under stationarity. Nevertheless, due to the "forgetting principle" of the `SW-UCB` algorithm, we cannot directly adopt this. Instead, we make use of the Sherman-Morrison formula to overcome the barrier. □

Remark 3. *When the variation budget* B_T *is known, Theorem 3 recommends choosing the size* w *of the sliding window to be decreasing with* B_T. *The recommendation is in agreement with the intuition that, when the learning environment becomes more volatile, the DM should focus on more recent observations. Indeed, if the underlying learning environment is changing at a higher rate, then the DM's past observations become obsolete faster. Theorem 3 pins down the intuition of forgetting past observation in face of drifting environments, by providing the mathematical definition of the sliding window size* w *that yields the optimal dynamic regret bound.*

2.5 Bandit-over-Bandit (`BOB`) Algorithm: Adapting to the Unknown Variation Budget

When B_T is not known, the DM can achieve the dynamic regret bound $\widetilde{O}\left(d^{2/3}(B_T+1)T^{2/3}\right)$ for the drifting linear bandit problem, by setting $w = \Theta((dT)^{2/3})$ (see Section 2.4). While

the bound is optimal in terms of T by Theorem 1, it becomes meaningless when $B_T = \Omega(T^{1/3})$, since then the resulting dynamic regret bound is linear in T.

To overcome this limitation, in this section we design an online algorithm whose dynamic regret grows sub-linearly in T, even when $B_T = o(T)$ is not known. Similar to the style of previous sections, the discussion in this section focuses on linear model. Nevertheless, we emphasize that the proposed framework applies to a variety of bandit models (see the forthcoming Section 2.6).

2.5.1 Design Challenges

Theorem 3 shows that running the SW-UCB algorithm for T with window size

$$w^* = \left\lfloor (dT)^{2/3} B_T^{-2/3} \right\rfloor \tag{2.11}$$

leads to an optimal dynamic regret. However, the choice of the window size w^* in (2.11) requires the crucial knowledge of B_T, which is not available to the DM. A natural attempt would be to "learn" the unknown B_T in order to properly tune the window size w. In a more restrictive setting in which the differences between consecutive θ_t's follow some underlying stochastic process, one possible approach is to apply a suitable machine learning technique to learn the underlying stochastic process and tune the parameter w accordingly. However, under the general setting of drifting environments (2.1), the differences between consecutive θ_t's need not follow any pattern, which challenges the use of statistical machine learning algorithms for identifying the patterns on the underlying changes.

2.5.2 Algorithm

The above mentioned observations as well as the established results motivate us to make use of the SW-UCB algorithm as a sub-routine, and "hedge" [18, 17] against the (possibly adversarial) changes of θ_t's to identify a reasonable fixed window size. Inspired by the heuristic envelop policy [38] and the bandit corralling technique [7, 134], we develop a novel Bandit-over-Bandit (BOB) algorithm that achieves a nearly optimal dynamic regret bound for drifting linear bandits. Specifically, we show In Section 2.5.4 that the BOB algo-

rithm has a dynamic regret sub-linear in T even when $B_T = o(T)$ is not known, unlike the SW-UCB algorithm.

As illustrated in Fig. 2-1, the BOB algorithm divides the whole time horizon into $\lceil T/H \rceil$ blocks of equal length H rounds (the last block can possibly have less than H rounds). In addition, the algorithm specifies a set of candidate window sizes J. For each block $i \in [\lceil T/H \rceil]$, the BOB algorithm first selects a window size $w_i \in J$. Then, the BOB algorithm restarts the SW-UCB algorithm from scratch (see Remark 6 for a discussion on the design of restarting) with the selected window size w_i for H rounds. On top of this, the BOB algorithm also maintains a separate bandit algorithm to determine each window size w_i based on the observed history in the previous $i-1$ blocks, and thus the name Bandit-over-Bandit. The choice of w_i is based on the EXP3 algorithm [18], which allows us to compete with the best window size in J (in the sense of minimizing dynamic regret), even when the θ_t's variation does not follow any pattern. The EXP3 algorithm is designed for adversarial multi-armed bandits, where the underlying reward function is designed by an oblivious adversary [18, 17]. Finally, to properly apply the EXP3 algorithm, we note that the total reward during each block is normalized so that the normalized reward lies in $[0,1]$ with high probability.

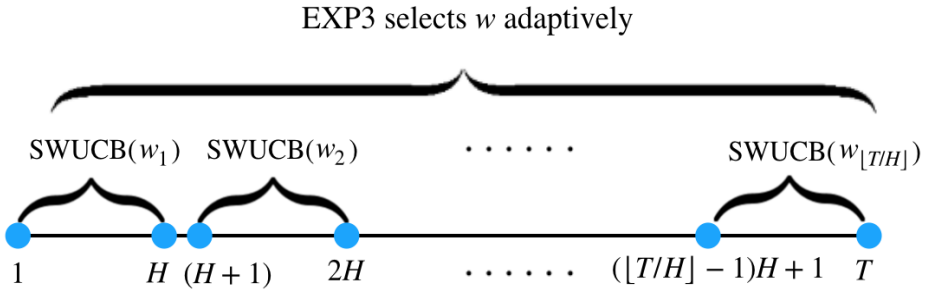


Figure 2-1: Structure of the BOB algorithm

To this end, we describe the details of the BOB algorithm for the linear model. Defining the parameters (we justify these choices in Section 3.4.1)

$$H = \left\lfloor dT^{\frac{1}{2}} \right\rfloor, \Delta = \lceil \ln H \rceil, J = \left\{ H^0, \left\lfloor H^{\frac{1}{\Delta}} \right\rfloor, \ldots, H \right\}, Q = 2H + 4R\sqrt{H \ln(T/\sqrt{H})}, \quad (2.12)$$

The BOB algorithm first divides the time horizon T into $\lceil T/H \rceil$ blocks of length H rounds

(except for the last block, which can be less than H rounds), and then initiates the parameters

$$\gamma = \min\left\{1, \sqrt{\frac{(\Delta+1)\ln(\Delta+1)}{(e-1)\lceil T/H\rceil}}\right\}, s_{j,1} = 1 \quad \forall j = 0,1,\ldots,\Delta. \tag{2.13}$$

for the EXP3 algorithm [18]. At the beginning of each block $i \in [\lceil T/H\rceil]$, the BOB algorithm first sets

$$p_{j,i} = (1-\gamma)\frac{s_{j,i}}{\sum_{u=0}^{\Delta} s_{u,i}} + \frac{\gamma}{\Delta+1} \quad \forall j = 0,1,\ldots,\Delta, \tag{2.14}$$

and then sets $j_i = j$ with probability $p_{j,i}$ for each $j = 0,1,\ldots,\Delta$. The selected window size is then $w_i = \left\lfloor H^{j_i/\Delta} \right\rfloor$. Afterwards, the BOB algorithm selects actions X_t by running the SW-UCB algorithm with window size w_i for each round t in block i, and the total collected reward is

$$\sum_{t=(i-1)H+1}^{i\cdot H\wedge T} Y_t = \sum_{t=(i-1)H+1}^{i\cdot H\wedge T} \langle X_t, \theta_t\rangle + \eta_t.$$

Finally, the total rewards is normalized by first dividing Q, and then added by $1/2$ so that it lies within $[0,1]$ with high probability. The parameter $s_{j_i,i+1}$ is set to

$$s_{j_i,i} \cdot \exp\left(\frac{\gamma}{(\Delta+1)p_{j_i,i}}\left(\frac{1}{2} + \frac{\sum_{t=(i-1)H+1}^{i\cdot H\wedge T} Y_t}{Q}\right)\right); \tag{2.15}$$

while $s_{u,i+1}$ is the same as $s_{u,i}$ for all $u \neq j_i$. The pseudo-code of the BOB algorithm is shown in Algorithm 2.

Algorithm 2 BOB algorithm for drifting linear bandits

1: **Input:** Time horizon T, the SW-UCB algorithm, parameters H, Δ, J, Q (as defined in 2.12).
2: **Initialize** parameters $\gamma, \{s_{j,1}\}_{j=0}^{\Delta}$ by eq. (2.13).
3: **for** $i = 1,2,\ldots,\lceil T/H\rceil$ **do**
4: Define distribution $(p_{j,i})_{j=0}^{\Delta}$ by eq. (2.14), and set $j_t \leftarrow j$ with probability $p_{j,i}$.
5: Set the window size $w_i \leftarrow \left\lfloor H^{j_t/\Delta} \right\rfloor$.
6: Restart the SW-UCB algorithm for H rounds with window size w_i.
7: Update $s_{j_i,i+1}$ according to eq. (2.15), and $s_{u,i+1} \leftarrow s_{u,i}\ \forall u \neq j_i$
8: **end for**

2.5.3 Choice of Parameters

We first justify the choice of Q. Note that Q is used to perform normalization, we thus prove high probability upper and lower bounds for the total rewards of each block (here, we prove a slightly more general result by allowing $\max_{t\in[T],x\in D_t}|\langle x,\theta_t\rangle|$ to be in $[-\nu,\nu]$ for some $\nu>0$).

Lemma 4. *Suppose* $\max_{t\in[T],x\in D_t}|\langle x,\theta_t\rangle|\in[-\nu,\nu]$ *for some* $\nu>0$ *and denote* M_i *as the absolute value of cumulative rewards for block* i*, then with probability at least* $1-2/T$, M_i *does not exceed* $H\nu+2R\sqrt{H\ln(T/\sqrt{H})}$ *for all* i*, i.e.,*

$$\Pr\left(\forall i\in\lceil T/H\rceil\quad M_i\le H\nu+2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right)\ge 1-\frac{2}{T}.$$

The complete proof of Lemma 4 is in Section A.4 of the appendix. With Lemma 4 and the choice of $Q=2H+4R\sqrt{H\ln(T/\sqrt{H})}$ (note that $\nu=1$ by our model assumption in Section 2.1), it is evident that $\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}Y_t/Q$ in eq. (2.15) lies in $[-1/2,1/2]$ with probability at least $1-2/T$. Adding this by $1/2$, we normalize the total rewards of each block to $[0,1]$ with probability at least $1-2/T$ for all the blocks.

To determine H,Δ, and J, we first consider the dynamic regret of the `BOB` algorithm. Here, we point out due to the design of restarting, any instance of the `SW-UCB` algorithm cannot last for more than H rounds. As a consequence, even if the EXP3 selects a window size $w_i>H$ for some block i, the effective window size is H. In other words, w^* is not necessarily attainable, *i.e.*, by definition in eq. (2.11), $w^*=\left\lfloor(dT)^{2/3}B_T^{-2/3}\right\rfloor$ might be larger than H when B_T is small. We thus have to denote the optimally (over J) tuned window size as $w^\dagger$, and derive the following result.

Proposition 5. *For the drifting linear bandit setting, the dynamic regret of the* `BOB` *algorithm is*

$$\mathscr{R}_T(\texttt{BOB algorithm})=\widetilde{O}\left(w^\dagger B_T+\frac{dT}{\sqrt{w^\dagger}}+Q\sqrt{\frac{|J|T}{H}}\right).\tag{2.16}$$

Proof Sketch. The complete proof is presented in Section A.5 of the appendix. The dy-

namic regret bound (2.16) can be decomposed as

$$\underbrace{\widetilde{O}\left(w^{\dagger}B_T + \frac{dT}{\sqrt{w^{\dagger}}}\right)}_{\mathscr{R}_T(\texttt{SW-UCB algorithm}) \text{ with } w^{\dagger}} + \underbrace{\widetilde{O}\left(Q\sqrt{\frac{|J|T}{H}}\right)}_{\text{Loss in learning } w^{\dagger}}. \tag{2.17}$$

The first term in (2.17) is due to the dynamic regret of the underlying SW-UCB algorithm under the optimally tuned window size $w^{\dagger}$. More precisely, we can view each block as a new non-stationary linear bandit instance, and the dynamic regret is due to the application of SW-UCB algorithm with window size $w^{\dagger}$ on each block. The second term is due to the loss by the EXP3 algorithm, which essentially treat each of the window size in J as an expert, and compete with the best expert. □

Eq. (2.16) exhibits a similar structure to the regret of the SW-UCB algorithm as stated in Theorem 3, and this immediately indicates a clear trade-off in the design of the block length H :

- On one hand, H should be small to control the regret incurred by the EXP3 algorithm in identifying $w^{\dagger}$, *i.e.*, the third term in eq. (2.16).

- On the others, H should also be large enough to allow $w^{\dagger}$ to get close to $w^* = \lfloor (dT)^{2/3} B_T^{-2/3} \rfloor$ so that the sum of the first two terms in eq. (2.16) is minimized.

A more careful inspection also reveals the tension in the design of J. Obviously, we hope that $|J|$ is small to minimize the third term in eq. (2.16), but we also wish J to be dense enough so that it forms a cover to the set $[H]$. Otherwise, even if H is large enough that $w^{\dagger}$ can approach w^*, approximating w^* with any element in J can cause a major loss.

These observations suggest the following choice of J.

$$J = \left\{H^0, \left\lfloor H^{\frac{1}{\Delta}} \right\rfloor, \ldots, H\right\} \tag{2.18}$$

for some positive integer Δ, and since the choice of H should not depend on B_T, we can set $H = \lfloor d^{\varepsilon} T^{\alpha} \rfloor$ with some $\alpha \in [0,1]$ and $\varepsilon > 0$ to be determined. We then distinguish two

cases depending on whether w^* is smaller than H or not (or alternatively, whether B_T is larger than $d^{(2-3\varepsilon)/2}T^{(2-3\alpha)/2}$ or not).

Case 1: $w^* \leq H$ **or** $B_T \geq d^{(2-3\varepsilon)/2}T^{(2-3\alpha)/2}$. Under this situation, $w^\dagger$ can automatically adapt to the nearly optimal window size $\mathrm{clip}_J(w^*)$, where $\mathrm{clip}_J(x)$ finds the largest element in J that does not exceed x. Notice that $|J| = \Delta + 1$, the dynamic regret of the BOB algorithm then becomes

$$\begin{aligned}
\mathscr{R}_T(\texttt{BOB algorithm}) =& \widetilde{O}\left(w^\dagger B_T + \frac{dT}{\sqrt{w^\dagger}} + \sqrt{H|J|T}\right) \\
=& \widetilde{O}\left(w^* H^{\frac{1}{\Delta}} B_T + \frac{dT}{\sqrt{w^* H^{-1/\Delta}}} + \sqrt{d^\varepsilon T^{\alpha+1}\Delta}\right) \\
=& \widetilde{O}\left(d^{\frac{2}{3}}\left(B_T+1\right)^{\frac{1}{3}} T^{\frac{2}{3}} H^{\frac{1}{\Delta}} + d^{\frac{\varepsilon}{2}} T^{\frac{\alpha+1}{2}} \Delta^{\frac{1}{2}}\right). \qquad (2.19)
\end{aligned}$$

Case 2: $w^* > H$ **or** $B_T < d^{(2-3\varepsilon)/2}T^{(2-3\alpha)/2}$. Under this situation, $w^\dagger$ equals to H, which is the window size closest to w^*, the regret of the BOB algorithm then becomes

$$\begin{aligned}
\mathscr{R}_T(\texttt{BOB algorithm}) =& \widetilde{O}\left(w^\dagger B_T + \frac{dT}{\sqrt{w^\dagger}} + \sqrt{H|J|T}\right) \\
=& \widetilde{O}\left(HB_T + \frac{dT}{\sqrt{H}} + \sqrt{H|J|T}\right) \\
=& \widetilde{O}\left(d^\varepsilon\left(B_T+1\right)T^\alpha + d^{1-\frac{\varepsilon}{2}} T^{\frac{2-\alpha}{2}} + + d^{\frac{\varepsilon}{2}} T^{\frac{\alpha+1}{2}} \Delta^{\frac{1}{2}}\right) \\
=& \widetilde{O}\left(d^{1-\frac{\varepsilon}{2}} T^{\frac{2-\alpha}{2}} + d^{\frac{\varepsilon}{2}} T^{\frac{\alpha+1}{2}} \Delta^{\frac{1}{2}}\right), \qquad (2.20)
\end{aligned}$$

where we have make use of the fact that $B_T < d^{(2-3\varepsilon)/2}T^{(2-3\alpha)/2}$ in the last step.

Now both eq. (2.19) and eq. (2.20) suggests that we should set $\Delta = \lceil \ln H \rceil$, and eq. (2.20) further reveals that we should take $\alpha = 1/2$ and $\varepsilon = 1$. These then lead to the choice of parameters presented in eq. (2.12), *i.e.*, $H = \left\lfloor dT^{\frac{1}{2}} \right\rfloor, \Delta = \lceil \ln H \rceil, J = \left\{H^0, \left\lfloor H^{\frac{1}{\Delta}} \right\rfloor, \ldots, H\right\}$. Here we have to emphasize that $w^\dagger, \alpha$, and ε are used only in the analysis, while the only parameters that we need to decide are H, Δ, J, and Q, which clearly do not depend on B_T.

2.5.4 Dynamic Regret Analysis

We are now ready to present the dynamic regret analysis of the BOB algorithm for the drifting linear bandits.

Theorem 6. *The dynamic regret of the BOB algorithm for drifting linear bandit is*

$$\mathscr{R}_T\left(\textit{BOB algorithm}\right) = \widetilde{O}\left(d^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}} + d^{\frac{1}{2}}T^{\frac{3}{4}}\right).$$

Proof Sketch. The proof of the theorem essentially follows from substituting the choice of H and J into the dynamic regret bound in Proposition 5, and the complete proof is presented in Section A.6 of the appendix. □

Remark 4 (Removing Assumption 1)**.** *To remove Assumption 1, one can apply a restarting strategy [38] together with an algorithm for adversarial linear bandit, e.g., Algorithm 15 of [125]. When B_T is known and D_t's are fixed, by an argument similar to Theorem 2 of [38], one can show that this restarting strategy can achieve the minimax-optimal dynamic regret bound $\widetilde{O}(d^{2/3}B_T^{1/3}T^{2/3})$; when B_T is unknown, we can apply the BOB algorithm to adaptively tune the restarting rate to achieve the dynamic regret bound $\widetilde{O}(d^{2/3}B_T^{1/3}T^{2/3} + d^{1/2}T^{3/4})$.*

Remark 5. *Compared to the lower bound in Theorem 1, the dynamic regret bound presented in Theorem 6 is optimal when $B_T \geq d^{-1/2}T^{1/4}$; while it also leaves a small $O(T^{1/12})$ gap in the worst case i.e., when $B_T = \Theta(1)$. This is because the smaller the non-stationarity, the harder the detection, and hence a worse dynamic regret bound.*

Remark 6. *The block structure and restarting the SW-UCB algorithm with a single window size for each block are essential for the correctness of the BOB algorithm. Otherwise, suppose the DM utilizes the EXP3 algorithm to select the window size w_t for each round t, and implements the SW-UCB algorithm with the selected window size without ever restarting it. Instead of eq. (A.33), the regret of the BOB algorithm is then decomposed as*

$$\sum_{t=1}^{T}\left(\textit{Reward of SW-UCB}\left(\left\{w^{\dagger}\right\}_{\tau=1}^{t}\right)\ \textit{in round } t - \textit{Reward of SW-UCB}\left(\{w_\tau\}_{\tau=1}^{t}\right)\ \textit{in round } t\right)$$

$$+\sum_{t=1}^{T}\left(\textit{Optimal reward in round } t - \textit{Reward of } \texttt{SW-UCB}\left(\left\{w^{\dagger}\right\}_{\tau=1}^{t}\right) \textit{ in round } t\right) \quad (2.21)$$

Here, with some abuse of notations, $\texttt{SW-UCB}(\{w^{\dagger}\}_{\tau=1}^{t})$ *(respectively* $(\texttt{SW-UCB}(\{w_{\tau}\}_{\tau=1}^{t}))$ *refers to in round* t*, the DM runs the* `SW-UCB` *algorithm with window size* $w^{\dagger}$ *(respectively* w_t*) and historical data, e.g., (action, reward) pairs, generated by running the* `SW-UCB` *algorithm with window size* $w^{\dagger}$ *(respectively* w_{τ}*) for rounds* $\tau = 1,\ldots,t-1$*. Same as before, the second term of eq. (2.21) can be upper bounded as a result of Theorem 3. It is also tempting to apply results from the EXP3 algorithm to upper bound the first term. Unfortunately, this is incorrect as it is required by the adversarial bandits protocol [18] that the DM and its competitor should receive the same reward if they select the same action, i.e., the reward of* $\texttt{SW-UCB}\left(\{w^{\dagger}\}_{\tau=1}^{t-1}, w_t = w\right)$ *in round* t *and the reward of* $\texttt{SW-UCB}\left(\{w_{\tau}\}_{\tau=1}^{t-1}, w_t = w^{\dagger}\right)$ *in round* t *should be the same for every* w*. Nevertheless, this is violated as running the* `SW-UCB` *algorithm with different window sizes for previous rounds can generate different (action,reward) pairs, and this results in possibly different estimated* $\hat{\theta}_t$*'s for the two* `SW-UCB` *algorithms even if both of them use the same window size in round* t*. Hence, the selected actions and the corresponding reward by these two instances might also be different. By the careful design of blocks as well as the restarting scheme, the* `BOB` *algorithm decouples the* `SW-UCB` *algorithm for a block from previous blocks, and thus fixes the above mentioned problem, i.e., the regret of the* `BOB` *algorithm is decomposed as eq. (A.33).*

Remark 7. *The bandit-over-bandit framework can go beyond the problem of non-stationary bandit optimization. In a high level, it provides us a viable approach to automatically optimize the performances of data-driven sequential decision-making algorithms. Although not always optimal, it can be applied to bandit model selection [84] as well as online meta-learning [32], in which the DM is trying to optimize the performances of her algorithms by selecting a correct model class or a set of proper parameters. Both of these are of great importance in the operations of data-driven decision-making algorithms.*

2.6 Applications to Other Bandit Settings

In this section, we demonstrate the generality of our established results. As illustrative examples, we apply our technique to several bandit settings, including multi-armed bandits [21], the generalized linear bandits [79, 128], and the combinatorial semi-bandits [88, 123]. Note that for generalized linear bandits, we need to impose Assumption 1. On the other hand, for multi-armed bandits, this assumption is always valid while for combinatorial semi-bandits, this assumption is not required. A preview of the results is shown in Table 2.1.

	Known B_T	Unknown B_T
d-armed bandit	$\widetilde{O}\left(d^{1/3}B_T^{1/3}T^{2/3}\right)$	$\widetilde{O}\left(d^{1/3}B_T^{1/3}T^{2/3}+d^{1/4}T^{3/4}\right)$
Generalized linear bandit	$\widetilde{O}\left(d^{2/3}B_T^{1/3}T^{2/3}\right)$	$\widetilde{O}\left(d^{2/3}B_T^{1/3}T^{2/3}+d^{1/2}T^{3/4}\right)$
Combinatorial semi-bandit	$\widetilde{O}\left(d^{1/3}m^{2/3}B_T^{1/3}T^{2/3}\right)$	$\widetilde{O}\left(d^{1/3}m^{2/3}B_T^{1/3}T^{2/3}+d^{1/4}m^{3/4}T^{3/4}\right)$

Table 2.1: Dynamic regret bounds of the `SW-UCB` algorithm and the `BOB` algorithm for different settings. Here m is an upper bound for the 1-norm of all the actions in the combinatorial semi-bandit problem.

2.6.1 An Algorithmic Template

The `SW-UCB` algorithm and the `BOB` algorithm developed in the previous sections can be viewed as an algorithmic template that allows us to extend the results from linear bandits to other bandit settings. Given a bandit setting `A`, we leverage the forgetting principle (similar to Section 2.3), and first modify the reward estimator used in the stationary setting to a sliding-window estimator. We then incorporate it into the UCB algorithm to arrive at the corresponding `SW-UCB` algorithm for the drifting environments. When the variation budget is known, we could optimally tune the window size to enjoy an optimal dynamic regret bound. To achieve low dynamic regret when the variation budget is unknown, we can proceed by plugging the `SW-UCB` algorithm for `A` into the `BOB` algorithm, *i.e.*, line 6 of Algorithm 2, and custom-tailor the parameters (as those listed in eq. (2.12)) to accommodate the need of `A`.

We note that the power of this algorithmic template is indeed entailed by a salient property, *i.e.*, the dynamic regret of the `SW-UCB` algorithm can be decomposed as "dynamic regret of drift" + "dynamic regret of uncertainty" (or eq. (2.10)), that actually holds for a variety of bandit learning models in addition to linear models. In what follows, we shall derive the `SW-UCB` algorithm as well as the parameters required by the `BOB` algorithm, *i.e.*, similar to those defined in eq. (2.12), for each of the above mentioned settings.

2.6.2 d-Armed Bandits

The d-armed bandit problem in drifting environments was first studied by [37], who proposed Rexp3, an innovative and interesting variant of the EXP3 algorithm [22]. When the underlying variation budget is known, their algorithm achieves the optimal dynamic regret bound. In this subsection, we provide an alternative derivation of the dynamic regret bound by our framework.

In the d-armed bandits setting, every action set D_t is comprised of d actions $e_1, \ldots, e_d$. The i^{th} action e_i has coordinate i equals to 1 and all other coordinates equal to 0. Therefore, the reward of choosing action $X_t = e_{I_t}$ in round t is $Y_t = \langle X_t, \theta_t \rangle + \eta_t = \theta_t(I_t) + \eta_t$, where $\theta_t(I_t)$ is the I_t^{th} coordinate of θ_t. We again assume $|\langle x, \theta_t \rangle| \in [-1, 1]$ for all $x \in D_t$ and all $t \in [T]$. Different than the linear bandit setting, we follow [37, 38] to define the tighter variation budget with the infinity norm, *i.e.*, $\sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|_\infty \leq B_T$. For a window size w, we also define $N_{t-1}(i)$ as the number of times that action i is chosen within rounds $(t-w), \ldots, (t-1)$, *i.e.*, for all $i \in [d]$, $N_{t-1}(i) = \sum_{s=1 \vee (t-w)}^{t-1} \mathbf{1}[X_t = e_i]$. Here $\mathbf{1}[\cdot]$ is the indicator function. Similar to the procedure in Section 2.3, we set the regularization parameter $\lambda = 0$, and compute the sliding window least squares estimate $\hat{\theta}_t$ for θ_t in each round, *i.e.*,

$$\hat{\theta}_t = V_{t-1}^* \left(\sum_{s=1 \vee (t-w)}^{t-1} X_s Y_s \right), \tag{2.22}$$

where V_{t-1}^* is Moore-Penrose pseudo-inverse of V_{t-1}. We can also derive the error bound for the latent expected reward of every action $x \in D_t$ in any round t.

Theorem 7. *For any $t \in [T]$ and any $i \in [d]$, we have with probability at least $1 - 1/T$,*

$$\left| e_i^\top (\hat{\theta}_t - \theta_t) \right| \leq \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_\infty + R\sqrt{2\ln(2dT^2)}\, \|e_i\|_{V_{t-1}^*}.$$

holds for all $x \in D_t$.

The complete proof is provided in Section A.7 of the appendix. We can now follow the same principle in Section 2.4 by choosing in each round the action X_t with the highest UCB, *i.e.*,

$$X_t = \arg\max_{x \in D_t} \left\{ \langle x, \hat{\theta}_t \rangle + R\sqrt{2\ln(2dT^2)}\, \|x\|_{V_{t-1}^*} \right\}, \tag{2.23}$$

and arrive at the following regret upper bound for the `SW-UCB` algorithm.

Theorem 8. *For the d-armed bandit setting, the dynamic regret of the `SW-UCB` algorithm is upper bounded as $\mathscr{R}_T$ (`SW-UCB` algorithm) $= \widetilde{O}\left(wB_T + \sqrt{d}T/\sqrt{w}\right)$. When B_T (> 0) is known, by taking $w = \Theta\left(d^{1/3}T^{2/3}B_T^{-2/3}\right)$, the dynamic regret of the `SW-UCB` algorithm is $\mathscr{R}_T$ (`SW-UCB` algorithm) $= \widetilde{O}\left(d^{1/3}B_T^{1/3}T^{2/3}\right)$. When B_T is unknown, by taking $w = \Theta\left(d^{1/3}T^{2/3}\right)$, the dynamic regret of the `SW-UCB` algorithm is $\mathscr{R}_T$ (`SW-UCB` algorithm) $= \widetilde{O}\left(d^{1/3}B_T T^{2/3}\right)$.*

Proof Sketch. The proof of this theorem is very similar to that of Theorem 3, and is thus omitted. The key difference is that β (defined in eq. (2.3) for the linear bandit setting) is now set to $R\sqrt{2\ln(2dT^2)}$, and this saves the extra $\sqrt{d}$ factor presented in eq. (A.29). Hence the dynamic regret bound can be obtained accordingly. □

Comparing the results obtained in Theorem 8 to the lower bound presented in [37], we can easily see that the dynamic regret bound is optimal when B_T is known. When B_T is unknown, we can implement the `BOB` algorithm with the following parameters:

$$H = \left\lfloor (dT)^{\frac{1}{2}} \right\rfloor, \Delta = \lceil \ln H \rceil, J = \left\{ H^0, \left\lfloor H^{\frac{1}{\Delta}} \right\rfloor, \ldots, H \right\}, Q = 2H + 4R\sqrt{H \ln(T/\sqrt{H})}. \tag{2.24}$$

The regret of the `BOB` algorithm for the MAB setting is characterized as follows.

Theorem 9. *The dynamic regret of the* BOB *algorithm for the d-armed bandit setting is* $\mathscr{R}_T(\texttt{BOB algorithm}) = \widetilde{O}\left(d^{1/3}B_T^{1/3}T^{2/3} + d^{1/4}T^{3/4}\right)$.

The proof of the theorem is very similar to Theorem 6's, and it is thus omitted.

2.6.3 Generalized Linear Bandits

For the generalized linear bandits model, we adopt the setup in [79, 128]: it is essentially the same as the linear bandit setting except that the decision set is time invariant, *i.e.*, $D_t = D$ for all $t \in [T]$, and the reward of choosing action $X_t \in D$ is $Y_t = \mu(\langle X_t, \theta_t \rangle) + \eta_t$.

Let $\dot{\mu}(\cdot)$ and $\ddot{\mu}(\cdot)$ denote the first derivative and second derivative of $\mu(\cdot)$, respectively, we follow [79] to make the following assumptions.

Assumption 2. *There exists a set of d actions* $a_1, \ldots, a_d \in D$ *such that the minimal eigenvalue of* $\sum_{i=1}^{d} a_i a_i^\top$ *is* λ_0 (> 0).

Assumption 3. *The link function* $\mu(\cdot) : \mathbb{R} \to \mathbb{R}$ *is strictly increasing, continuously differentiable, Lipschitz with constant* k_μ, *and we define* $c_\mu = \inf_{x \in D, \theta \in \mathbb{R}^d : \|\theta\| \le S} \dot{\mu}(\langle x, \theta \rangle)$.

Assumption 4. *There exists* $Y_{\max} > 0$ *such that for any* $t \in [T]$, $Y_t \in [0, Y_{\max}]$.

Similar to the procedure in Section 2.3, we compute the maximum quasi-likelihood estimate $\hat{\theta}_t$ for θ_t in each round $t \in [T]$ by solving the equation

$$\sum_{s=1 \vee (t-w)}^{t-1} \left(Y_s - \mu\left(\left\langle X_s, \hat{\theta}_t \right\rangle\right)\right) X_s = 0. \tag{2.25}$$

Defining $\beta = 2k_\mu Y_{\max} \sqrt{2d \ln(w) \ln(2dT^2) \left(3 + 2\ln\left(1 + 2L^2/\lambda_0\right)\right)} / c_\mu$, we can also derive the deviation inequality type bound for the latent expected reward of every action $x \in D_t$ in any round t.

Theorem 10. *For any* $t \in [T]$, *we have with probability at least* $1 - 1/T$,

$$\left|\mu\left(x^\top \hat{\theta}_t\right) - \mu\left(x^\top \theta_t\right)\right| \le \frac{k_\mu^2 L}{c_\mu} \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \beta \|x\|_{V_{t-1}^{-1}}$$

holds for all $x \in D_t$.

Proof Sketch. The proof is a consequence of Proposition 1 of [79] and Theorem 2. Please refer to Section A.8 of the appendix for the complete proof. □

We can now follow the same principle in Section 2.4 to design the `SW-UCB` algorithm. Note that in order for V_{t-1} to be invertible for all t, our algorithm should select the actions $a_1, \ldots, a_d$ every w rounds for some window size w. For each of the remaining round t, it chooses the action X_t with the highest UCB, *i.e.*,

$$X_t = \arg\max_{x \in D_t} \left\{ \langle x, \hat{\theta}_t \rangle + \beta \|x\|_{V_{t-1}^*} \right\}, \tag{2.26}$$

and arrive at the following regret upper bound.

Theorem 11. *For the drifting generalized linear bandit setting, the dynamic regret of the* `SW-UCB` *algorithm is upper bounded as* $\mathscr{R}_T(\texttt{SW-UCB algorithm}) = \widetilde{O}(wB_T + dT/\sqrt{w})$. *When* B_T (> 0) *is known, by taking* $w = \Theta\left((dT)^{2/3} B_T^{-2/3}\right)$, *the dynamic regret of the* `SW-UCB` *algorithm is* $\mathscr{R}_T(\texttt{SW-UCB algorithm}) = \widetilde{O}\left(d^{2/3} B_T^{1/3} T^{2/3}\right)$. *When* B_T *is unknown, by taking* $w = \Theta\left((dT)^{2/3}\right)$, *the dynamic regret of the* `SW-UCB` *algorithm is*

$$\mathscr{R}_T(\texttt{SW-UCB algorithm}) = \widetilde{O}\left(d^{2/3} B_T T^{2/3}\right).$$

Proof Sketch. The proof of this theorem is similar to that of Theorem 3, and is thus omitted. The only difference is that we need to include the regret contributed by selecting actions $a_1, \ldots, a_d$ every w rounds. But these sums to $\widetilde{O}(dT/w)$, which is dominated by the term $\widetilde{O}(dT/\sqrt{w})$. Hence the dynamic regret bounds can be obtained similarly as the linear bandit setting. □

We can now implement the `BOB` algorithm with the same set of parameters as eq. (2.12), except that Q is set to $H \cdot Y_{\max}$, *i.e.*,

$$H = \left\lfloor (dT)^{\frac{1}{2}} \right\rfloor, \Delta = \lceil \ln H \rceil, J = \left\{ H^0, \left\lfloor H^{\frac{1}{\Delta}} \right\rfloor, \ldots, H \right\}, Q = 2H \cdot Y_{\max}. \tag{2.27}$$

This is because the total rewards of each block is deterministically bounded by $[-H \cdot Y_{\max}, H \cdot Y_{\max}]$. The dynamic regret bound when B_T is unknown thus follows.

Theorem 12. *The dynamic regret bound of the* `BOB` *algorithm for the drifting generalized linear bandit setting is* $\mathscr{R}_T$ (`BOB` *algorithm*) $= \widetilde{O}\left(d^{2/3}B_T^{1/3}T^{2/3} + d^{1/2}T^{3/4}\right)$.

The proof of the theorem is similar to Theorem 6's, and it is thus omitted.

2.6.4 Combinatorial Semi-Bandits

Finally, we consider the drifting combinatorial semi-bandit problem. For ease of presentation, we use $X(i)$ to denote the i^{th} coordinate of a vector X. Following the setup in Kveton et al. [123], an instance of combinatorial semi-bandit is represented by the tuple $(E, \mathscr{E}, \{P_t\}_{t=1}^T)$, where the ground set E consist of d items, and $\mathscr{E}$ is a family of indicator vectors of subsets of E. Each P_t is a latent distribution on the reward vector $W_t = (W_t(1), \dots W_t(d))$ on each and every item $i \in E$ in round $t \in [T]$. The DM only knows that $W_t(i)$ belongs to $[0,1]$ for each $i \in [d]$ and $t \in [T]$, but she does not know $\theta_t(i) = \mathbb{E}[W_t(i)]$ for any $i \in [d]$ and $t \in [T]$. We can thus know from Lemma 1.8 of Rigollet and Hütter [154] that $W_t(i) - \theta_t(i)$ is $R = 1/2$ sub-Gaussian for all $t \in [T]$ and $i \in [d]$. The sequence $\{P_t\}_{t=1}^T$ are generated by an oblivious adversary before the online process begins.

In each round t, a reward vector W_t is sampled according to the latent distribution P_t. Then, the DM pulls an action $X_t \in \mathscr{E}_t$, and earns a reward $Y_t = \langle X_t, W_t \rangle = \sum_{i \in E} X_t(i) W_t(i)$ that corresponds to the items indicated by X_t. Under the semi-bandit feedback model, the DM observes the realized rewards $\{W_t(i) : X_t(i) = 1\}$ for the indicated items, but she does not observe $W_t(i)$ for $X_t(i) = 0$. The DM desires to minimize the dynamic regret $\mathbb{E}\left[\sum_{t=1}^T \max_{x_t^* \in \mathscr{E}} \langle x_t^* - X_t, \theta_t \rangle\right]$. Similar to the d-armed bandit setting, we define the variation budget B_T with the infinity norm: $\sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|_\infty \le B_T$. For the subsequent discussion, we denote $m = \max_{X \in \mathscr{E}} \sum_{i \in E} X(i)$ as the maximum arm size of the underlying instance.

We first show a lower bound for this setting.

Theorem 13. *Let* (d, m, T, B_T) *be a tuple that satisfies inequalities* $d \ge 2m \ge 2$, $T \ge 1$, $m/d \le B_T \le Tm/d$. *For any non-anticipatory policy, there exists a drifting combinatorial bandit instance* $(E, \mathscr{E}, \{P_t\}_{t=1}^T)$, *with* d *items, maximum arm size* m, *and variation budget* B_T *such that the dynamic regret in* T *rounds is* $\Omega(d^{1/3}m^{2/3}B_T^{1/3}T^{2/3})$.

The complete proof is presented in Section A.9 of the appendix. For a window size w, we define $N_{t-1}(i)$ as the number of times that coordinate i of the chosen action is set to 1 within rounds $(t-w),\dots,(t-1)$, *i.e.*, for all $i\in[d]$, $N_{t-1}(i)=\sum_{s=1\vee(t-w)}^{t-1}\mathbf{1}[X_s(i)=1]$. Here $\mathbf{1}[\cdot]$ is the indicator function. In each round t, the DM also maintains the sliding-window estimates for each coordinate $i\in[d]$ of θ_t:

$$\hat{\theta}_t(i)=\frac{\sum_{s=1\vee(t-w)}^{t-1}W_s(i)\cdot\mathbf{1}[X_s(i)=1]}{\max\{N_{i,t-1},1\}}.$$

Thanks to the semi-bandit feedback, the outcome $W_s(i)$ is observed when $X_s(i)=1$, so $\hat{\theta}_{t,i}$ can be constructed from the observations in the previous w rounds. We can thus reuse the Theorem 7 derived for the d-armed bandit case:

Theorem 14. *For all $t\in[T]$ and all $i\in[d]$, we have with probability at least $1-1/T$,*

$$\left|\hat{\theta}_t(i)-\theta_t(i)\right|\le\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_\infty+4R\sqrt{\frac{\ln(2dT^2)}{N_{t-1}(i)+1}},$$

holds for all $x\in D_t$.

The complete proof is presented in Section A.10. Following the rationale of UCB algorithm for stochastic combinatorial semi-bandit [123] as well as that of Section 2.4, we consider the `SW-UCB` algorithm which selects a combinatorial action X_t with highest UCB in each round t, *i.e.*,

$$\max_{X\in\mathscr{E}_t}\left\{\sum_{i\in E}X(i)\cdot\left[\hat{\theta}_{t,i}+4R\sqrt{\frac{\ln(2dT^2)}{N_{t-1}(i)+1}}\right]\right\}.$$

Denoting $m:=\max_{t\in[T],X\in\mathscr{E}_t}\|X\|_1$, we can now arrive at the following regret upper bound.

Theorem 15. *For any window size $w\ge d/m$, the dynamic regret of the `SW-UCB` algorithm for the drifting combinatorial semi-bandit setting is upper bounded as $\mathscr{R}_T$ (`SW-UCB` algorithm) $=\widetilde{O}\left(wmB_T+\sqrt{dm}T/\sqrt{w}\right)$. When $B_T<mT/d$, is known, by taking $w=\Theta\left(d^{1/3}m^{-1/3}T^{2/3}B_T^{-2/3}\right)$, the dynamic regret of the `SW-UCB` algorithm is $\mathscr{R}_T$ (`SW-UCB` algorithm) $=\widetilde{O}\left(d^{1/3}m^{2/3}B_T^{1/3}T^{2/3}\right)$. When B_T is unknown, by taking $w=\Theta\left(d^{1/3}m^{-1/3}T^{2/3}\right)$, the dynamic regret of the `SW-UCB` algorithm*

is

$$\mathscr{R}_T\left(\texttt{SW-UCB}\ algorithm\right) = \widetilde{O}\left(d^{1/3}m^{2/3}B_T T^{2/3}\right).$$

The complete proof is presented in Section A.11 of the appendix. When B_T is unknown, we can implement the BOB algorithm with the following parameters:

$$H = \left\lfloor (dT)^{\frac{1}{2}} m^{-\frac{1}{2}} \right\rfloor, \Delta = \lceil \ln H \rceil, J = \left\{ H^0, \left\lfloor H^{\frac{1}{\Delta}} \right\rfloor, \ldots, H \right\}, Q = 2H \cdot m \tag{2.28}$$

This is because the total rewards of each block is deterministically bounded by $[-H \cdot m, H \cdot m]$. The dynamic regret bound of the BOB algorithm for the combinatorial semi-bandit setting is characterized as follows.

Theorem 16. *The dynamic regret of the* BOB *algorithm for the drifting combinatorial semi-bandit setting is* $\mathscr{R}_T\left(\texttt{BOB}\ algorithm\right) = \widetilde{O}\left(d^{1/3}m^{2/3}B_T^{1/3}T^{2/3} + d^{1/4}m^{3/4}T^{3/4}\right)$.

The complete proof is presented in Section A.12.

2.7 Numerical Experiments

As a complement to our theoretical results, we conduct numerical experiments on synthetic datasets and the CPRM-12-001: On-Line Auto Lending dataset provided by the Center for Pricing and Revenue Management at Columbia University to compare the dynamic regret performances of the SW-UCB algorithm and the BOB algorithm with several existing non-stationary bandit algorithms.

2.7.1 Experiments on Synthetic Dataset

For synthetic dataset, in Section 2.7.1, we first evaluate the growth of dynamic regret when T increases. We follow the setup of [38] for fair comparisons. Then, in Section 2.7.1, we fix $T = 10^5$, and evaluate the behavior of the algorithms across rounds.

The Trend of Dynamic Regret with Varying T

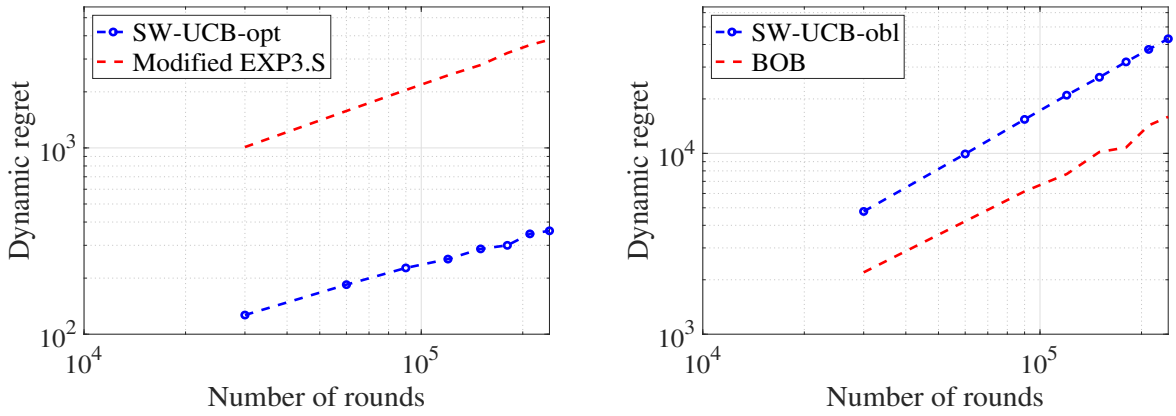
We consider a 2-armed bandit setting, and we vary T from 3×10^4 to 2.4×10^5 with a step size of 3×10^4. We set θ_t to be the following sinusoidal process, *i.e.*, $\forall t \in [T]$, $\theta_t = \left(0.5 + 0.3\sin\left(5B_T\pi t/T\right), 0.5 + 0.3\sin\left(\pi + 5B_T\pi t/T\right)\right)^\top$. The total variation of the θ_t's across the whole time horizon is upper bounded by $\sqrt{2}B_T$. We also use i.i.d. normal distribution with $R = 0.1$ for the noise terms.

Known Constant Variation Budget. We start from the known constant variation budget case, *i.e.*, $B_T = 1$, to measure the regret growth of the two optimal algorithms, *i.e.*, the optimally tuned (*i.e.*, knowing B_T) `SW-UCB` algorithm and the modified EXP3.S algorithm [37], with respect to the total number of rounds. The log-log plot is shown in Fig. 2-2(a). From the plot, we can see that the regret of `SW-UCB` algorithm is only about 20% of the regret of EXP3.S algorithm.

Unknown Time-Dependent Variation Budget. We then turn to the more realistic time-dependent variation budget case, *i.e.*, $B_T = T^{1/3}$. As the modified EXP3.S algorithm does not apply to this setting, we compare the performances of the obliviously tuned (*i.e.*, not knowing B_T) `SW-UCB` algorithm and the `BOB` algorithm. The log-log plot is shown in Fig. 2-2(b). From the results, we verify that the slope of the regret growth of both algorithms roughly match the established results, and the regret of `BOB` algorithm's is much smaller than that of the `SW-UCB` algorithm's.

A Further Study on the Algorithms' Behavior

We provide additional numerical evaluation, by considering *piecewise linear instances*, where the reward vector $\theta_t \in \mathbb{R}^d$ is a randomly generated piecewise linear function of t. To generate such an instance, we first set $T = 10^5$, and then we randomly sample 30 time points in $\tau_1, \tau_2, \ldots, \tau_{30} \in \{2, \ldots, T-1\}$ without replacement. We further denote $\tau_0 = 1, \tau_{31} = T$. After that, we randomly sample 32 random unit length vectors $v_0, \ldots, v_{31} \in \mathbb{R}^d$. Finally, for each $t \in [T]$, we define θ_t as the linear interpolation between v_s, v_{s+1}, where $\tau_s \leq t\tau_{s+1}$.



(a) Log-log plot for known $B_T = O(1)$.

(b) Log-log plot for unknown $B_T = O(T^{1/3})$.

Figure 2-2: Results for gradually change environment with 2 arms

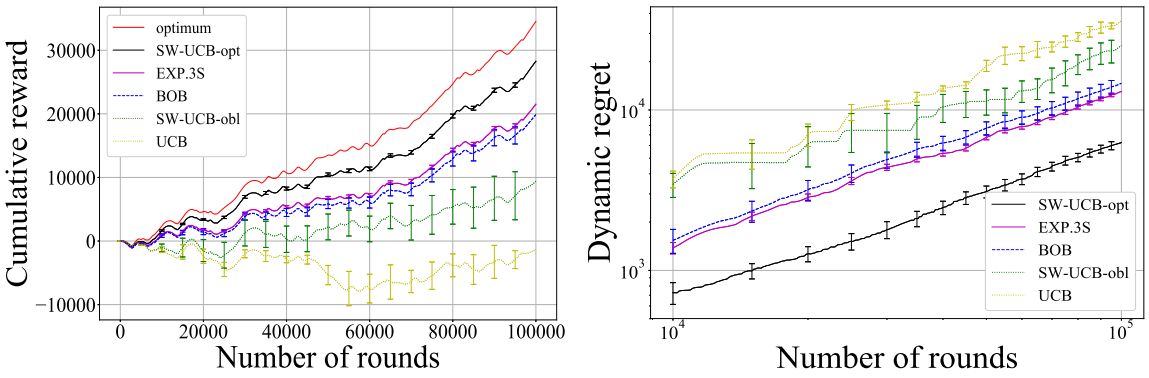
More precisely, we have $\theta_t = ((\tau_{s+1} - t)v_s + (t - \tau_s)v_{s+1})/(\tau_{s+1} - \tau_s)$. Note that the random reward in each period can be negative.

In what follows, we first evaluate the performance of the algorithms by [38] as well as our algorithms in a 2-armed bandit piece-wise linear instance. Then, we evaluate the performance of our algorithms in a linear bandit piece-wise linear instance, where $d = 5$, and each D_t is a random subset of 40 unit length vectors in $\mathbb{R}^d$. We do not evaluate the algorithms by [38] in the second instance, since the algorithms by [38] are only designed for the non-stationary K-armed bandit setting. For each instance, each algorithm is evaluated 50 times.

Two armed bandits. We first evaluate the performance of the modified EXP.3S in [38] as well as the performance of the `SW-UCB` algorithm, `BOB` algorithmin a randomly generated 2-armed bandit instance. Fig 2-3(a) illustrates the average cumulative reward earned by each algorithm in the 50 trials, and Fig 2-3(b) depicts the average dynamic regret incurred by each algorithm in the 50 trials. In Figs 2-3(a), 2-3(b), shorthand SW-UCB-opt is the `SW-UCB` algorithm, where B_T is known and $w = w^{\text{opt}}$ is set to further optimized the log factors of the dynamic regret bound (see Appendix A.13 for the expression of w^{opt}). Shorthand EXP3.S stands for the modified EXP3.S algorithm by [38], where B_T is known and the window size is set to optimized the dynamic regret bound. Shorthand BOB stands

for the BOB algorithm. Shorthand SW-UCB-obl is the SW-UCB algorithm, where B_T is not known, and $w = w^{\text{obl}}$ is obliviously set (see Appendix A.13 for the expression of w^{obl}). Finally, shorthand UCB stands for the UCB algorithm by [3], which is applicable to the stationary K-armed bandit problem. Note that B_T is known to SW-UCB-opt, EXP3.S, but not to BOB, SW-UCB-obl, UCB.

Overall, we observe that SW-UCB-opt is the better performing algorithm when B_T is known, and BOB is the best performing when B_T is not known. It is evident from Fig 2-3(a) that SW-UCB-opt, EXP3.S and BOB are able to adapt to the change in the reward vector θ_t across time t. We remark that BOB, which does not know B_T, achieves a comparable amount of cumulative reward to EXP3.S, which does know B_T, across time. It is also interesting to note that UCB, which is designed for the stationary setting, fails to converge (or even to achieve a non-negative total reward) in the long run, signifying the need of an adaptive UCB algorithm in a non-stationary setting.



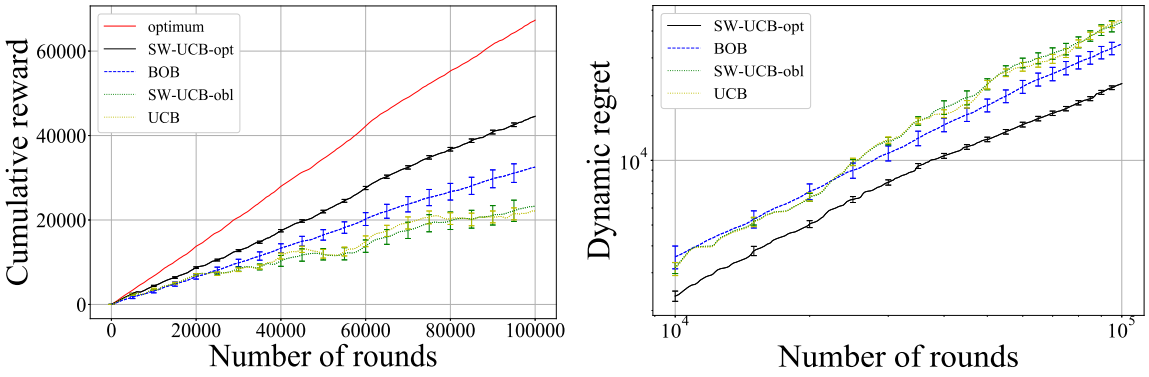
(a) Cumulative reward

(b) Dynamic regret

Figure 2-3: Results for piecewise linear environment with 2 arms

Linear bandits. Next, we move to the linear bandit case, and we consider the performance of SW-UCB-opt, SW-UCB-obl, BOB and UCB, as illustrated in Figs 2-4(a), 2-4(b). While the performance of the algorithms ranks similarly to the previous 2-armed bandit case, we witness that UCB, which is designed for the stationary setting, has a much better performance in the current case than the 2-armed case. We surmise that the relatively larger size of the action space D_t here allows UCB to choose an action that performs well even

when the reward vector is changing.



(a) Cumulative reward

(b) Dynamic Regret

Figure 2-4: Results for piecewise linear environment with linear action set.

2.7.2 Experiments on Online Auto-Lending Dataset

We now conduct experiments on the on-line auto lending dataset, which was first studied by [145], and subsequently used to evaluate dynamic pricing algorithms by [27]. The dataset records all auto loan applications received by a major online lender in the United States from July 2002 through November 2004. For each piece of data, it presents the borrower's feature (*e.g.*, date of an application, the term and amount of loan requested, and some personal information), the lender's decision (*e.g.*, the monthly payment for the borrower), and whether or not this offer is accepted by the borrower. Please refer to Columbia University Center for Pricing and Revenue Management [64] for a detailed description of the dataset.

Similar to [27], we use the first $T = 5 \times 10^4$ arrivals that span 276 days for this experiment. We adopt the commonly used [127, 40] linear model to interpolate the response of each customer: for the t^{th} customer with feature x_t, if price p_t is offered, she accepts the offer with "probability" $\langle \theta_t, [x_t; p_t x_t] \rangle$. Although the customers' responses are binary, *i.e.*, whether or not she accepts the loan, [40] theoretically justified that the revenue loss caused by using this misspecified model is negligible. For the changing environment, we assume that the θ_t's remain stationary in a single day period, but can change across days. We also use the feature selection results in [27] to pick FICO score, the term of contract, the loan

amount approved, prime rate, the type of car, and the competitor's rate as the feature vector for each customer.

As a first step, we recover the unknown θ_t's from the dataset with linear regression method. But since the lender's decisions, *i.e.*, the price for each customer, is not presented in the dataset, 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, please refer to the cited references for more details). The resulted B_T is 1.9×10^2 $\left(\approx T^{0.48}\right)$, which means we are in the moderately non-stationary environment. Since the maximum of the imputed prices is ≈ 400, the range of price in our experiment is thus set to $[0, 500]$ with a step size of 10.

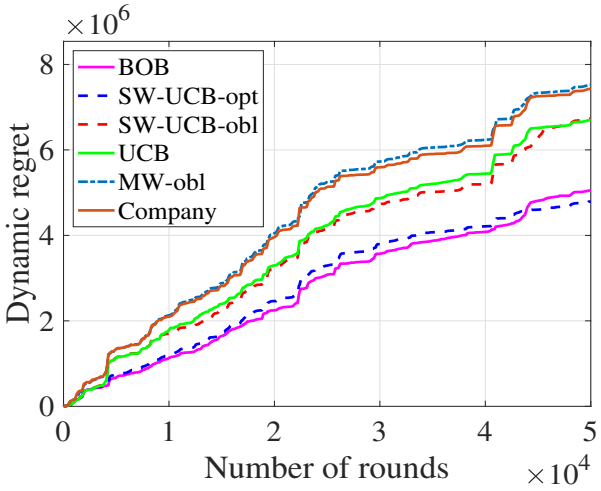


Figure 2-5: Results for the on-line auto lending dataset.

We then run the experiment with the recovered parameters, and measure the dynamic regrets of the `SW-UCB` algorithm (known B_T and unknown B_T), the `BOB` algorithm, the UCB algorithm, the Moving Window (MW) algorithm [115] without knowing B_T, as well as the company's original decisions. Here, we note that the MW algorithm does not permit customer features, and hence its dynamic regret should scale linearly in T. The results are shown in Fig. 4-4. The plot shows that the `SW-UCB` algorithm with known B_T (SW-UCB-opt) and the `BOB` algorithm have the lowest dynamic regrets. Besides, the dynamic regret of the parameter-free `BOB` algorithm is $\geq 24\%$ less than those of the obliviously tuned `SW-UCB` algorithm (SW-UCB-obl) and the UCB algorithm. It also saves $\geq 32\%$ dynamic

regret when compared to the MW algorithm and the company's original decisions. The results clearly indicate that the SW-UCB algorithm and the BOB algorithm can deal with the drift while the UCB algorithm fails to keep track of the dynamic environment. More importantly, the results validate our theoretical findings regarding the parameter-free adaptation of the BOB algorithm.

Chapter 3

Non-Stationary Reinforcement Learning

3.1 Problem Formulation

In this section, we introduce the notations to be used and introduce the learning protocol.

3.1.1 Notation

Throughout this chapter, all vectors are column vectors, unless specified otherwise. We define $[n]$ to be the set $\{1,2,\ldots,n\}$ for any positive integer n. We denote $\mathbf{1}[\cdot]$ as the indicator function. For $p \in [1,\infty]$, we use $\|x\|_p$ to denote the p-norm of a vector $x \in \mathbb{R}^d$. We denote $x \vee y$ and $x \wedge y$ as the maximum and minimum between $x, y \in \mathbb{R}$, respectively. We adopt the asymptotic notations $O(\cdot), \Omega(\cdot)$, and $\Theta(\cdot)$ [65]. When logarithmic factors are omitted, we use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$, respectively. With some abuse, these notations are used when we try to avoid the clutter of writing out constants explicitly.

3.1.2 Learning Protocol

Model Primitives: An instance of non-stationary MDP is specified by the tuple $(\mathscr{S}, \mathscr{A}, T, r, p)$. The set $\mathscr{S}$ is a finite set of states. The collection $\mathscr{A} = \{\mathscr{A}_s\}_{s\in\mathscr{S}}$ contains a finite action set $\mathscr{A}_s$ for each state $s \in \mathscr{S}$. We say that (s,a) is a state-action pair if $s \in \mathscr{S}, a \in \mathscr{A}_s$. We denote $S = |\mathscr{S}|$, $A = (\sum_{s\in\mathscr{S}} |\mathscr{A}_s|)/S$. We denote T as the total number of time steps, and $r = \{r_t\}_{t=1}^T$ as the sequence of mean rewards. For each t, we have $r_t = \{r_t(s,a)\}_{s\in\mathscr{S}, a\in\mathscr{A}_s}$,

and $r_t(s,a) \in [0,1]$ for each state-action pair (s,a). In addition, we denote $p = \{p_t\}_{t=1}^T$ as the sequence of state transition distributions. For each t, we have $p_t = \{p_t(\cdot|s,a)\}_{s\in\mathscr{S},a\in\mathscr{A}_s}$, where $p_t(\cdot|s,a)$ is a probability distribution over $\mathscr{S}$ for each state-action pair (s,a).

Non-Stationarity: The quantities r_t's and p_t's vary across different t's in general. Following [36], we quantify the variations on r_t's and p_t's in terms of their respective *variation budgets* B_r, B_p (> 0):

$$\begin{aligned} B_r &= \sum_{t=1}^{T-1} B_{r,t}, \text{ where } B_{r,t} = \max_{s\in\mathscr{S},a\in\mathscr{A}_s} |r_{t+1}(s,a) - r_t(s,a)|, \\ B_p &= \sum_{t=1}^{T-1} B_{p,t}, \text{ where } B_{p,t} = \max_{s\in\mathscr{S},a\in\mathscr{A}_s} \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1. \end{aligned} \tag{3.1}$$

We emphasize although B_r and B_p might be used as inputs by the DM, individual $B_{r,t}$'s and $B_{p,t}$'s are unknown to the DM throughout.

Remark 8 (Definition of Variation Budgets). *For brevity of exposition, we choose to define the variation budgets (see eqn. (3.1)) for reward and state transition distributions with the infinity norm and 1-norm, respectively. One can also define them with respect to other commonly used metrics, such as the 2-norm [59], and the this would only affect the dependence on S and A for the established dynamic regret bounds in the subsequent sections.*

Model Dynamics: The DM faces a non-stationary MDP instance $(\mathscr{S},\mathscr{A},T,r,p)$. She knows $\mathscr{S},\mathscr{A},T$, but not r,p. The DM starts at an arbitrary state $s_1 \in \mathscr{S}$. At time t, three events happen. First, the DM observes her current state s_t. Second, she takes an action $a_t \in \mathscr{A}_{s_t}$. Third, given s_t, a_t, she stochastically transitions to another state $s_{t+1} \sim p_t(\cdot|s_t,a_t)$, and receives a stochastic reward $R_t(s_t,a_t)$, which is 1-sub-Gaussian with mean $r_t(s_t,a_t)$. In the second event, the choice of a_t is based on a *non-anticipatory* policy Π_t. That is, the choice only depends on the current state s_t and the previous observations $\mathscr{H}_{t-1} := \{s_q, a_q, R_q(s_q,a_q)\}_{q=1}^{t-1}$. We denote $\Pi = \{\Pi_1,\ldots,\Pi_T\}$ as the policy of the DM throughout the entire time horizon.

Dynamic Regret: The DM aims to maximize the cumulative expected reward $\mathbb{E}[\sum_{t=1}^T r_t(s_t,a_t)]$. To measure the her performance, we consider an equivalent objective of minimizing the *dy-*

namic regret [36, 108]. Formally, the dynamic regret of a policy Π is defined as

$$\text{Dyn-Reg}_T(\Pi) = \sum_{t=1}^{T} \left\{ \mathbb{E}[r_t(s_t^{\Pi^*}, a_t^{\Pi^*}) - \mathbb{E}[r_t(s_t^{\Pi}, a_t^{\Pi})] \right\}, \tag{3.2}$$

which is the difference between the optimal policy Π^* (that knows all the r and p) and Π in terms of cumulative expected reward. Here, we denote $\{(s_t^{\Pi^*}, a_t^{\Pi^*})\}_{t=1}^T$ and $\{(s_t^{\Pi}, a_t^{\Pi})\}_{t=1}^T$ as the trajectory under policy Π^* and Π, respectively.

Alternative Oracle: Note that in the optimal policy, the expected cumulative reward are intertwined due to endogenous dynamics, *i.e.*, $s_{t+1}^{\Pi} \sim p_t(\cdot|s_t^{\Pi}, a_t^{\Pi})$. To ease our analysis, we introduce an intermediate oracle $\sum_{t=1}^T \rho_t^*$, where the summand ρ_t^* is the optimal long-term average reward of the stationary MDP with state transition distribution p_t and mean reward r_t. The optimum ρ_t^* can be computed by solving linear program (B.1) provided in Section B.1.1 of the appendix. Since the quantity $\sum_{t=1}^T \rho_t^*$ can be decomposed to summations across different intervals, it is more convenient for analysis than the expected cumulative reward of the optimal policy. We point out that the same oracle is used for RL in piecewise-stationary MDPs [108]. To understand the difference between the two oracles (and to ensure learnability), we begin by reviewing the concept of *diameter* of a MDP.

Definition 17 (Communicating MDPs and Diameter [108]). *Consider a set of states $\mathscr{S}$, a collection $\mathscr{A} = \{\mathscr{A}_s\}_{s\in\mathscr{S}}$ of action sets, and a state transition distribution $\bar{p} = \{\bar{p}(\cdot|s,a)\}_{s\in\mathscr{S}, a\in\mathscr{A}_s}$. For any $s, s' \in \mathscr{S}$ and stationary policy π, the hitting time from s to s' under π is the random variable $\Lambda(s'|\pi, s) := \min\{t : s_{t+1} = s', s_1 = s, s_{\tau+1} \sim \bar{p}(\cdot|s_\tau, \pi(s_\tau))\ \forall \tau\}$, which can be infinite. We say that $(\mathscr{S}, \mathscr{A}, \bar{p})$ is a* communicating MDP *iff*

$$D := \max_{s,s'\in\mathscr{S}} \min_{\text{stationary } \pi} \mathbb{E}\left[\Lambda(s'|\pi, s)\right]$$

is finite. The quantity D is the diameter *associated with $(\mathscr{S}, \mathscr{A}, \bar{p})$.*

As shown in [108], the concept of diameter plays a fundamental role in characterizing the complexity of RL in MDPs because it captures the "hardness" of transitioning between states in this MDP.

Remark 9 (Diameter and RL in MDPs). *In order to make informative decisions, the DM*

has to have accurate estimates of the quantities $r_t(s,a)$'s and $p_t(\cdot|s,a)$.'s. In other words, she must visit every state $s \in \mathscr{S}$ and choose each of its available actions $a \in \mathscr{A}_s$ frequently enough to collect relevant samples. Consequently, the harder to transition from a state s to another state s' (i.e., the diameter is larger), the more the DM would pay for learning.

With the above remark, we assume that the diameter induced by each $(\mathscr{S}, \mathscr{A}, p_t)$ is bounded to enable learning.

Assumption 5 (Bounded Diameters)**.** *For each $t \in [T]$, the tuple $(\mathscr{S}, \mathscr{A}, p_t)$ constitutes a communicating MDP with diameter at most D_t. We denote the* maximum diameter *as $D_{max} = \max_{t\in\{1,\dots,T\}} D_t$. Here, $D_{\max}$ is unknown ahead.*

With these, the following proposition upper bounds the difference between the optimal expected cumulative reward and $\sum_{t=1}^{T} \rho_t^*$.

Proposition 18. *Consider an instance $(\mathscr{S}, \mathscr{A}, T, p, r)$ that satisfies Assumption 5 with maximum diameter D_{max}, and has variation budgets B_r, B_p for the rewards and transition distributions respectively. Then, it holds*

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t^{\Pi^*}, a_t^{\Pi^*})\right] - \sum_{t=1}^{T} \rho_t^* \leq 4\sqrt{D_{\max}(B_r + 2D_{max}B_p)T} + (B_r + 2B_p).$$

The Proposition is proved in section B.1.2 of the appendix. With Proposition 18, we can focus on comparing the performance of the `SWUCRL2-CW` algorithm against the quantity $\sum_{t=1}^{T} \rho_t^*$.

Lower Bound: Before proceeding, we also characterize the minimax lower bound of our problem to understand the limit of this setting.

Proposition 19. *For any natural numbers $S, A \geq 10$, $D_{\max} \geq \log_A S$, $T \geq D_{\max} SA$, $B_r \in [S^{-1}A^{-1}, S^{-1}A^{-1}T]$, and $B_p \in [D^{-1/2}S^{1/2}A^{1/2}T^{-1/2}, D_{\max}^{-2}S^{-1}A^{-1}T]$, there exists an non-stationary MDP instance $(\mathscr{S}, \mathscr{A}, T, r, p)$ such that the dynamic regret of any non-anticipatory policy Π satisfies Dyn-Reg$_T(\Pi) = \Omega(D_{\max}^{2/3} B_p^{1/3} S^{1/3} A^{1/3} T^{2/3} + B_r^{1/3} S^{1/3} A^{1/3} T^{2/3})$.*

The proof is provided in Section B.2 of the appendix.

3.2 Related Works

3.2.1 RL in Stationary MDPs

RL in stationary (discounted and un-discounted reward) MDPs has been widely studied in [48, 29, 108, 11, 85, 87, 167, 166, 172, 187, 86, 175]. For the discounted reward setting, [167, 172, 166] proposed (nearly) optimal algorithms in terms of sample complexity. For the un-discounted reward setting, [108] established a minimax lower bound $\Omega(\sqrt{D_{\max}SAT})$ on the regret when both the reward and state transition distributions are time-invariant. They also designed the UCRL2 algorithm and showed that it attains a regret bound $\tilde{O}(D_{\max}S\sqrt{AT})$. [86] proposed the UCRL2B algorithm, which is an improved version of the UCRL2 algorithm. The regret bound of the UCRL2B algorithm is $\tilde{O}(S\sqrt{D_{\max}AT}+D_{\max}^2S^2A)$. The minimax optimal algorithm is provided in [187] although it is not computationally efficient.

3.2.2 RL in Non-Stationary MDPs

In a parallel work [144], the authors considered a similar setting to ours by applying the "forgetting principle" from non-stationary bandit settings [92, 58] to design a learning algorithm. To achieve its dynamic regret bound, the algorithm by [144] partitions the entire time horizon $[T]$ into time intervals $\mathscr{I}=\{I_k\}_{k=1}^K$, and crucially requires the access to $\sum_{t=\min I_k}^{\max I_k-1}B_{r,t}$ and $\sum_{t=\min I_k}^{\max I_k-1}B_{p,t}$, *i.e.*, the variations in both reward and state transition distributions of each interval $I_k\in\mathscr{I}$ (see Theorem 3 in [144]). In contrast, the `SWUCRL2-CW` algorithm and the `BORL` algorithm require significantly less information on the variations. Specifically, the `SWUCRL2-CW` algorithm does not need any additional knowledge on the variations except for B_r and B_p, *i.e.*, the variation budgets over the entire time horizon as defined in eqn. (3.1), to achieve its dynamic regret bound (see Theorem 21). This is similar to algorithms for the non-stationary bandit settings, which only require the access to B_r [36]. More importantly, the `BORL` algorithm (built upon the `SWUCRL2-CW` algorithm) enjoys the same dynamic regret bound even without knowing either of B_r or B_p (see Theorem 22).

There also exists some settings that are closely related to, but different than our set-

ting (in terms of exogeneity and feedback). [108, 90] proposed solutions for the RL in piecewise-stationary MDPs setting. But as discussed in Section 1.2, simply applying their techniques to the general RL in non-stationary MDPs may result in undesirable dynamic regret bounds (see Section 3.3.3 for more details). In [182, 139, 16, 72, 110, 51], the authors considered RL in MDPs with changing reward distributions but fixed transition distributions. [73, 181, 140, 2, 157, 130] considered RL in non-stationary MDPs with full information feedback.

3.2.3 Non-Stationary Multi-Armed Bandits (MAB)

For online learning and bandit problems where there is only one state, the works by [18, 92, 36, 115] proposed several "forgetting" strategies for different non-stationary MAB settings. More recently, the works by [113, 134, 59, 58, 57] designed parameter-free algorithms for non-stationary MAB problems. Another related but different setting is the Markovian bandit [118, 135], in which the state of the chosen action evolve according to an independent time-invariant Markov chain while the states of the remaining actions stay unchanged. In [189], the authors also considered the case when the states of all the actions are governed by the same (uncontrollable) Markov chain.

3.3 Sliding Window UCRL2 with Confidence Widening Algorithm

In this section, we first present the `SWUCRL2-CW` algorithm, which incorporates our novel confidence widening technique and sliding window estimates [92] into UCRL2 [108], and motivate this design by formally present the unique challenge of RL in non-stationary MDP.

3.3.1 Design Overview

The `SWUCRL2-CW` algorithm first specifies a pair of sliding window parameters $W \in \mathbb{N}$ and a confidence widening parameter $\eta \geq 0$. Parameter W specifies the number of previous time steps to look at when estimating the reward and state transition distributions, respec-

tively. Parameter η quantifies the amount of additional optimistic exploration, on top of the conventional optimistic exploration using upper confidence bounds. The latter turns out to be helpful for handling the temporal drifts in the state transition distributions (see Section 3.3.3) and is capable of ensuring the MDP output by the EVI has a bounded diameter most of the time.

The algorithm runs in a sequence of episodes that partitions the T time steps. Episode m starts at time $\tau(m)$ (in particular $\tau(1) = 1$), and ends at the end of time step $\tau(m+1) - 1$. Throughout an episode m, the DM follows a certain stationary policy $\tilde{\pi}_{\tau(m)}$. The DM ceases the m^{th} episode if at least one of the following two criteria is met:

- The time index t is a multiple of W. Consequently, each episode last for at most W time steps. The criterion ensures that the DM switches the stationary policy $\tilde{\pi}_{\tau(m)}$ frequently enough, in order to adapt to the exogenous dynamics.

- There exists some state-action pair (s,a) such that $\nu_{\tau(m)}(s,a)$, the number of time step t's with $(s_t,a_t) = (s,a)$ within episode m, is at least as many as the total number of counts for it within the W time steps prior to $\tau(m)$, *i.e.*, from $(\tau(m) - W) \vee 1$ to $(\tau(m) - 1)$. This is similar to the doubling criterion in [108], which ensures that each episode is sufficiently long so that the DM can focus on learning.

The combined effect of these two criteria allows the DM to learn a low dynamic regret policy with historical data from an appropriately sized time window and confidence widening parameter.

3.3.2 Policy Construction

To describe `SWUCRL2-CW` algorithm, we first define for each state-action pair (s,a) and each time t in episode m,

$$N_t(s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} \mathbf{1}((s_q,a_q) = (s,a)), \quad N_t^+(s,a) = \max\{1, N_t(s,a)\}. \tag{3.3}$$

Confidence Region for Rewards

For each state-action pair (s,a) and each time t in episode m, we consider the empirical mean estimator

$$\hat{r}_t(s,a) = \frac{1}{N_t^+(s,a)} \left(\sum_{q=(\tau(m)-W)\vee 1}^{t-1} R_q(s,a)\, \mathbf{1}(s_q = s, a_q = a) \right),$$

which serves to estimate the average reward

$$\bar{r}_t(s,a) = \frac{1}{N_t^+(s,a)} \left(\sum_{q=(\tau(m)-W)\vee 1}^{t-1} r_q(s,a) \mathbf{1}(s_q = s, a_q = a) \right).$$

The confidence region $H_{r,t} = \{H_{r,t}(s,a)\}_{s\in\mathscr{S}, a\in\mathscr{A}_s}$ is defined as

$$H_{r,t}(s,a) = \{\dot{r} \in [0,1] : |\dot{r} - \hat{r}_t(s,a)| \leq \mathsf{rad}\text{-}_{r,t}(s,a)\}, \tag{3.4}$$

with confidence radius $\mathsf{rad}\text{-}_{r,t}(s,a) = 2\sqrt{2\log(SAT/\delta)/N_t^+(s,a)}$. Here, δ is an input parameter (to be set to $1/T$ in the subsequent results).

Confidence Widening for State Transition Distributions.

For each state-action pair s,a and each time step t in episode m, we consider the empirical mean estimator

$$\hat{p}_t(s'|s,a) = \frac{1}{N_t^+(s,a)} \left(\sum_{q=(\tau(m)-W)\vee 1}^{t-1} \mathbf{1}(s_q = s, a_q = a, s_{q+1} = s') \right),$$

which serves to estimate the average transition probability

$$\bar{p}_t(s'|s,a) = \frac{1}{N_t^+(s,a)} \sum_{q=(\tau(m)-W)\vee 1}^{t-1} p_q(s'|s,a) \mathbf{1}(s_q = s, a_q = a). \tag{3.5}$$

Different from the case of estimating reward, the confidence region

$$H_{p,t}(\eta) = \{H_{p,t}(s,a;\eta)\}_{s\in\mathscr{S}, a\in\mathscr{A}_s}$$

Algorithm 3 `SWUCRL2-CW` algorithm

Input: Time horizon T, state space $\mathscr{S}$, and action space $\mathscr{A}$, window size W, confidence widening parameter $\eta, \delta \leftarrow 1/T$
Initialize $t \leftarrow 1$, initial state s_1.
for episode $m = 1, 2, \ldots$ **do**
  Set $\tau(m) \leftarrow t$, $v_{\tau(m)}(s,a) \leftarrow 0$, and $N_{\tau(m)}(s,a)$ according to eqn (3.3), for all s, a.
  Compute the confidence regions $H_{r,\tau(m)}$, $H_{p,\tau(m)}(\eta)$ according to eqns (3.4, 3.6).
  Compute a $(1/\sqrt{\tau(m)})$-optimal optimistic policy $\tilde{\pi}_{\tau(m)}$:

$$\mathsf{EVI}(H_{r,\tau(m)}, H_{p,\tau(m)}(\eta); 1/\sqrt{\tau(m)}) \to (\tilde{\pi}_{\tau(m)}, \tilde{r}_{\tau(m)}, \tilde{p}_{\tau(m)}, \tilde{\rho}_{\tau(m)}, \tilde{\gamma}_{\tau(m)}).$$

  while t is not a multiple of W and $v_m(s_t, \tilde{\pi}_{\tau(m)}(s_t)) < N^+_{\tau(m)}(s_t, \tilde{\pi}_{\tau(m)}(s_t))$ **do**
    Choose action $a_t = \tilde{\pi}_{\tau(m)}(s_t)$, observe reward $R_t(s_t, a_t)$ and the next state s_{t+1}.
    Update $v_{\tau(m)}(s_t, a_t) \leftarrow v_{\tau(m)}(s_t, a_t) + 1, t \leftarrow t + 1$.
    if $t > T$ **then**
      The algorithm is terminated.
    end if
  end while
end for

for the transition probability involves a widening parameter $\eta \geq 0$:

$$H_{p,t}(s,a;\eta) = \left\{ \dot{p} \in \Delta^{\mathscr{S}} : \|\dot{p}(\cdot|s,a) - \hat{p}_t(\cdot|s,a)\|_1 \leq \mathsf{rad}\text{-}_{p,t}(s,a) + \eta \right\}, \tag{3.6}$$

with confidence radius $\mathsf{rad}\text{-}_{p,t}(s,a) = 2\sqrt{2S\log(SAT/\delta)/N_t^+(s,a)}$. We shall provide a suitable choice of η when we discuss our main results (see Theorem 21).

Extended Value Iteration (EVI) [108].

The `SWUCRL2-CW` algorithm relies on the EVI, which solves MDPs using the OFU principle to near-optimality. We extract and rephrase a description of EVI in Section B.1.3 of the appendix. EVI inputs the confidence regions H_r, H_p for the rewards and the state transition distributions. The algorithm outputs an "optimistic MDP model", which consists of reward vector $\tilde{r}$ and state transition distribution $\tilde{p}$ under which the optimal average gain $\tilde{\rho}$ is the largest among all $\dot{r} \in H_r, \dot{p} \in H_p$:

- **Input:** Confidence regions H_r for r, H_p for p, and an error parameter $\varepsilon > 0$.

- **Output:** The returned policy $\tilde{\pi}$ and the auxiliary output $(\tilde{r}, \tilde{p}, \tilde{\rho}, \tilde{\gamma})$. In the latter, $\tilde{r}$, $\tilde{p}$, and $\tilde{\rho}$ are the selected optimistic reward vector, state transition distribution, and the corresponding long term average reward. The output $\tilde{\gamma} \in \mathbb{R}_+^{\mathscr{S}}$ is a *bias vector* [108]. For each $s \in \mathscr{S}$, the quantity $\tilde{\gamma}(s)$ is indicative of the short term reward when the DM starts at state s and follows the optimal policy. By the design of EVI, for the output $\tilde{\gamma}$, there exists $s \in \mathscr{S}$ such that $\tilde{\gamma}(s) = 0$. Altogether, we express

$$\mathrm{EVI}(H_r, H_p; \varepsilon) \to (\tilde{\pi}, \tilde{r}, \tilde{p}, \tilde{\rho}, \tilde{\gamma}).$$

Combining these components, a formal description of the `SWUCRL2-CW` algorithm is shown in Algorithm 3.

3.3.3 The Perils of Drift in Learning Markov Decision Processes

Before analyzing the performance of the `SWUCRL2-CW` algorithm, we first take a detour to provide a formal justification for why we need to widen the confidence regions. To analyze the loss due to using EVI, existing works (see *e.g.*, Section 4.3 of [108] or Section 4.1 of [86]) typically argue that there exists a state transition distribution $\mathring{p}$ in the confidence region such that the diameter of $(\mathscr{S}, \mathscr{A}, \mathring{p})$ is small (*i.e.*, $\leq D_{\max}$) and then show that the loss scales with the diameter of $(\mathscr{S}, \mathscr{A}, \mathring{p})$ (as well as other instance dependent parameters).

In the case of stationary MDPs, where $\forall\, t \in [T]\ p_t = p_0$, one can easily show that the un-widened confidence region $H_{p,t}(0)$ contains p_0 with high probability (see Section 4.2 of [108]). Leveraging the fact that the underlying MDP remains stationary between changes, this type of argument was further extended to RL in piecewise-stationary MDPs by [108],

However, simply inputting $H_{p,t}(0)$ to EVI might lead to unfavorable dynamic regret bound in general non-stationary MDPs. In the non-stationary environment where $p_{t-W}, \ldots, p_{t-1}$ are generally distinct, we show in the proposition below that the diameter of any $\tilde{p} \in H_{p,t}(0)$ can grow as large as $\Omega(\sqrt{W/\log W})$ despite each of $p_{t-W}, \ldots, p_{t-1}$'s diameters are just 1. To ease the notation, we put $t = W + 1$ without loss of generality.

Proposition 20. *There exists a sequence of non-stationary MDP transition distributions $p_1, \ldots, p_W$ such that 1) The diameter of $(\mathscr{S}, \mathscr{A}, p_n)$ is 1 for each $n \in [W]$; 2) The total*

variations in state transition distributions is $O(1)$. Nevertheless, there exists a policy such that if the DM follows this policy to act,

1. *The empirical MDP $(\mathscr{S}, \mathscr{A}, \hat{p}_{W+1})$ has diameter $\Theta(W)$*
2. *Further, for every $\tilde{p} \in H_{p,W+1}(0)$, the MDP $(\mathscr{S}, \mathscr{A}, \tilde{p})$ has diameter $\Omega(\sqrt{W/\log W})$*

The proof of Proposition 20 is provided in Section B.3 of the appendix.

3.3.4 Performance Analysis: The Blessing of More Optimism

We are now ready to analyze the performance of the `SWUCRL2-CW` algorithm assuming the knowledge of B_r, B_p to set W, η.

Theorem 21. *Assuming $S > 1$, the `SWUCRL2-CW` algorithm with window size W, confidence widening parameter $\eta > 0$, and $\delta = T^{-1}$ enjoys a dynamic regret (as defined in (3.2)) of order*

$$\tilde{O}\left(\frac{B_p W}{\eta} + B_r W + \frac{\sqrt{SA}T}{\sqrt{W}} + D_{max}\left[B_p W + \frac{S\sqrt{A}T}{\sqrt{W}} + T\eta + \frac{SAT}{W} + \sqrt{T}\right] + \sqrt{D_{\max}(B_r + 2D_{max}B_p)T}\right).$$

If we further put $W = W^* = S^{2/3}A^{1/2}T^{1/2}(B_r + B_p)^{-1/2}$ and $\eta = \eta^* := \sqrt{B_p W^* T^{-1}}$, this is $\tilde{O}\left(D_{max}(B_r + B_p)^{1/4} S^{2/3} A^{1/2} T^{3/4}\right)$.

Proof. Proof Sketch. The complete proof of Theorem 21 is provided in Section B.4 of the appendix. We begin by introducing two events $\mathscr{E}_r, \mathscr{E}_p$, which state that $\bar{r}_t$'s and $\bar{p}_t$'s lie in the respective (un-widened) confidence regions, *i.e.*,

$$\mathscr{E}_r = \{\bar{r}_t(s,a) \in H_{r,t}(s,a)\ \forall s,a,t\}, \quad \mathscr{E}_p = \{\bar{p}_t(\cdot|s,a) \in H_{p,t}(s,a;0)\ \forall s,a,t\}.$$

We prove that $\mathscr{E}_r, \mathscr{E}_p$ hold with probability at least $1-\delta$ (see Lemma 43 in Section B.4 of the appendix). Conditioned on $\mathscr{E}_r$ and $\mathscr{E}_p$, we distinguish two cases for each episode m (as shown in Fig. 3-1):

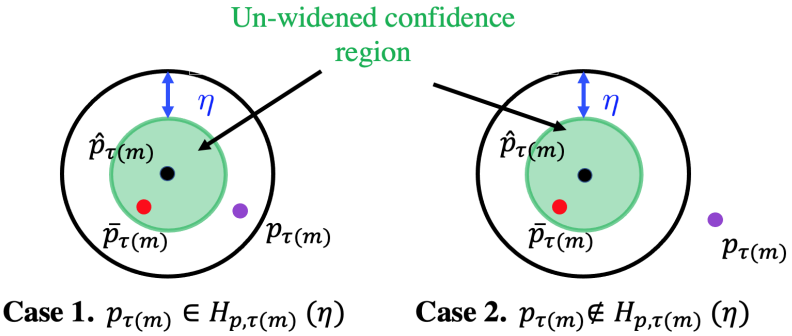
Case 1. $p_{\tau(m)} \in H_{p,\tau(m)}(\eta)$ (left panel of Fig. 3-1): In this case, we can show the difference between ρ_t^* and `SWUCRL2-CW` algorithm's reward for any time step t of episode m scales

with $D_{\max}$ (see Proposition 44 in Section B.4 of the appendix) and the price we are paying is a confidence region widened by η per time step.

Case 2. $p_{\tau(m)} \notin H_{p,\tau(m)}(\eta)$ (right panel of Fig. 3-1): In this case, note that event $\mathscr{E}_p$ assures us that $\bar{p}_{\tau(m)} \in H_{p,\tau(m)}(0)$. Hence, by virtue of confidence widening, we have that there must exists a state-action pair s,a such that $\|\bar{p}_{\tau(m)}(\cdot|s,a) - p_{\tau(m)}(\cdot|s,a)\|_1 \geq \eta$. Recall from eqn. (3.5) that $\bar{p}_{\tau(m)}(\cdot|s,a)$ is the average of several $p_{t'}(\cdot|s,a)$ for $t' \in [1 \vee (t-W), t-1]$, which implies there exists at least one $t' \in [1 \vee (t-W), t-1]$ such that $\|p_{t'}(\cdot|s,a) - p_{\tau(m)}(\cdot|s,a)\|_1 \geq \eta$. Through a triangle inequality, we have

$$\sum_{t=t'}^{\tau(m)-1} \|p_t(\cdot|s,a) - p_{t+1}(\cdot|s,a)\|_1 \geq \eta$$

In other words, the variation budget B_p is consumed by η.



Case 1. $p_{\tau(m)} \in H_{p,\tau(m)}(\eta)$ **Case 2.** $p_{\tau(m)} \notin H_{p,\tau(m)}(\eta)$

Figure 3-1: In the case of $p_{\tau(m)} \notin H_{p,\tau(m)}(\eta)$, the widened confidence regions forces an η consumption of the variation budget B_p.

□

Remark 10 (Importance of Confidence Widening). *Similar to the regret analysis of the UCRL2 algorithm (Section 4 of [108]) and the UCRL2B algorithm (Section 4 of [86]), case 1 in the proof of Theorem 21 states that, if the confidence region $H_{p,\tau(m)}(\eta)$ contains a state transition distribution with diameter at most D, then the EVI provided with $H_{p,\tau(m)}(\eta)$ returns a policy with dynamic regret bound that scales at most linearly with D during episode m. Without the widened confidence region, the smallest upper bound we can get for D is $\Omega(\sqrt{W/\log W})$ (as shown in Proposition 20), which would result in linear in T dynamic regret bound. In contrast, although the confidence widening technique cannot guarantee*

that all $H_{p,\tau(m)}(\eta)$ contains a MDP with small diameter, it enforces that the underlying environment has to consume at least η variation budget whenever this is violated (case 2) and thus makes sure that such violation could happen with limited number of times.

Remark 11 (Comparison with Conventional Optimistic Exploration). *Inspecting the prevalent OFU guided approach for stochastic MAB and RL in MDPs settings [21, 3, 108, 46, 125], one usually concludes that a tighter design of confidence region can result in a lower (dynamic) regret bound. In [6], this insights has been formalized in stochastic K-armed bandit settings via a potential function type argument. Nevertheless, Proposition 20 (together with Theorem 21) demonstrates that using the tightest confidence region may not be enough to ensure low dynamic regret bound for RL in non-stationary MDPs. This demonstrates the critical difference between RL in non-stationary MDPs and prior settings.*

Remark 12 (Improved Dynamic Regret Bound). *In [176], the authors extends our fixed confidence widening technique to an adaptive confidence widening schedule that tunes η adaptively based on historical observations to achieve an improved upper bound of order $O(D_{\max}(B_p+B_r)^{1/3}S^{2/3}A^{1/3}T^{2/3}+D_{\max}SA^{1/2}T^{1/2})$ when both B_p and B_r are known ahead.*

3.4 Bandit-over-Reinforcement Learning Algorithm: Towards Parameter-Free

Similar to [59, 58], if B_p, B_r are not known, we can set W and η obliviously as $W = S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{1}{2}}$, $\eta = \sqrt{W/T} = S^{\frac{2}{3}}A^{\frac{1}{2}}T^{-\frac{1}{2}}$ to obtain a dynamic regret bound

$$\tilde{O}\left(D_{\max}(B_r+B_p+1)S^{2/3}A^{1/2}T^{3/4}\right)$$

for the `SWUCRL2-CW` algorithm. This means, in the case of unknown B_r and B_p, the dynamic regret of `SWUCRL2-CW` algorithm scales linearly in B_r and B_p. However, by Theorem 21, we are assured a fixed pair of parameters (W^*, η^*) can ensure low dynamic regret. For the bandit setting, [58, 59] propose the Bandit-over-Bandit (BOB) framework that

uses a separate copy of EXP3 algorithm to tune the window size. Inspired by it, we develop a novel Bandit-over-Reinforcement Learning (BORL) algorithm with parameter-free $\tilde{O}\left(D_{\max}(B_r+B_p+1)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$ dynamic regret here.

3.4.1 Design Overview

Similar to the BOB framework developed in [58], we make use of the SWUCRL2-CW algorithm as a sub-routine, and "hedge" [46] against the (possibly adversarial) changes of r_t's and p_t's to identify a reasonable fixed window size and confidence widening parameter.

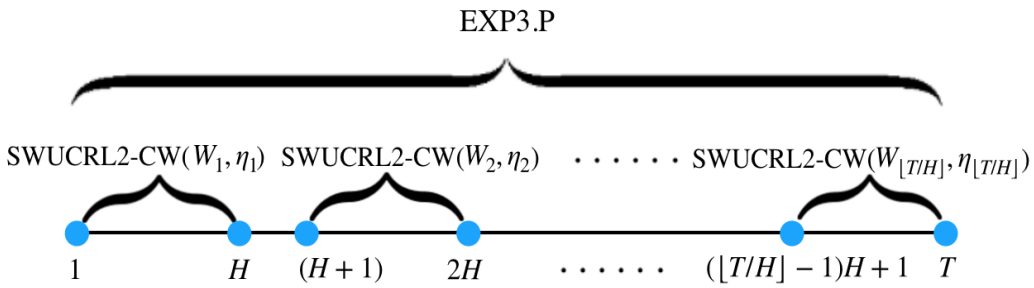


Figure 3-2: Structure of the BORL algorithm

As illustrated in Fig. 3-2, the BORL algorithm divides the whole time horizon into $\lceil T/H \rceil$ blocks of equal length H rounds (the length of the last block can $\leq H$), and specifies a set J from which each pair of (window size, confidence widening) parameter are drawn from. For each block $i \in [\lceil T/H \rceil]$, the BORL algorithm first calls the master algorithm to select a pair of (window size, confidence widening) parameters (W_i, η_i) $(\in J)$, and restarts the SWUCRL2-CW algorithm with the selected parameters as a sub-routine to choose actions for this block. Afterwards, the total reward of block i is fed back to the master algorithm, and the "posterior" of these parameters are updated accordingly. Here, we use the EXP3.P algorithm (see Section 3.2 of [46]), which is an adversarial bandit algorithm against *adaptive* adversaries, as the master algorithm.

Remark 13 (Comparison with the Bandit-over-Bandit Framework). *Even though the BORL algorithm is heavily inspired by the BOB algorithm [58], its master algorithm is critically different than the BOB algorithm's. In the BOB algorithm, the EXP3 algorithm*

[18] against oblivious *adversaries is used as the master algorithm (while the* `BORL` *algorithm uses the EXP3.P algorithm against* adaptive *adversaries). This is because the starting state of each block is determined by previous actions of the DM. Hence, the master algorithm is not facing a simple oblivious environment as the case of MAB [59], and we cannot use the EXP3 [18] algorithm as the master. Fortunately, the starting state of each block is observed before this block begins. Therefore, the regret guarantee of the EXP3.P algorithm can be leveraged here.*

3.4.2 Design Details

Algorithm 4 `BORL` algorithm

Input: Time horizon T, state space $\mathscr{S}$, and action space $\mathscr{A}$, initial state s_1.
Initialize $H, \Phi, \Delta_W, \Delta_\eta, \Delta, J_W, J_\eta$ according to eqn. (3.7), and α, β, γ according to eqn. (3.8).
$M \leftarrow \{(j',k') : j' \in \{0,1,\ldots,\Delta_W\}, k' \in \{0,1,\ldots,\Delta_\eta\}\}, q_{(j,k),1} \leftarrow 0 \; \forall (j,k) \in M$.
for $i = 1,2,\ldots,\lceil T/H \rceil$ **do**
    Define distribution $(u_{(j,k),i})_{(j,k)\in M}$ according to eqn. (3.9), and set $(j_i,k_i) \leftarrow (j,k)$ with probability $u_{(j,k),i}$.
    $W_i \leftarrow \left\lfloor H^{j_i/\Delta_W} \right\rfloor, \eta_i \leftarrow \left\lfloor \Phi^{k_i/\Delta_\eta} \right\rfloor$.
    for $t = (i-1)H+1,\ldots,i\cdot H \wedge T$ **do**
        Run the `SWUCRL2-CW` algorithm with window size W_i, $\delta = T^{-1}$, and confidence widening parameter η_i, and observe the total rewards $\mathsf{R}\left(W_i, \eta_i, s_{(i-1)H+1}\right)$.
    end for
    Update $q_{(j,k),i+1}$ according to eqn. (3.10).
end for

We are now ready to state the details of the `BORL` algorithm. For some fixed choice of block length H (to be determined later), we first define a couple of additional notations:

$$H = \lfloor 3S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{1}{2}} \rfloor, \Phi = \frac{1}{2T^{\frac{1}{2}}}, \Delta_W = \lfloor \ln H \rfloor, \Delta_\eta = \left\lfloor \ln \Phi^{-1} \right\rfloor, \Delta = (\Delta_W+1)(\Delta_\eta+1), \quad (3.7)$$
$$J_W = \left\{H^0, \left\lfloor H^{\frac{1}{\Delta_W}} \right\rfloor, \ldots, H\right\}, J_\eta = S^{\frac{1}{3}}A^{\frac{1}{4}} \times \left\{\Phi^0, \Phi^{\frac{1}{\Delta_\eta}}, \ldots, \Phi\right\}, J = \left\{(W,\eta) : W \in J_W, \eta \in J_\eta\right\}.$$

Here, J_W and J_η are all possible choices of window size and confidence widening parameter, respectively, and J is the Cartesian product of them with $|J| = \Delta$. We also let $\mathsf{R}_i(W,\eta,s)$

be the total rewards for running the SWUCRL2-CW algorithm with window size W and confidence widening parameter η for block i starting from state s,

The EXP3.P algorithm treats each element of J as an arm. It begins by initializing

$$\alpha = 0.95\sqrt{\frac{\ln\Delta}{\Delta\lceil T/H\rceil}}, \beta = \sqrt{\frac{\ln\Delta}{\Delta\lceil T/H\rceil}}, \gamma = 1.05\sqrt{\frac{\Delta\ln\Delta}{\lceil T/H\rceil}}, q_{(j,k),1} = 0 \ \ \forall\ (j,k) \in M, \tag{3.8}$$

where $M = \{(j',k') : j' \in \{0,1,\ldots,\Delta_W\}, k' \in \{0,1,\ldots,\Delta_\eta\}\}$. At the beginning of each block $i \in [\lceil T/H\rceil]$, the BORL algorithm first sees the state $s_{(i-1)H+1}$, and computes

$$\forall\ (j,k) \in M, \quad u_{(j,k),i} = (1-\gamma)\frac{\exp(\alpha q_{(j,k),i})}{\sum_{(j',k')\in M}\exp(\alpha q_{(j',k'),i})} + \frac{\gamma}{\Delta}. \tag{3.9}$$

Then it sets $(j_i,k_i) = (j,k)$ with probability $u_{(j,k),i}\ \forall\ (j,k) \in M$. The selected pair of parameters are thus $W_i = \left\lfloor H^{j_i/\Delta_W} \right\rfloor$ and $\eta_i = \left\lfloor \Phi^{k_i/\Delta_\eta} \right\rfloor$. Afterwards, the BORL algorithm starts from state $s_{(i-1)H+1}$, selects actions by running the SWUCRL2-CW algorithm with window size W_i and confidence widening parameter η_i for each round t in block i. At the end of the block, the BORL algorithm observes the total rewards $\mathsf{R}\left(W_i,\eta_i,s_{(i-1)H+1}\right)$. As a last step, it rescales $\mathsf{R}\left(W_i,\eta_i,s_{(i-1)H+1}\right)$ by dividing it by H so that it is within $[0,1]$, and updates

$$\forall\ (j,k) \in M \quad q_{(j,k),i+1} = q_{(j,k),i} + \frac{\beta + \mathbf{1}_{(j,k)=(j_i,k_i)} \cdot \mathsf{R}_i\left(W_i,\eta_i,s_{(i-1)H+1}\right)/H}{u_{(j,k),i}}. \tag{3.10}$$

The formal description of the BORL algorithm (with H defined in the next subsection) is shown in Algorithm 4.

3.4.3 Performance Analysis

The dynamic regret guarantee of the BORL algorithm can be presented as follows

Theorem 22. *Assume $S > 1$, the dynamic regret bound of the BORL algorithm is*

$$\tilde{O}\left(D_{\max}(B_r + B_p + 1)^{1/4} S^{2/3} A^{1/2} T^{3/4}\right)$$

3.5 Alternative for Confidence Widening with Application in Inventory Control

As demonstrated in previous sections, running the proposed algorithms with the widened confidence regions can help the DM to attain provably low dynamic regret in general RL in non-stationary MDPs. Nevertheless, confidence widening is not always necessary if the state transition distributions bear a special structure. In particular, we consider the following assumption on the state transition distributions $p_1, \ldots, p_T$.

Assumption 6. *There exists a positive quantity (not necessarily known to the DM) $\zeta \in \mathbb{R}_+$, such that for any pair of states $s, s' \in \mathscr{S}$, there is an action $a_{(s,s')} \in \mathscr{A}_s$ that satisfies $p_t\left(s'|s, a_{(s,s')}\right) \geq \zeta$ for all $t \in [T]$.*

We can now analyze the dynamic regret bound of the `SWUCRL2-CW` algorithm under Assumption 6. Here, we follow the notations introduced in Section 3.1 for consistency. In general, Assumption 6 ensures that for every time step $t \in [T]$, there exists a state transition distribution $p \in H_{p,t}(0)$ such that the induced diameter of the MDP $(\mathscr{S}, \mathscr{A}, p)$ is upper bounded by the constant $\bar{D} := 1/\zeta$ with high probability.

Proposition 23. *Under Assumption 6 and conditioned on the event $\mathscr{E}_{\surd}$, there exists a state transition distribution p in the confidence region $H_{p,t}(0)$, such that the induced diameter of the MDP $(\mathscr{S}, \mathscr{A}, p)$ is at most $\bar{D} := 1/\zeta$ for all $t \in [T]$.*

The proof of Proposition 23 is provided in Section B.14 of the appendix. The proposition indicates that the DM can achieve a bounded dynamic regret by implementing the `SWUCRL2-CW` algorithm with $\eta = 0$. We are now ready to state the dynamic regret bound of the `SWUCRL2-CW` algorithm when Assumption 6 holds (we omit the proof since it is similar to that of Theorem 21.).

Theorem 24. *Under Assumption 6 and assuming $S > 1$, the `SWUCRL2-CW` algorithm with window size W, confidence widening parameter $\eta = 0$, and $\delta = T^{-1}$ satisfies the dynamic*

regret bound

$$\textit{Dyn-Reg}_T(\texttt{SWUCRL2-CW}) = \tilde{O}\left(B_r W + \bar{D}\left[B_p W + \frac{S\sqrt{A}T}{\sqrt{W}} + \frac{SAT}{W} + \sqrt{T}\right]\right)$$

If we further put $W = W^* = S^{2/3}A^{1/2}T^{2/3}(B_r + B_p + 1)^{-2/3}$, *this dynamic regret bound is* $\tilde{O}\left(\bar{D}(B_r + B_p + 1)^{1/3}S^{2/3}A^{1/2}T^{2/3}\right)$.

3.5.1 An Application to Inventory Control

In this subsection, we first elaborate on Assumption 6 in the context of *single non-perishable item inventory control problem with zero lead time, fixed cost, and lost sales* similar to [183], and then demonstrate how to implement the `SWUCRL2-CW` algorithm for this problem. For each time step $t \in [T]$ of the inventory control problem (with some abuse of notations), the following sequence of events happens:

1. The seller first observes her stock level s_t, and decides the quantity a_t to order.

2. If $a_t > 0$, a fixed cost f and a c per-unit ordering cost are incurred, and the order arrives instantaneously. The stock level then becomes $s_t + a_t$.

3. The demand X_t is realized, and the seller observes the censored demand $Y_t = \min\{X_t, s_t + a_t\}$. The DM faces non-stationary demands, in the sense that the demand distributions of $X_1, \ldots, X_T$ at time steps $1, \ldots T$ are independent but not identically distributed.

4. Unfulfilled demand incurs a l per-unit lost sales cost, while excess inventory leads to a h per-unit holding cost. The total cost for time step t is

$$C_t(s_t, a_t) = f \cdot \mathbf{1}[a_t > 0] + c \cdot a_t + l \cdot [X_t - s_t - a_t]^+ + h \cdot [s_t + a_t - X_t]^+. \tag{3.11}$$

   Due to demand censoring, the cost is not *observable*.

The seller's objective is to minimize the cumulative total cost $\sum_{t=1}^{T} C_t(s_t, a_t)$. To map this into the non-stationary MDP model we described in Section 3.1, we represent the level

of stock at the beginning of each time step as the state. Same as [183] (and similar to [106, 186, 12]), we assume the DM has a limited shelf capacity, and she can hold at most S units of inventory at any time. Consequently, $\mathscr{S} = \{0, \ldots, S\}$, and $\mathscr{A}_s = \{0, \ldots, S-s\}$ for each $s \in \mathscr{S}$. We also define the reward and state transition distributions for all $t \in [T]$, $s, s' \in \mathscr{S}$, and $a \in \mathscr{A}_s$ as follows,

$$R_t(s,a) = -C_t(s,a) \qquad \text{and} \qquad p_t(s'|s,a) = \Pr\left(s + a - \min\{s+a, X_t\} = s'\right).$$

However, it is worth emphasizing that, different than our setup in Section 3.1, $R_t(s,a)$ is not observable as $C_t(s,a)$ is not observable. Nevertheless, we shall demonstrate in Section 3.5.1 that one could use the technique of pseudo-reward proposed in [12] to bypass this issue.

Following Assumption 6, we make the *strictly positive probability mass function (PMF) assumption* on $X_1, \ldots, X_T$.

Assumption 7 (Strictly Positive PMF)**.** *There is a $\zeta > 0$ such that $\Pr(X_t = s) \geq \zeta > 0$ for all $t \in [T]$ and $s \in \{0, \ldots, S\}$.*

Remark 14. *It can be readily verified that if the demands satisfy the strictly positive PMF assumption, the underlying inventory control problem satisfies Assumption 6. Indeed, the DM could transit from a state $s \in \mathscr{S}$ to another state $s' \in \mathscr{S}$ with probability at least ζ by ordering $S - s$ units of the item, since then $p_t(s'|s, S-s) = \Pr(X_t = S - s') \geq \zeta$.*

Comparisons to Existing Inventory Control Models

We first compare our setting and existing ones on single non-perishable item inventory control problem with lost sales.

Similar to [106, 186, 183, 12], the model presented in this section studies the single non-perishable item inventory control problem with lost sales. However, there are several key differences between ours and the existing works in terms of cost functions, demand distributions, and lead time:

- **Cost Functions:** In [106], the authors assume a linear purchasing cost function without fixed cost, linear lost sales and holding cost functions. In [183], the authors

	Cost functions	Demand distributions	Lead time
[106]	no fixed cost	stationary, continuous or discrete	zero
[186, 12]	no fixed cost	stationary, continuous or discrete	positive
[183]	with fixed cost	stationary, continuous or discrete	zeros
Ours	with fixed cost	non-stationary, discrete, with strictly positive PMF	zero

Table 3.1: Comparisons between our inventory control model and existing works'

additionally allow fixed cost. In [186, 12], the authors assume the lost sales cost function and the holding cost function are linear, and there is no purchasing cost. In our setting, our cost function is the same as that of [183].

- **Demand Distributions:** In [106, 186, 183, 12], the authors assume stationary demand distributions, but they admit both continuous or discrete demand distributions. In contrast, we allow non-stationary demand distributions, but we impose that the demand distribution has to be discrete, and satisfies the strictly positive PMF assumption described above.

- **Lead Time:** In [186, 12], the authors allow the lead time to be positive; while in [106, 183] and our setting, we assume the lead time is zero.

A summary of the comparisons is provided in Table 3.1.

Implementation of the `SWUCRL2-CW` algorithm

As pointed out in Section 3.5.1, different than the model we present in Section 3.1, the reward in each time step t is not directly observable due to the censored demand. Nevertheless, we can follow the pseudo-reward technique proposed in [12] to implement the `SWUCRL2-CW` algorithm on a sequence of suitably designed pseudo-reward distributions.

In particular, we define the pseudo-reward following [12] for each time step $t \in [T]$, every state s, and every action $a \in \mathscr{A}_s$ as

$$R_t^{\text{pseudo}}(s,a) := R_t(s,a) + l \cdot X_t = -f \cdot \mathbf{1}[a > 0] - c \cdot a_t - h \cdot [s + a - Y_t]^+ + l \cdot Y_t,$$

where we recall $Y_t = \min\{s + a, X_t\}$ is the censored demand. We note that the pseudo-

reward is perfectly *observable*. We also define the mean pseudo-reward or each time step $t \in [T]$, every state s, and every action $a \in \mathscr{A}_s$ as

$$r_t^{\text{pseudo}}(s,a) := \mathbb{E}\left[R_t^{\text{pseudo}}(s,a)\right] = \mathbb{E}\left[R_t(s,a) + l \cdot X_t\right] = r_t(s,a) + l \cdot \mathbb{E}[X_t]. \tag{3.12}$$

This indicates regardless of state and action, the mean pseudo-reward of a time step t can be obtained from shifting the corresponding mean reward uniformly by $l \cdot \mathbb{E}[X_t]$. Without loss of generality, we assume for all $t \in [T]$, $s \in \mathscr{S}$, and $a \in \mathscr{A}_s$, the mean pseudo-reward is bounded, *i.e.*, $r_t^{\text{pseudo}}(s,a) \in [0,1]$, and the pseudo-reward $R_t^{\text{pseudo}}(s,a)$ is 1-sub-Gaussian with mean $r_t^{\text{pseudo}}(s,a)$. Defining $\rho_t^{*\text{pseudo}}$ as the optimal long-term average reward of the stationary MDP with state transition distribution p_t and mean reward $r_t^{\text{pseudo}} = \{r_t^{\text{pseudo}}(s,a)\}_{s\in\mathscr{S},a\in\mathscr{A}_s}$, we can show that for any policy Π, the dynamic regret of the non-stationary MDP instance specified by the tuple $\mathscr{M} = (\mathscr{S}, \mathscr{A}, T, r, p)$ and the dynamic regret of the non-stationary MDP instance specified by the tuple $\mathscr{M}^{\text{pseudo}} = (\mathscr{S}, \mathscr{A}, T, r^{\text{pseudo}} = \{r_t^{\text{pseudo}}\}_{t=1}^T, p)$ are the same.

Proposition 25. *For any policy* Π, *we denote the sample path for following* Π *on* $\mathscr{M}$ *as* $\{s_t(\mathscr{M}), a_t(\mathscr{M})\}_{t=1}^T$, *and the sample path for following* Π *on* $\mathscr{M}^{pseudo}$ *as*

$$\{s_t(\mathscr{M}^{pseudo}), a_t(\mathscr{M}^{pseudo})\}_{t=1}^T,$$

we have

$$\sum_{t=1}^{T} \{\rho_t^* - \mathbb{E}[r_t(s_t(\mathscr{M}), a_t(\mathscr{M}))]\} = \sum_{t=1}^{T} \left\{\rho_t^{*pseudo} - \mathbb{E}[r_t^{pseudo}(s_t(\mathscr{M}^{pseudo}), a_t(\mathscr{M}^{pseudo}))]\right\}.$$

The proof of Proposition 25 is provided in Section B.15 in the appendix. Together with Theorem 24, we have the following dynamic regret bound guarantee for the `SWUCRL2-CW` algorithm on the the single non-perishable item inventory control problem with zero lead time, fixed cost, and lost sales.

Theorem 26. *For the inventory control model in Section 3.5.1, under Assumption 7 and assuming* $S > 1$, *the* `SWUCRL2-CW` *algorithm with window size* W, *confidence widening*

parameter $\eta = 0$, *and* $\delta = T^{-1}$ *satisfies the dynamic regret bound*

$$\textit{Dyn-Reg}_T(\texttt{SWUCRL2-CW}) = \tilde{O}\left(B_r W + \bar{D}\left[B_p W + \frac{S^{\frac{3}{2}}T}{\sqrt{W}} + \frac{S^2 T}{W} + \sqrt{T}\right]\right)$$

If we further put $W = W^* = ST^{2/3}(B_r + B_p + 1)^{-2/3}$, *this dynamic regret bound is*

$$\tilde{O}\left(\bar{D}(B_r + B_p + 1)^{1/3} S T^{2/3}\right).$$

Remark 15. *To interpret the dynamic regret bound of the* `SWUCRL2-CW` *algorithm in the context of inventory control, we note that in Theorem 26, we normalize the cost functions so that the cost incurs in each time period is in* $[0,1]$. *This is slightly different than the setups in [106, 186, 183, 12], where the upper bound of the cost functions are of order* $O(S)$.

3.6 Numerical Experiments

As a complement to our theoretical results, we conduct numerical experiments on synthetic datasets to compare the dynamic regret performances of our algorithms with the UCRL2 algorithm [108], which is one of the most widely used benchmarks for RL in MDPs due to its nearly-optimal regret bound in stationary environments [175], and also the restarting UCRL2 (denoted as UCRL2.S) algorithm for RL in piecewise-stationary MDPs [108]

Setup: We consider a MDP with 2 states $\{s_1, s_2\}$ and 2 actions $\{a_1, a_2\}$, and set $T = 5000$. The mean rewards are set to

$$r_t(s_1,a_1) = 0.2 + 3\cos\left(5V_r \pi t/T\right), \quad r_t(s_1,a_2) = 0.2 + \cos\left(5V_r \pi t/T\right),$$
$$r_t(s_2,a_1) = 0.2 - \cos\left(5V_r \pi t/T\right), \quad r_t(s_2,a_2) = 0.2 - 3\cos\left(5V_r \pi t/T\right).$$

The total variations in mean rewards is thus $B_r = 15V_r = \Theta(V_r)$. An illustration of the reward process of state s_2 and action a_2 is provided in Fig. 3-3 (the mean rewards of other (state,action) pairs are similar). The state transition distributions are set to

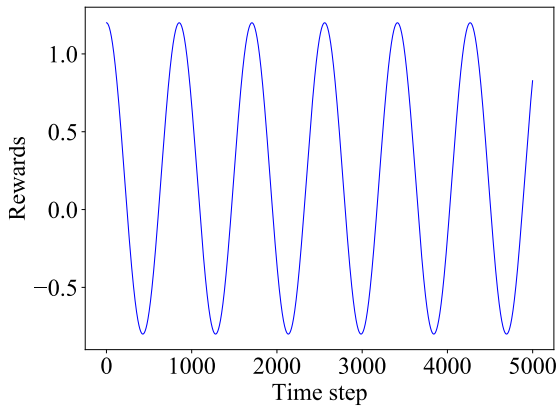
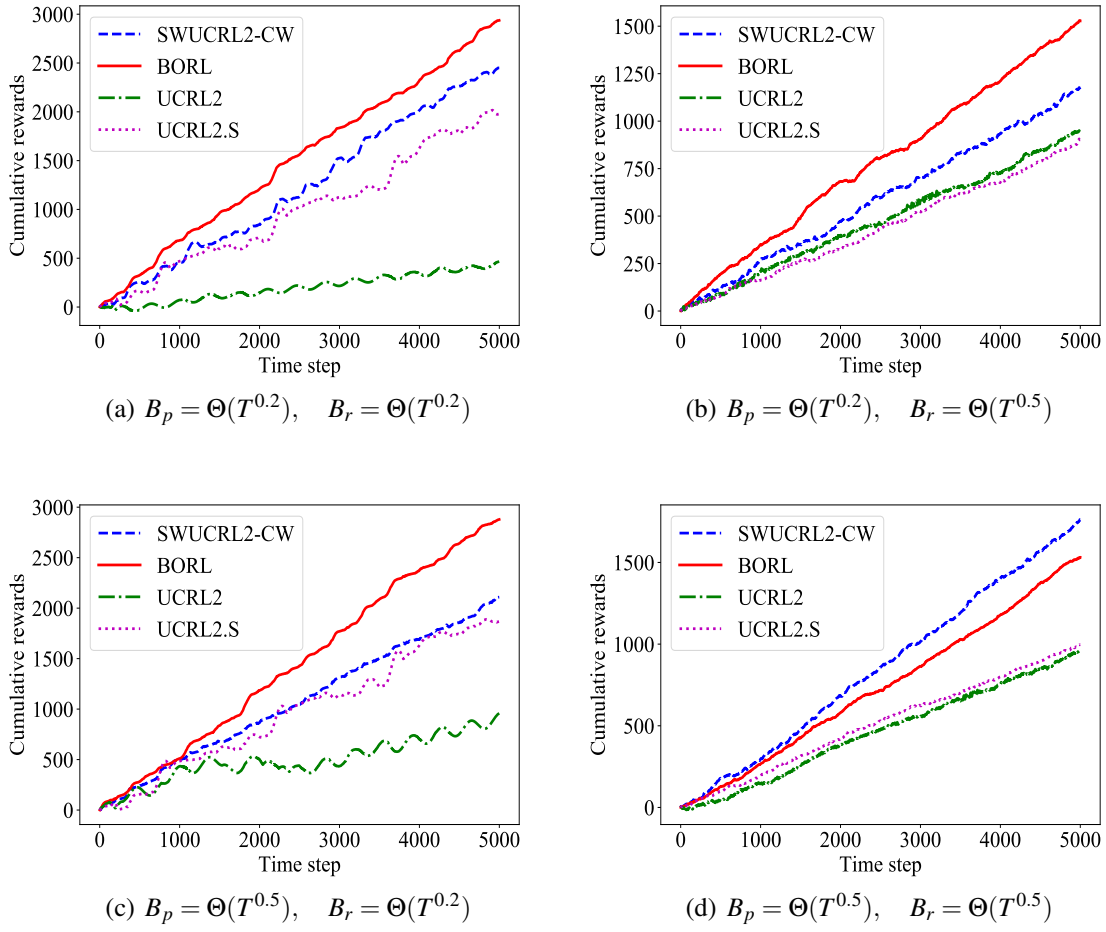


Figure 3-3: Illustrations of mean rewards $r_t(s_2,a_2)$ (the mean rewards of other state-action pairs are similar)

$$p_t(s_1|s_1,a_1)=1,\quad p_t(s_2|s_1,a_1)=0,\quad p_t(s_1|s_1,a_2)=1-\beta_t,\quad p_t(s_1|s_1,a_2)=\beta_t,$$
$$p_t(s_1|s_2,a_1)=0,\quad p_t(s_2|s_2,a_1)=1,\quad p_t(s_1|s_2,a_2)=\beta_t,\quad p_t(s_1|s_2,a_2)=1-\beta_t.$$

where β_t is governed by the process: $\beta_t = 0.5+0.3\sin(5V_p\pi t/T)$. The total variations in the state transition distributions is thus $B_p = 12V_p = \Theta(V_p)$. In this simulation, we allow both V_r and V_p to take values from $\{T^{0.2}, T^{0.5}\}$ to evaluate the performances of the algorithms in low and high variations scenarios. Here, we assume the `SWUCRL2-CW` algorithm knows the variation budgets, and the UCRL2.S algorithm restarts the UCRL2 algorithm every $\lfloor T^{2/3} \rfloor$ time steps. All the results are averaged over 50 runs.

Results: The cumulative rewards of the algorithms under various variation budgets are shown in Fig. 3-4. The results show that both the `SWUCRL2-CW` algorithm and the `BORL` algorithm are able to collect at least 20% more rewards then the UCRL2 algorithm and the UCRL2.S algorithm except for the case when $B_p = \Theta(T^{0.5})$ and $B_r = \Theta(T^{0.2})$, the percentage improvement is 12%. Comparing the results in Figs. 3-4(a), 3-4(b), and 3-4(c), we can see that both the `SWUCRL2-CW` algorithm and the `BORL` algorithm are more robust to variations in the state transition distributions than that in reward distributions. This demonstrate the power of our confidence widening technique. Interestingly, we can see that in Figs. 3-4(a), 3-4(b), and 3-4(c), the cumulative rewards of the `BORL` algorithm (does not know the variation budgets) are higher than those of the `SWUCRL2-CW` algorithm (knows the



(a) $B_p = \Theta(T^{0.2}), \quad B_r = \Theta(T^{0.2})$

(b) $B_p = \Theta(T^{0.2}), \quad B_r = \Theta(T^{0.5})$

(c) $B_p = \Theta(T^{0.5}), \quad B_r = \Theta(T^{0.2})$

(d) $B_p = \Theta(T^{0.5}), \quad B_r = \Theta(T^{0.5})$

Figure 3-4: Cumulative rewards of the algorithms

variation budgets). This indeed has no contradiction to our theoretical results. Theorems 21 and 22 state that the `SWUCRL2-CW` algorithm and the `BORL` algorithm enjoy the same (in the sense of $\tilde{O}(\cdot)$) worst case dynamic regret bound. Nevertheless, the environments we construct in Fig. 3-3 are not the worst case scenario, and the results indicate that the adaptive master algorithm (*i.e.*, the EXP3.P algorithm) of the `BORL` algorithm is able to leverage this more benign environment to attain higher rewards.

Chapter 4

Meta Dynamic Pricing

4.1 Problem Formulation

For ease of exposition, we primarily focus on a seller offering a *single* product at a time. Our approach and results generalize straightforwardly when *multiple* products are offered simultaneously, where a seller must also learn cross-product elasticities to capture substitution effects (see extension in Appendix F).

Notation: Throughout this chapter, 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 $\|x\|_u$ to denote the ℓ_u norm of a vector $x \in \mathscr{R}^d$, but we often omit the subscript when we refer to the ℓ_2 norm. For a matrix $X \in \mathscr{R}^{d\times d}$ $\|X\|_{op} := \max_{v\in\mathscr{R}^d:\|v\|=1} |v^\top X v|$ is the operator norm of X. For a positive definite matrix $A \in \mathscr{R}^{d\times d}$ and vectors $x, \mathbf{y} \in \mathscr{R}^d$, let $\|x\|_A$ denote the matrix norm $\sqrt{x^\top A x}$ and $\langle x, \mathbf{y}\rangle$ denote the inner product $x^\top \mathbf{y}$. For two matrices A and B, we use $A \otimes B$ to denote their Kronecker product. We also denote $x \vee y$ and $x \wedge y$ as the maximum and minimum between $(x,y) \in \mathscr{R}$, respectively. We use the standard notation $O(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$ to characterize the asymptotic growth rate of a function [65]; when logarithmic factors are omitted, we use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$. Finally, let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a matrix respectively.

4.1.1 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.

In epoch $i \in [N]$ at time $t \in [T]$, the seller observes a random feature vector $x_{i,t} \in \mathscr{R}^d$, which is independently and identically distributed from a known distribution $\mathscr{P}_i$. 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

[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 §4.3.4.

$0 < p_{\min} < p_{\max} < \infty$. The seller then observes the resulting induced demand

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

where $\alpha_i \in \mathscr{R}^d$ and $\beta_i \in \mathscr{R}^d$ are unknown fixed constants throughout epoch i, and $\varepsilon_{i,t} \sim \mathscr{N}(0, \sigma^2)$ is i.i.d. Gaussian noise with variance σ^2. This demand model was recently proposed by [26], and captures several salient aspects. In particular, the observed feature vector $x_{i,t}$ 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.

Example 4 (Rue La La)**.** *Rue La La sells a limited set of new products in multi-day "events" [78]. In this case, T is the number of price changes during the event (events are typically 1-4 days, and prices are updated no more than a few times a day), N is the number of events offered so far by the seller (note that $N \gg T$), and K is the number of simultaneously-offered products in an event. For ease of exposition, we primarily consider $K = 1$, but Appendix F provides a straightforward extension to general values of K, accounting for substitution effects.*

Remark 16 (Alternative Demand Models)**.** *Our demand model utilizes a continuous outcome variable, motivated by the setting where many customers simultaneously view the same product with the same price in a given time unit. One can alternatively modify the demand model to follow a generalized linear model (e.g., logistic) to consider a binary purchase outcome variable for each customer. Our proposed algorithms easily generalize by appropriately modifying our Bayesian posterior update rules; however, we restrict our regret analysis to the linear case since OLS Bayesian posterior updates have a closed form, yielding a tractable analysis.*

Shared Structure: For ease of notation, we denote $\theta_i = \begin{pmatrix} \alpha_i^\top & \beta_i^\top \end{pmatrix}^\top \in \mathscr{R}^{2d}$; 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 from previously offered 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 $i \in [N]$.[2] 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 [105, 132]. 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 17 (Product Features)**.** *A complementary form of shared structure can be captured through product features. However, even after conditioning on observed product features, the demand functions for two products may behave very differently, e.g., two black dresses may cater to very different types of customers or have very different price elasticities due to attributes like fit or design that may be hard to capture as features. To capture product-specific (i.e., SKU-level) demand behaviors, we allow the coefficients of the demand function (e.g., price-elasticity) to differ.*

4.1.2 Assumptions

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

Assumption 8 (Boundedness)**.** *The support of the features are bounded, i.e.,*

$$\forall i \in [N], \forall t \in [T] \quad \|x_{i,t}\| \leq x_{\max}.$$

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

[2] 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 [163]).

Our first assumption is that the observed feature vectors $\{x_{i,t}\}$ 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 expected regret at any time step is bounded. This is likely satisfied since features and outcomes are typically bounded in practice.

Assumption 9 (Positive-Definite Feature Covariance)**.** *The minimum eigenvalue of the feature covariance matrix* $\mathbb{E}_{x_{i,t}\sim\mathscr{P}_i}\left[x_{i,t}x_{i,t}^\top\right]$ *in every epoch* $i \in [N]$ *is lower bounded by some positive constant* λ_0*, i.e.,*

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

Our second assumption imposes that the covariance matrix of the observed feature vectors $\mathbb{E}\left[x_{i,t}x_{i,t}^\top\right]$ 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 10 (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.,*

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

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

4.1.3 Background on Thompson Sampling with Known Prior

In this subsection, we consider the setting where the true prior $\mathscr{N}(\theta_*, \Sigma_*)$ 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 [168]. The Thompson sampling algorithm adapted to our dynamic pricing setting for a single epoch $i \in [N]$ is formally given in Algorithm 5 below. Since the prior is known, there is no additional shared structure to exploit across products, so we can treat each epoch independently.

We denote $\mathrm{TS}(\mathcal{N}(\theta_*, \Sigma_*), \lambda_e)$, as the Thompson sampling algorithm with prior $\mathcal{N}(\theta_*, \Sigma_*)$ and a positive input parameter λ_e for initialization. In line with pricing algorithms in the literature (see, *e.g.*, [117, 26]), to ensure that we can obtain a well-defined OLS estimate of the underlying parameter at the end of an epoch, our algorithm initially performs random price exploration (alternating between $p_{\min}$ and $p_{\max}$) until the Fisher information matrix $V_{i,t} = \sum_{s=1}^{t} \begin{pmatrix} x_{i,s}^\top & p_{i,s} x_{i,s}^\top \end{pmatrix}^\top \begin{pmatrix} x_{i,s}^\top & p_{i,s} x_{i,s}^\top \end{pmatrix}$ has minimum eigenvalue of at least λ_e. Let $\mathcal{T}_i$ be the (random) length of this initialization period in epoch i,

$$\mathcal{T}_i \;=\; \arg\min_{t} \lambda_{\min}(V_{i,t}) \;\geq\; \lambda_e \,. \tag{4.1}$$

We show that $\mathcal{T}_i = \tilde{O}(1)$ with high probability (see Lemma 4 in Appendix A), and therefore this initialization period forms a negligible portion of the epoch.

For each time step after initialization, $t \geq \mathcal{T}_i + 1$, the 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 $\mathcal{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}^{\mathrm{TS}}\right)$, and (2) solves and offers the resulting optimal price based on the demand function given by the sampled parameters

$$p_{i,t}^{\mathrm{TS}} = \arg\max_{p \in [p_{\min}, p_{\max}]} \; p \cdot \langle \mathring{\alpha}_{i,t}, x_{i,t} \rangle + p^2 \cdot \left\langle \mathring{\beta}_{i,t}, x_{i,t} \right\rangle . \tag{4.2}$$

Upon observing the actual realized demand $D_{i,t}\left(p_{i,t}^{\mathrm{TS}}, x_{i,t}\right)$, the algorithm computes the posterior $\mathcal{N}\left(\theta_{i,t+1}^{\mathrm{TS}}, \Sigma_{i,t+1}^{\mathrm{TS}}\right)$ for round $t+1$. Specifically, using the update rule for Bayesian linear regression [43] and letting $m_{i,t}^{\mathrm{TS}} = (x_{i,t}^\top, p_{i,t}^{\mathrm{TS}} x_{i,t}^\top)^\top$, the posterior at time t is

$$\theta_{i,t}^{\mathrm{TS}} = \left(\Sigma_*^{-1} + \sigma \sum_{s=1}^{t-1} m_{i,s}^{\mathrm{TS}} (m_{i,s}^{\mathrm{TS}})^\top\right)^{-1} \left(\Sigma_*^{-1}\theta_* + \sigma \sum_{s=1}^{t-1} m_{i,s}^{\mathrm{TS}} D_{i,s}\right),$$

$$\Sigma_{i,t}^{\mathrm{TS}} = \left(\Sigma_*^{-1} + \sigma \sum_{s=1}^{t-1} m_{i,s}^{\mathrm{TS}} (m_{i,s}^{\mathrm{TS}})^\top \right)^{-1}.$$

The same algorithm is applied independently to each epoch $i \in [N]$.

Algorithm 5 $\mathrm{TS}(\mathcal{N}(\theta_*, \Sigma_*), \lambda_e)$: Thompson Sampling Algorithm

Input: The prior mean vector θ_* and covariance matrix Σ_*, the index i of epoch, the length of each epoch T, the noise parameter σ, exploration parameter λ_e.
Initialization: $t \leftarrow 1$, $\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}^{\mathrm{TS}}\right) \leftarrow (\theta_*, \Sigma_*)$.
while $\lambda_{\min}\left(\sum_{s=1}^{t-1} \begin{pmatrix} x_{i,s}^\top \\ p_{i,s} x^\top \end{pmatrix}^\top \begin{pmatrix} x_{i,s}^\top \\ p_{i,s} x^\top \end{pmatrix}\right) < \lambda_e$ **do**
    Observe feature vector $x_{i,t}$, and offer price $p_{i,t}^{\mathrm{TS}} \leftarrow \begin{cases} p_{\max} & \text{if } t \text{ is even,} \\ p_{\min} & \text{otherwise.} \end{cases}$
    Observe demand $D_{i,t}\left(p_{i,t}^{\mathrm{TS}}, x_{i,t}\right)$, and compute the posterior $\mathcal{N}\left(\theta_{i,t+1}^{\mathrm{TS}}, \Sigma_{i,t+1}^{\mathrm{TS}}\right)$.
    $t \leftarrow t+1$
end while
while $t \le T$ **do**
    Observe feature vector $x_{i,t}$.
    Sample parameter $\mathring{\theta}_{i,t} \leftarrow \left[\mathring{\alpha}_{i,t}; \mathring{\beta}_{i,t}\right] \sim \mathcal{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}^{\mathrm{TS}}\right)$.
    $p_{i,t}^{\mathrm{TS}} \leftarrow \arg\max_{p \in [p_{\min}, p_{\max}]} \; p \cdot \langle \mathring{\alpha}_{i,t}, x_{i,t} \rangle + p^2 \cdot \left\langle \mathring{\beta}_{i,t}, x_{i,t} \right\rangle$.
    Observe demand $D_{i,t}\left(p_{i,t}^{\mathrm{TS}}, x_i\right)$, and compute the posterior $\mathcal{N}\left(\theta_{i,t+1}^{\mathrm{TS}}, \Sigma_{i,t+1}^{\mathrm{TS}}\right)$.
    $t \leftarrow t+1$
end while

As evidenced by the large literature on the practical success of Thompson sampling [55, 162, 77], Algorithm 5 is a very attractive choice for implementation in practice.

Algorithm 5 attains a strong performance guarantee under the classical formulation compared to an *oracle* that knows all N product demand parameters $\{\theta_i\}_{i=1}^N$ in advance. In particular, the oracle would offer the expected optimal price in each period $t \in [T]$ in epoch $i \in [N]$, *i.e.*,

$$\begin{aligned} p_{i,t}^* &= \arg\max_{p \in [p_{\min}, p_{\max}]} \; p \cdot \mathbb{E}_\varepsilon[D_{i,t}(p, x_{i,t})] \\ &= \arg\max_{p \in [p_{\min}, p_{\max}]} \; p\langle \alpha_i, x_{i,t} \rangle + p^2 \langle \beta_i, x_{i,t} \rangle. \end{aligned} \tag{4.3}$$

The resulting *Bayes regret* [162] of a policy π relative to the oracle is:

$$\text{Bayes Regret}_{N,T}(\pi) = \mathbb{E}_{\theta,x,\varepsilon}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} p^*_{i,t} D(p^*_{i,t}, x_{i,t}) - \sum_{i=1}^{N}\sum_{t=1}^{T} p^{\pi}_{i,t} D(p^{\pi}_{i,t}, x_{i,t})\right], \tag{4.4}$$

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 27. *When the prior over the demand parameters is known, Algorithm 5 satisfies*

$$\text{Bayes Regret}_{N,T}(\pi) = \tilde{O}\left(d^{\frac{3}{2}} N\sqrt{T}\right),$$

Theorem 27 follows from a similar argument used for the linear bandit setting presented in [162], 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 18. *Prior-independent Thompson sampling [10] achieves a Bayes regret of $\tilde{O}(d^2 N\sqrt{T})$, which is comparable to the performance of Algorithm 5. However, we document a substantial gap in empirical performance between the two approaches in §4.4, motivating our study of learning the prior.*

4.1.4 Meta Oracle and Meta Regret

We cannot directly implement Algorithm 5 in our setting, since the prior over the product demand parameters $\mathcal{N}(\theta_*, \Sigma_*)$ is unknown. In this work, 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 5 (which knows the true prior), we define this policy as our *meta oracle*.[3] Comparing the revenue of our policy relative to the meta oracle

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

leads naturally to the definition of *meta regret* $\mathscr{R}_{N,T}$ for a policy π, *i.e.*,

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

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. Recall that Theorem 27 bounds the Bayes regret of Thompson sampling with a known prior as $\tilde{O}\left(N\sqrt{T}\right)$. Thus, if our meta regret (*i.e.*, the performance of our meta-learning policy relative to Algorithm 5) grows sublinearly in N, then the price of not knowing the prior $\mathcal{N}(\theta_*, \Sigma_*)$ in advance is negligible in experiment-rich environments (large N) compared to the cost of learning the demand parameter for each product (the Bayes regret of Algorithm 5).

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).

Remark 19 (Choice of meta oracle)**.** *To the best of our knowledge, the optimal prior to use for Thompson sampling remains a difficult, open problem. Existing theory shows (in limited settings) that priors that fail to place sufficient mass on the true parameter fare poorly: the closest setting to ours is the linear bandit construction in Proposition 3.1 of [97], which shows that prior-dependent Thompson sampling with a mis-specified prior can achieve regret that scales exponentially in d; Theorem 1 of [132] and Theorem 2 of [105] also provide illustrative constructions with the same insight. In the other extreme, many empirical evaluations suggest that overly conservative priors (such as prior-independent approaches) also fare poorly relative to using the true prior (see, e.g., Section 6 of [31], the discussions in [55], or our numerical results in Section 4.4). As a result, we choose Thompson Sampling with the true prior as our meta oracle. However, one can choose alternative meta oracles — e.g., one that "widens" the true prior to place more weight on parameters that may induce higher regret — implementing such a meta oracle would still*

likely require learning the true prior, which is our primary contribution.

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 $\mathscr{H}_{0,0} = (x_{1,1})$, and $\mathscr{H}_{i,t} = (p_{1,1}, p_{1,2}, \cdots, p_{i,t}, D_{1,1}, D_{1,2}, \cdots, D_{i,t}, x_{1,1}, x_{1,2}, \cdots, x_{i,t+1})$ 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 $\mathscr{F}_{i,t}$ denote the σ-field generated by $\mathscr{H}_{i,t}$. Then, we impose that $\pi_{i,t+1}$ is $\mathscr{F}_{i,t}$ measurable.

4.2 `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.

4.2.1 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 Thompson sampling algorithm to ensure no more than $\tilde{O}(d^2\sqrt{T})$ meta regret for each epoch. After this initial exploration period, our algorithm sequentially updates the estimated prior and leverages this estimate in each subsequent epoch. The key technical challenge is that the estimated prior has finite-sample estimation error, resulting in a Thompson sampling instance with a mis-specified prior. We introduce a prior alignment proof technique to show that, *despite* prior mis-specification, our `Meta-DP` algorithm still achieves meta regret that grows sublinearly in N.

4.2.2 Algorithm

The Meta-DP algorithm is presented in Algorithm 6. We first define some additional notation, and then describe the algorithm in detail.

Additional Notation: Throughout the rest of the chapter, we use $m_{i,t} = \begin{pmatrix} x_{i,t}^\top & p_{i,t}x_{i,t}^\top \end{pmatrix}^\top$ to denote the price and feature information and $V_{i,t} = \sum_{\tau=1}^{t} m_{i,t}m_{i,t}^\top$ to denote the Fisher information matrix of round t in epoch i for all $i \in [N]$ and $t \in [T]$.

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

$$N_0 \;=\; 4c_2^2 d\mathscr{T}_e^2 \log_e(4dN^2T)\log_e(2NT) \;=\; \tilde{O}(d)\,, \tag{4.5}$$

where $\mathscr{T}_e = \max\left\{6\log_{e/2}(dNT)/c_1, 2\lambda_e/c_0\right\} = \tilde{O}(1)$ ($\mathscr{T}_e$ is a high probability upper bound on all $\mathscr{T}_i$'s, see Lemma 4 in Appendix A), and the constant is given by

$$c_2 = \frac{32\sqrt{x_{\max}^2(1+p_{\max}^2)(\sigma^2\lambda_e^{-1}+5\overline{\lambda})}}{\lambda_e\underline{\lambda}\sigma^2}\,.$$

As described in the overview, the Meta-DP algorithm proceeds in two phases. In particular, we distinguish the following two cases for each epoch i:

1. **Epoch $i \le N_0$** : the Meta-DP algorithm runs the prior-independent Thompson sampling algorithm [10, 5] $\mathrm{TS}(\mathscr{N}(0,\Psi I_{2d}),\lambda_e)$, where

$$\Psi = p_{\max}\sigma\sqrt{2d\log_e(T(1+x_{\max}^2p_{\max}^2(1+p_{\max}^2)T))} + \sqrt{20\overline{\lambda}d\log_e(2T)}\,.$$

   This is simply Algorithm 5 with a conservative prior (variance is a function of the horizon T).

2. **Epoch $i > N_0$** : the Meta-DP algorithm first computes the OLS estimate of the true parameter for each previous epoch $j < i$. It then averages these parameter estimates

to form an estimator $\hat{\theta}_i$ of the prior mean θ_*, *i.e.*,

$$\hat{\theta}_i = \frac{\sum_{j=1}^{i-1} V_{j,T}^{-1} \left(\sum_{t=1}^{T} D_{j,t}(p_{j,t}, x_{j,t}) m_{j,t} \right)}{i-1}. \tag{4.6}$$

Then, the `Meta-DP` algorithm runs Thompson Sampling (Algorithm 5) with the estimated prior $\mathcal{N}(\hat{\theta}_i, \Sigma_*)$, *i.e.*, $\mathrm{TS}(\mathcal{N}(\hat{\theta}_i, \Sigma_*), \lambda_e)$. Specifically, after some random initialization steps (these steps are identical to our meta oracle), our `Meta-DP` algorithm (1) samples the unknown product demand parameters $\mathring{\theta}_{i,t} = \left[\mathring{\alpha}_{i,t}; \mathring{\beta}_{i,t}\right]$ from its posterior $\mathcal{N}\left(\theta_{i,t}^{\mathrm{MD}}, \Sigma_{i,t}^{\mathrm{MD}}\right)$, and (2) solves and offers the resulting optimal price based on the demand function given by the sampled parameters

$$p_{i,t} = \underset{p \in [p_{\min}, p_{\max}]}{\arg\max} \; p \cdot \langle \mathring{\alpha}_{i,t}, x_{i,t} \rangle + p^2 \cdot \left\langle \mathring{\beta}_{i,t}, x_{i,t} \right\rangle. \tag{4.7}$$

Upon observing the actual realized demand $D_{i,t}(p_{i,t}, x_{i,t})$, the algorithm computes the posterior $\mathcal{N}\left(\theta_{i,t+1}^{\mathrm{MD}}, \Sigma_{i,t+1}^{\mathrm{MD}}\right)$ for round $t+1$.

Algorithm 6 Meta-Dynamic Pricing Algorithm

Input: The prior covariance matrix Σ_*, the total number of epochs N, the length of each epoch T, the noise parameter σ, and the set of feasible prices $[p_{\min}, p_{\max}]$.
Initialization: N_0 as defined in Eq. (4.5).
for each epoch $i = 1, \ldots, N$ **do**
    if $i \leq N_0$ **then**
        Run $\mathrm{TS}(\mathcal{N}(0, \Psi), \lambda_e)$.
    else
        Update $\hat{\theta}_i$ according to Eq. (4.6), and run $\mathrm{TS}\left(\mathcal{N}\left(\hat{\theta}_i, \Sigma_*\right), \lambda_e\right)$.
    end if
end for

We now state our main result upper bounding the meta regret of our `Meta-DP` algorithm (Algorithm 6). The proof is provided in Section 4.2.3 and Appendix C.

Theorem 28. *The meta regret of the proposed `Meta-DP` algorithm satisfies*

$$\mathscr{R}_{N,T}(\texttt{Meta-DP algorithm}) = \begin{cases} \tilde{O}(d^2 N\sqrt{T}) & \text{when } N < N_0 \\ \tilde{O}(d^2\sqrt{NT}) & \text{otherwise} \end{cases} = \tilde{O}\left(d^2\sqrt{NT} + d^3\sqrt{T}\right).$$

It is worthwhile to compare the bound in Theorem 28 to the $\tilde{O}(d^2N\sqrt{T})$ meta regret bound for prior-independent Thompson Sampling (Lemma 11 in Appendix C). When $N \lesssim \tilde{O}(d)$, our bound matches that of prior-independent Thompson Sampling, since we simply treat all our epochs as exploration epochs. In the large N regime, our meta regret scales as $\tilde{O}(d^2\sqrt{NT})$. Thus, our approach of learning the prior is particularly valuable in experiment-rich settings ($N \gg d$). Combining the two regimes yields a bound that is sublinear in both N and T.

Theorem 28 is somewhat surprising in the context of a growing theoretical literature that suggests that a mis-specified prior can result in very poor regret for prior-dependent Thompson Sampling (see, *e.g.*, [105, 132, 97]). Indeed, one may expect that the mis-specification induced by using the prior $\mathcal{N}(\hat{\theta}_i, \Sigma_*)$ instead of $\mathcal{N}(\theta_*, \Sigma_*)$ can be substantial, since the ratio between these two probability density functions is unbounded when $\hat{\theta}_i \neq \theta_*$. Yet, using our prior alignment proof strategy (described in the next subsection), we establish that Thompson Sampling is remarkably robust to mis-specification of the prior *mean*, lending theoretical support to previous empirical observations [31].

4.2.3 "Prior Alignment" Proof Strategy

Since we only have a logarithmic number (in N and T) of exploration epochs, the meta regret accrued from these epochs is $\tilde{O}(d^2N_0\sqrt{T})$ (see Lemma 11 in Appendix C).

In each non-exploration epoch $i > N_0$, the meta oracle starts with the true prior $\mathcal{N}(\theta_*, \Sigma_*)$ while our algorithm `Meta-DP` starts with the estimated prior $\mathcal{N}(\hat{\theta}_i, \Sigma_*)$. The following lemma (whose proof is in Appendix B) bounds the error of the estimated prior mean with high probability:

Lemma 29. *For any fixed $i \geq 2$ and $\delta \in [0, 2/e]$, with probability at least $1 - \delta - 2/(N^2T^2)$,*

$$\left\|\hat{\theta}_i - \theta_*\right\| \leq 8\sqrt{\frac{2(\sigma^2/\lambda_e + 5\overline{\lambda})d\log_e(4d/\delta)}{i}}.$$

Thus, the key challenge in proving Theorem 28 is bounding the difference in regret incurred by using a Thompson Sampling algorithm with a boundedly mis-specified prior.

We introduce a new "prior alignment" proof technique to address this challenge. At a high level, we show that after the $\mathscr{T}_i$ exploration time steps, the distributions of the meta oracle's (random) posterior estimate $\theta^{\text{TS}}_{i,\mathscr{T}_i+1}$ and `Meta-DP`'s (random) posterior estimate $\theta^{\text{MD}}_{i,\mathscr{T}_i+1}$ are close. More specifically, there is a continuum of realizations of the stochastic noise (in the observed demands) such that `Meta-DP` achieves the *same* posterior estimate $\theta^{\text{MD}}_{i,\mathscr{T}_i+1} = \theta^{\text{TS}}_{i,\mathscr{T}_i+1}$ despite starting with a different prior; when such a match occurs, the expected regret moving forward from time $\mathscr{T}_i+1,\cdots,T$ is the same for both policies. Using this approach, the regret of our `Meta-DP` algorithm can be expressed as a weighted distribution of the regret of the meta oracle (which we bounded in Theorem 27).

More specifically, the following lemma (whose proof is in Appendix C) establishes the difference in Bayesian posteriors between the meta oracle and our `Meta-DP` algorithm. Note that only the means of the posterior differ but the variance is the same.

Lemma 30. *Conditioned on θ_i and $x_{i,1},\ldots,x_{i,\mathscr{T}_i}$, the posteriors of the meta oracle and our `Meta-DP` algorithm satisfy*

$$\theta^{\text{TS}}_{i,\mathscr{T}_i+1} - \theta^{\text{MD}}_{i,\mathscr{T}_i+1} = \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top\right)^{-1} \left(\Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right) + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t}\left(\varepsilon^{\text{TS}}_{i,t} - \varepsilon^{\text{MD}}_{i,t}\right)\right),$$
$$\Sigma^{\text{TS}}_{i,\mathscr{T}_i+1} = \Sigma^{\text{MD}}_{i,\mathscr{T}_i+1}.$$

Now, consider any non-exploration epoch $i \geq N_0+1$. If upon completion of all exploration steps at time $\mathscr{T}_i+1$, we have that the posteriors of the meta oracle and our `Meta-DP` algorithm coincide — *i.e.*, $(\theta^{\text{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MD}}_{i,\mathscr{T}_i+1}) = (\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1})$ — then both policies would achieve the *same* expected revenue over the time periods $\mathscr{T}_i+1,\cdots,T$. By Lemma 30, we know that $\Sigma^{\text{TS}}_{i,\mathscr{T}_i+1} = \Sigma^{\text{MD}}_{i,\mathscr{T}_i+1}$ always, so all that remains is establishing when $\theta^{\text{TS}}_{i,\mathscr{T}_i+1} = \theta^{\text{MD}}_{i,\mathscr{T}_i+1}$.

Since the two algorithms begin with different priors but encounter the same covariates $\{x_{i,t}\}_{t=1}^T$ and take the same decisions in $t \in \{1,\cdots,\mathscr{T}_i\}$, their posteriors can only align at time $\mathscr{T}_i+1$ due to the stochasticity in the observations $\varepsilon_{i,t}$. For convenience, denote the noise terms from $t \in \{1,\cdots,\mathscr{T}_i\}$ of the meta oracle and the `Meta-DP` algorithm respectively

as

$$\chi_i^{\mathrm{TS}} = \begin{pmatrix} \varepsilon_{i,1}^{\mathrm{TS}} & \dots & \varepsilon_{i,\mathscr{T}_i}^{\mathrm{TS}} \end{pmatrix}^\top, \tag{4.8}$$

$$\chi_i^{\mathrm{MD}} = \begin{pmatrix} \varepsilon_{i,1}^{\mathrm{MD}} & \dots & \varepsilon_{i,\mathscr{T}_i}^{\mathrm{MD}} \end{pmatrix}^\top. \tag{4.9}$$

Furthermore, let $M_i = \begin{pmatrix} m_{i,1} & \dots & m_{i,\mathscr{T}_i} \end{pmatrix} \in \mathscr{R}^{2d \times \mathscr{T}_i}$. Lemma 30 indicates that if

$$\chi_i^{\mathrm{MD}} - \chi_i^{\mathrm{TS}} = \frac{1}{\sigma}(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1} \left(\theta_* - \hat{\theta}_i\right), \tag{4.10}$$

then the posteriors of both algorithms align with $\theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}} = \theta_{i,\mathscr{T}_i+1}^{\mathrm{MD}}$. Thus for every realization of the meta oracle's noise terms χ_i^{TS} and the prior mean estimation error $\theta_* - \hat{\theta}_i$, there exists a well-defined and feasible choice of `Meta-DP` algorithm's error χ_i^{MD} that allows the two posteriors to coincide. Furthermore, by Lemma 29, $\|\theta_* - \hat{\theta}_i\|$ is bounded as a function of $\sqrt{1/i}$ with high probability, ensuring that the difference in noise terms $\chi_i^{\mathrm{MD}} - \chi_i^{\mathrm{TS}}$ needed to achieve alignment is small for later epochs (as i grows large). With this observation, we can perform a change of measure over our noise terms and integrate over the resulting distributions, yielding the desired bound on the meta regret. The proof is provided in Appendix C.

Remark 20. *Our prior alignment approach may be of general interest for analyzing the regret of mis-specified Thompson Sampling instances. [162] propose a related but different approach in Section 3.1 of their paper. Specifically, they relate the regret of implementing $TS(\mathcal{N}(\hat{\theta}_i, \Sigma_*), \lambda_e)$ in an environment with true prior $\mathcal{N}(\theta_*, \Sigma_*)$ to the regret of $TS(\mathcal{N}(\hat{\theta}_i, \Sigma_*), \lambda_e)$ in an environment with a different true prior $\mathcal{N}(\hat{\theta}_i, \Sigma_*)$. In contrast, we wish to compare the regret of implementing $TS(\mathcal{N}(\hat{\theta}_i, \Sigma_*), \lambda_e)$ (`Meta-DP`, Algorithm 6) and $TS(\mathcal{N}(\theta_*, \Sigma_*), \lambda_e)$ (meta oracle, Algorithm 5) in the same environment with true prior $\mathcal{N}(\theta_*, \Sigma_*)$. We cannot adopt their approach since one must additionally quantify the difference in regret between TS algorithms learning in environments with different true priors; while this regret difference clearly scales sublinearly in T, we require a bound that limits to 0 as the difference in priors $\|\hat{\theta}_i - \theta_*\| \to 0$ (as $i \to \infty$). This requirement is because even a constant nonzero difference in regret between the meta oracle and our*

Meta-DP algorithm would result in $O(N)$ meta regret over N epochs. To our knowledge, it is an open problem to derive such a bound. Our "prior alignment" sidesteps this issue by directly relating $TS(\mathcal{N}(\hat{\theta}_i, \Sigma_*), \lambda_e)$ and $TS(\mathcal{N}(\theta_*, \Sigma_*), \lambda_e)$ in an environment with true prior $\mathcal{N}(\theta_*, \Sigma_*)$.

4.3 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_*.

4.3.1 Overview

The Meta-DP++ algorithm also begins by using initial product epochs as an exploration phase to initialize our estimate of the prior mean θ_* and covariance Σ_*. After this initial exploration period, our algorithm sequentially updates the estimated prior and leverages this estimate in each subsequent epoch. Once again, the estimated prior has finite-sample estimation error, resulting in a Thompson sampling instance with a mis-specified prior. The key challenge compared to the previous section is that we can no longer exactly "align" our algorithm's posterior with that of the meta oracle when Σ_* is also estimated. We leverage importance sampling arguments from off-policy evaluation to bound the additional meta regret accrued due to this mismatch. Importantly, to ensure that our importance weights remain well-behaved, we *widen* the estimated covariance via a correction term that scales as the finite-sample estimation error of estimating $\hat{\Sigma}_*$.

4.3.2 Algorithm

The Meta-DP++ algorithm is presented in Algorithm 7. We first define some additional notation, and then describe the algorithm in detail.

Additional Notation: As with the Meta-DP algorithm, at the beginning of each epoch $i \in [N]$, we update our estimate $\hat{\theta}_i$ of the prior mean θ_* according to Eq. (4.6). To estimate

Σ_*, we need unbiased and *independent* estimates for the unknown true demand parameter realizations θ_i across epochs.[4] We use the initialization steps $t \in [\mathscr{T}_i]$ to produce an estimate $\dot{\theta}_i$ for θ_i, *i.e.*,

$$\dot{\theta}_i = V_{i,\mathscr{T}_i}^{-1} \left(\sum_{t=1}^{\mathscr{T}_i} D_{i,t}(p_{i,t}, x_{i,t}) m_{i,t} \right).$$

Algorithm Description: The first N_1 epochs are treated as exploration epochs, where we employ the prior-independent Thompson Sampling algorithm. We define

$$N_1 \;=\; \max\left\{N_0,\; 256 c_3^2 d^3 \mathscr{T}_e^2 \log_e^3(4dN^2T),\; c_4^2 d^4 T^2 \log_e^3(2N^2T)\right\} \;=\; \tilde{O}(d^4T^2), \quad (4.11)$$

and the constants are given by

$$c_3 = \frac{16\sqrt{\sigma^2\lambda_e^{-1} + 5\overline{\lambda}}}{\sigma\lambda_e\underline{\lambda}} + \frac{256(\overline{\lambda}\lambda_e^2 + 16\sigma^2)}{\lambda_e^2\underline{\lambda}^2}\left(\frac{8p_{\max}x_{\max}\sqrt{(1+p_{\max}^2)}}{\lambda_e} + \frac{S}{\sigma\lambda_e}\right),$$
$$c_4 = \frac{10^4\sigma(\overline{\lambda}\lambda_e^2 + 16\sigma^2)}{\lambda_e^2\underline{\lambda}^2}.$$

Note that we now require $\tilde{O}(\min\{N, d^4T^2\})$ exploration epochs, whereas we only required $\tilde{O}\left(d^2\right)$ exploration epochs for the `Meta-DP` algorithm.

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

1. **Epoch i $\leq$ N_1:** the `Meta-DP++` algorithm runs the prior-independent Thompson sampling algorithm [10, 5] $\mathrm{TS}(\mathscr{N}(0, \Psi I_{2d}), \lambda_e)$, where

$$\Psi = p_{\max}\sigma\sqrt{2d\log_e(T(1+x_{\max}^2p_{\max}^2(1+p_{\max}^2)T))} + \sqrt{20\overline{\lambda}d\log_e(2T)}.$$

   This is simply Algorithm 5 with a conservative prior (variance is a function of the horizon T).

2. **Epoch i $>$ N_1:** the `Meta-DP++` algorithm computes an estimator $\hat{\theta}_i$ of the prior mean

[4] When estimating the prior covariance, we cannot use an estimator of θ_i that uses all T observations from epoch i (as we do when estimating the prior mean). This is because the use of the learned prior from past epochs renders observations from later epochs non-independent. We avoid this issue by restricting our estimator of θ_i to observations from the initialization periods in each epoch, $t \in [\mathscr{T}_i]$.

θ_* using Eq. (4.6) (same as `Meta-DP` algorithm), and an estimator $\hat{\Sigma}_i$ of the prior covariance Σ_* as

$$\hat{\Sigma}_i = \frac{1}{i-2}\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \frac{\sum_{k=1}^{i-1}\dot{\theta}_k}{i-1}\right)\left(\dot{\theta}_j - \frac{\sum_{k=1}^{i-1}\dot{\theta}_k}{i-1}\right)^\top - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1}. \quad (4.12)$$

The second term $\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]/(i-1)$ accounts for the estimation error in $\{\dot{\theta}_j\}_{j=1}^{i-1}$.

As noted earlier, we then *widen* our estimator to account for finite-sample estimation error:

$$\hat{\Sigma}_i^w \;=\; \hat{\Sigma}_i + \frac{256(\bar{\lambda}\lambda_e^2 + 16\sigma^2 d)}{\lambda_e^2}\sqrt{\frac{5d\log_e(2N^2T)}{i}}\cdot I_{2d}\,, \quad (4.13)$$

where I_{2d} is the $(2d)$-dimensional identity matrix.

Then, the `Meta-DP++` algorithm runs Thompson Sampling (Algorithm 5) with the estimated prior $\mathscr{N}(\hat{\theta}_i, \hat{\Sigma}_i^w)$, *i.e.*, $\mathrm{TS}(\mathscr{N}(\hat{\theta}_i, \hat{\Sigma}_i^w), \lambda_e)$. Specifically, after some random initialization steps (these steps are identical to our meta oracle), our `Meta-DP++` 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 $\mathscr{N}\left(\theta_{i,t}^{\mathrm{MDP}}, \Sigma_{i,t}^{\mathrm{MDP}}\right)$, and (2) solves and offers the resulting optimal price based on the demand function given by the sampled parameters

$$p_{i,t} = \underset{p\in[p_{\min},p_{\max}]}{\arg\max}\; p\cdot\langle\mathring{\alpha}_{i,t}, x_{i,t}\rangle + p^2\cdot\left\langle\mathring{\beta}_{i,t}, x_{i,t}\right\rangle. \quad (4.14)$$

Upon observing the actual realized demand $D_{i,t}\,(p_{i,t}, x_{i,t})$, the algorithm computes the posterior $\mathscr{N}\left(\theta_{i,t+1}^{\mathrm{MDP}}, \Sigma_{i,t+1}^{\mathrm{MDP}}\right)$ for round $t+1$.

We now state our main result upper bounding the meta regret of our `Meta-DP++` algorithm (Algorithm 7). The proof is provided in Section 4.3.3 and Appendix E.

Theorem 31. *The meta regret of the proposed `Meta-DP++` algorithm satisfies*

$$\mathscr{R}_{N,T}(\texttt{Meta-DP++ algorithm}) = \tilde{O}\left(\min\left\{d^2NT^{\frac{1}{2}},\, d^4N^{\frac{1}{2}}T^{\frac{3}{2}}\right\}\right) = \tilde{O}\left(d^3(NT)^{\frac{5}{6}}\right).$$

Algorithm 7 Meta-Dynamic Pricing++ Algorithm

Input: The total number of products N, the length of each epoch T, the noise parameter σ, and the set of feasible prices $[p_{\min}, p_{\max}]$.
for epoch $i = 1, \ldots, N$ **do**
  if $i \leq N_1$ **then**
    Run TS$(\mathcal{N}(0, \Psi), \lambda_e)$.
  else
    Update $\hat{\theta}_i$ and $\hat{\Sigma}_i$ according to Eqs. (4.6) and (4.12) respectively.
    Compute widened prior mean estimate $\hat{\Sigma}_i^w$ according to Eq. (4.13).
    Run TS$\left(\mathcal{N}\left(\hat{\theta}_i, \hat{\Sigma}_i^w\right), \lambda_e\right)$.
  end if
end for

It is worthwhile to compare the bound in Theorem 31 to the $\tilde{O}(d^2 N\sqrt{T})$ meta regret bound for prior-independent Thompson Sampling (Lemma 11 in Appendix C). When $N \lesssim \tilde{O}(d^4 T^2)$, our bound matches that of prior-independent Thompson Sampling, since we simply treat all our epochs as exploration epochs. In the large N regime, our meta regret scales as $\tilde{O}(d^4 N^{\frac{1}{2}} T^{\frac{3}{2}})$. Thus, our approach of learning the prior is particularly valuable in settings with many short-horizon experiments ($N \gg T$). For instance, as discussed in Example 4, sellers like Rue La La host many events, offering new items with short selling seasons. Combining the two regimes yields a bound that is sublinear in both N and T.

4.3.3 Proof Strategy

The number of exploration epochs N_1 is logarithmic in N but quadratic in T. This motivates the analysis of two cases: (i) when the number of epochs $N < N_1 = \tilde{O}(d^4 T^2)$, the meta regret guarantees given by existing prior-independent approaches is already good; (ii) when we transition to an experiment rich environment with $N > N_1$, the meta regret accrued from these epochs is small since their cardinality scales logarithmically in N (see argument in Appendix E). We now focus on the latter case where N is large.

Once again, following the proof strategy employed for `Meta-DP` algorithm, we employ "prior alignment" to match the means of the meta oracle's (random) posterior estimate and `Meta-DP++`'s (random) posterior estimates. However, since Σ_* was known in the previous section, matching the posterior means $\theta^{\text{MD}}_{i,\mathcal{T}_i+1} = \theta^{\text{TS}}_{i,\mathcal{T}_i+1}$ implied equality of the *entire dis-*

tribution of the posterior (see Lemma 30). This equivalence allowed us to exactly equate the expected regret (after alignment) for the meta oracle and our `Meta-DP` algorithm.

However, when Σ_* is unknown, matching the posterior means $\theta^{\text{MDP}}_{i,\mathscr{T}_i+1} = \theta^{\text{TS}}_{i,\mathscr{T}_i+1}$ no longer implies that the posterior distributions are equal. Furthermore, since the Bayesian update for the covariance matrix does not depend on the noise terms (it depends only on the observed covariates and chosen prices), we cannot use any alignment strategy based on χ_i^{TS} and χ_i^{MDP} to get exact equivalence of the posterior distributions. Thus, the key added challenge in proving Theorem 31 is bounding the difference in regret between our `Meta-DP++` algorithm and the meta oracle *after* alignment of the means of their posteriors at time $t = \mathscr{T}_i$.

Specifically, in each non-exploration epoch $i > N_1$, the meta oracle starts with the true prior $\mathscr{N}(\theta_*, \Sigma_*)$ while our algorithm `Meta-DP++` starts with the (widened) estimated prior $\mathscr{N}(\hat{\theta}_i, \hat{\Sigma}_i^w)$. Lemma 29 from the previous section already provides a bound on $\|\hat{\theta}_i - \theta_*\|$, and the following lemma (whose proof is in Appendix D) bounds the error of the estimated covariance $\|\hat{\Sigma}_i - \Sigma_*\|$ (and thus the error of our widened covariance $\|\hat{\Sigma}_i^w - \Sigma_*\|$) with high probability:

Lemma 32. *For any fixed $i \geq 3$ and $\delta \in [0, 2/e]$, with probability at least $1 - 2\delta - 2/(N^2T^2)$,*

$$\left\|\hat{\Sigma}_i - \Sigma_*\right\|_{op} \leq \frac{128(\bar{\lambda}\lambda_e^2 + 16\sigma^2 d)}{\lambda_e^2}\left(\sqrt{\frac{5d\log_e(2/\delta)}{i}} \vee \frac{5d\log_e(2/\delta)}{i}\right).$$

At time $t = \mathscr{T}_i + 1$, we use a change of measure to "align" our `Meta-DP++` algorithm's prior $\mathscr{N}(\theta^{\text{MDP}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})$ to $\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})$. Combining Lemma 32 and the fact that both policies offer the same prices in the random exploration periods, we know that $\Sigma^{\text{TS}}_{i,\mathscr{T}_i+1}$ and $\Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}$ are close with high probability for later epochs. However, it remains to bound the regret difference between the meta oracle's policy, which employs the prior $\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1})$, and our `Meta-DP++` algorithm, which employs the prior $\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})$. We leverage importance sampling arguments from off-policy evaluation [146, 138] to bound this remaining term. Prior widening is instrumental in this last step, ensuring that our importance weights do not diverge.

Remark 21. *While our `Meta-DP` algorithm does not require prior widening, we widen our*

prior for our `Meta-DP++` *algorithm as described above. This allows us to shave off some extra factors of the dimension d in our analysis, by ensuring that the importance weights are well-behaved post-alignment. This is consistent with recent work by [97], who show that Thompson sampling can in general incur a worst-case regret that scales exponentially in d, unless it uses a widened posterior variance at each step. Furthermore, we observe (often significantly) improved empirical performance on both synthetic and real datasets by employing our* `Meta-DP++` *algorithm compared to its non-widened analog (see Section 4.4).*

4.3.4 Additional Remarks

Hierarchical Model: An alternative heuristic to leverage shared structure is to use hierarchical Thompson Sampling, maintaining a posterior on the shared prior and updating it after each epoch. In Appendix G.1, we compare the `Meta-DP` algorithm to a hierarchical approach; while the hierarchical algorithm outperforms prior-independent Thompson Sampling by leveraging shared structure, we find that it still significantly underperforms compared to the `Meta-DP` algorithm for moderate to large values of N due to excessive exploration.

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 [54] for details. For the `Meta-DP` algorithm, we would simply continue to update the estimated prior mean; for the `Meta-DP++` algorithm, we would need to also follow the prior widening schedule. 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, both the `Meta-DP` algorithm and the

`Meta-DP++` algorithm can be modified to only use samples from the *initialization period* $t \in [\mathscr{T}_i]$ in each epoch for estimating the prior mean (note that our estimation of the prior covariance already only uses samples from initialization periods) without affecting the meta regret bounds and analysis. Therefore, when epochs overlap, we will update our estimate of the prior as soon as we see $\tilde{O}(1)$ customer responses for any product.

4.4 Numerical Experiments

We now validate our theoretical results by empirically comparing the performance of our proposed algorithms against prior-independent Thompson Sampling [10]. As discussed earlier, this approach ignores learning shared structure (the prior) across products, and achieves $\tilde{O}(d^2 N\sqrt{T})$ meta regret (see Lemma 11 in Appendix C). When the prior covariance is unknown, we illustrate the benefits of prior widening by additionally comparing against a version of the `Meta-DP++` algorithm that greedily uses the estimated covariance matrix (*i.e.*, $\Sigma_i = \hat{\Sigma}_i$).

In addition to meta regret, we present results on Bayes regret (relative to the classical oracle) to illustrate that our transfer learning approach significantly increases performance under the standard metric. 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.

A number of additional numerical results are presented in Appendix G, including comparison to a hierarchical Thompson Sampling heuristic (G.1), examining the estimation error of the prior as a function of N (G.2), as well as results under a revenue metric (G.3).

4.4.1 Synthetic Data

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

Parameters: We consider $N = 700$ products, each with a selling horizon of $T = 300$ periods. We set the feature dimension $d = 5$, the prior mean $\theta_* = [1.2 \times \mathbf{1}_d; -0.3 \times \mathbf{1}_d]^\top$, and the prior covariance $\Sigma_* = 0.2 \times I_{2d}$. In each epoch $i \in [N]$ and each round $t \in [T]$, each

entry of the observed feature vector $x_{i,t}$ 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 lie in $(0, 5]$. Finally, the noise distribution is the standard normal distribution, *i.e.*, $\sigma = 1$.

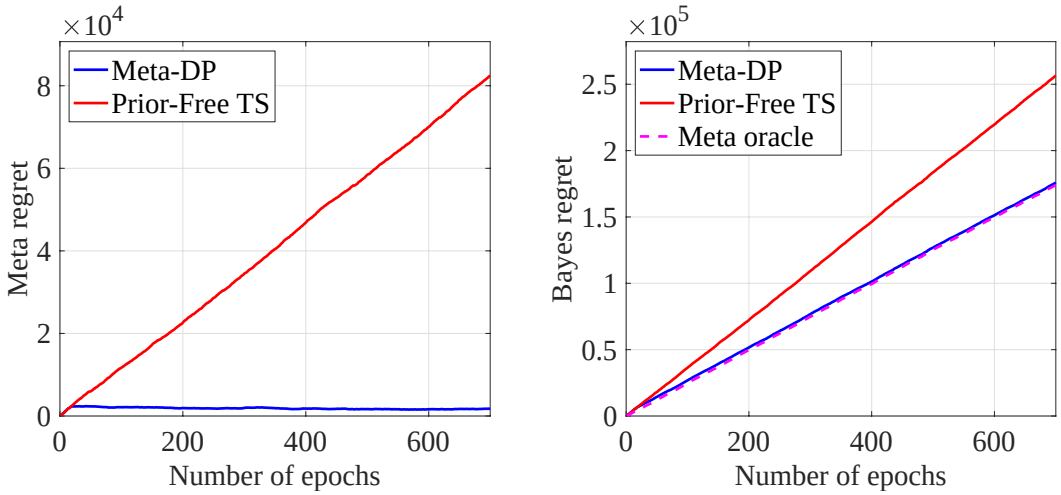


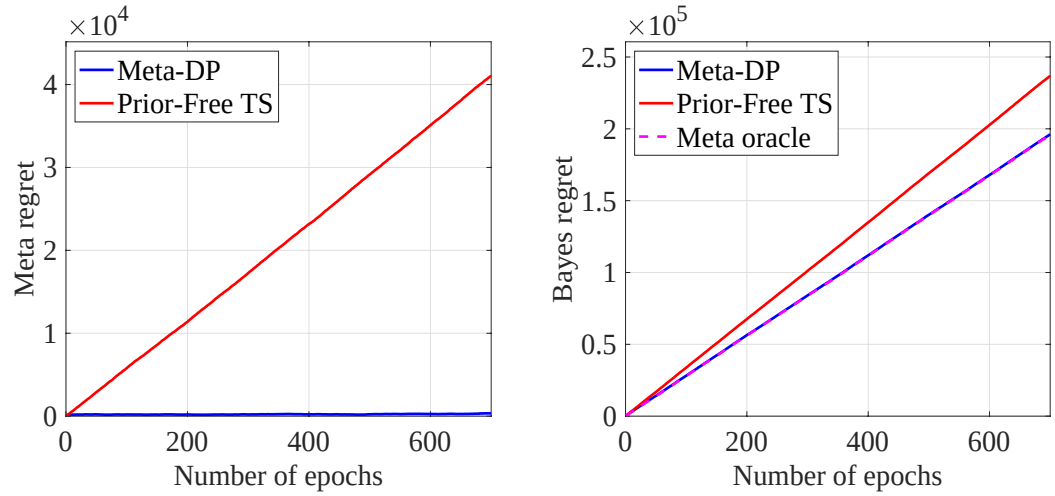
Figure 4-1: Cumulative meta regret and Bayes regret for `Meta-DP` and prior-independent Thompson Sampling.

Results: We plot the cumulative meta regret and Bayes regret of each algorithm, averaged over 20 random trials, as a function of the number of epochs N (recall that each epoch lasts for T periods). The results are shown in Figure 4-1. Both algorithms are identical during the initial exploration epochs.

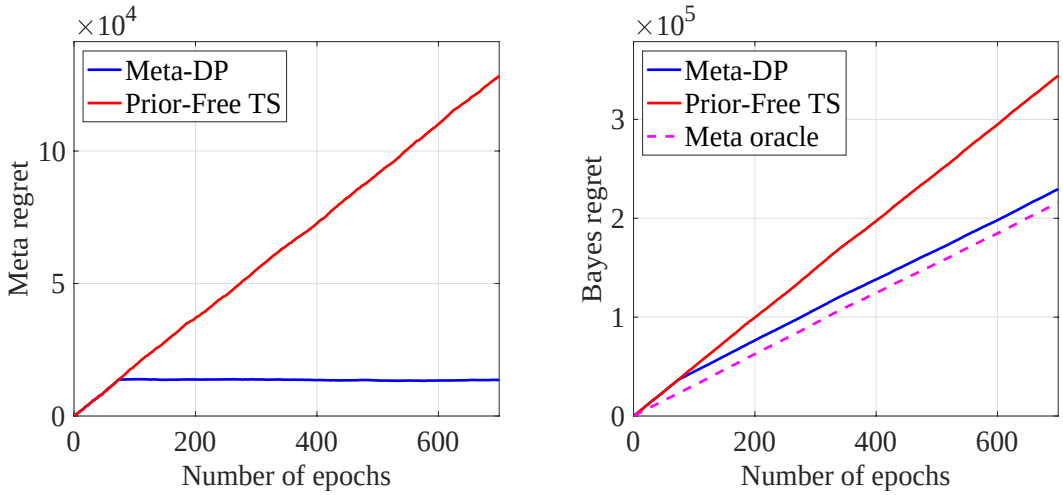
As expected, the prior-independent approach achieves meta regret that scales linearly in N, since each epoch is treated independently. In contrast, the left panel of Figure 4-1 shows that `Meta-DP` achieves nearly zero meta regret after the exploration epochs as it has learned the prior.

The right panel of Figure 4-1 examines Bayes regret; note that even the meta-oracle achieves $O(N)$ Bayes regret (Theorem 27). However, the *slope* of `Meta-DP` closely matches that of the meta-oracle after the initial exploration epochs, *i.e.*, we do not accrue additional regret (relative to the meta oracle) as N grows large. In contrast, the slope of prior-independent Thompson Sampling is significantly larger, resulting in additional regret con-

tinually accruing as N grows large. In particular, when $N = 700$, the Bayes regret of prior independent Thompson Sampling is 39% larger than that of `Meta-DP` and 48% larger than that of the meta oracle. Thus, our approach of learning shared structure is particularly valuable in experiment-rich environments.



(a) $d = 1$



(b) $d = 10$

Figure 4-2: Cumulative meta regret and Bayes regret for `Meta-DP` and prior-independent Thompson Sampling for different values of the feature dimension 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 $x_{i,t} = 1$ for all $i \in [N]$ and $t \in [T]$.

2. *Many features, $d = 10$:* Each entry of the observed feature vector $x_{i,t}$ 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 20 random trials, are shown in Figures 4-2(a) and 4-2(b) respectively. Again, we see that `Meta-DP` substantially outperforms prior-independent Thompson sampling algorithm in both meta regret and Bayes regret, regardless of the choice of feature dimension d. Note that we require more exploration epochs when d is larger (since N_0 scales as d).

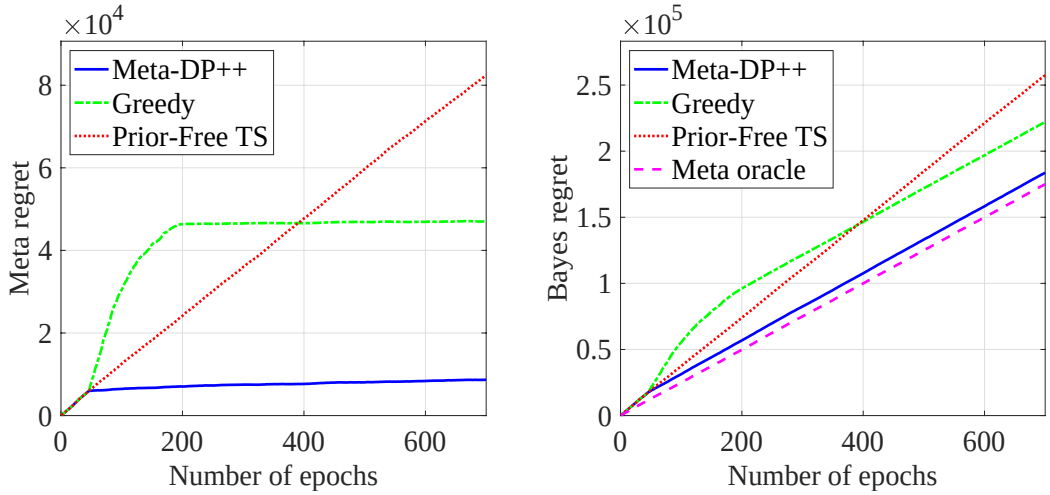


Figure 4-3: Cumulative meta regret and Bayes regret for `Meta-DP++` and benchmark algorithms.

Unknown prior covariance Σ_*: We now shift our attention to the `Meta-DP++` algorithm, and follow the same setup described earlier. To quantify the benefit of prior widening, we additionally consider a version of the `Meta-DP++` algorithm that *greedily* uses the estimated covariance matrix, *i.e.*, $\Sigma_i = \hat{\Sigma}_i$. The results, averaged over 20 random trials, are shown in Figure 4-3. We see that the `Meta-DP++` algorithm significantly outperforms both the prior-independent Thompson sampling algorithm as well as the non-widened greedy benchmark in meta regret (left panel) and Bayes regret (right panel). Interestingly, the greedy approach performs significantly worse in earlier epochs after the initial exploration epochs (when it relies on a prior that is likely to be significantly mis-specified); in later epochs, the greedy

approach's slope begins to match that of `Meta-DP++` as it starts learning the true prior. Thus, prior widening appears critical to ensure good performance on *each* pricing problem — particularly earlier ones, where we should be careful not to over-rely on a prior is likely to be significantly mis-specified. The overall success of `Meta-DP++` suggests that the price of not knowing the prior in advance is negligible in experiment-rich environments (large N).

4.4.2 Real Data on Online Auto-Lending

We now turn to the on-line auto lending dataset. This dataset was first studied by [145], and subsequently used to evaluate dynamic pricing algorithms by [26]. 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 [64] for a detailed description of the dataset.

Algorithms: We consider the setting where both the prior mean and prior covariance are unknown. Thus, we compare the performance of our `Meta-DP++` algorithm against that of prior-independent Thompson Sampling, the ILSX algorithm proposed in [26],[5] and the greedy version of `Meta-DP++` that does not employ prior widening.

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

[5] We do not use a ℓ_1 penalty in our implementation because the "true support" identified by [26] with all observations is precisely the set of features we consider.

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.

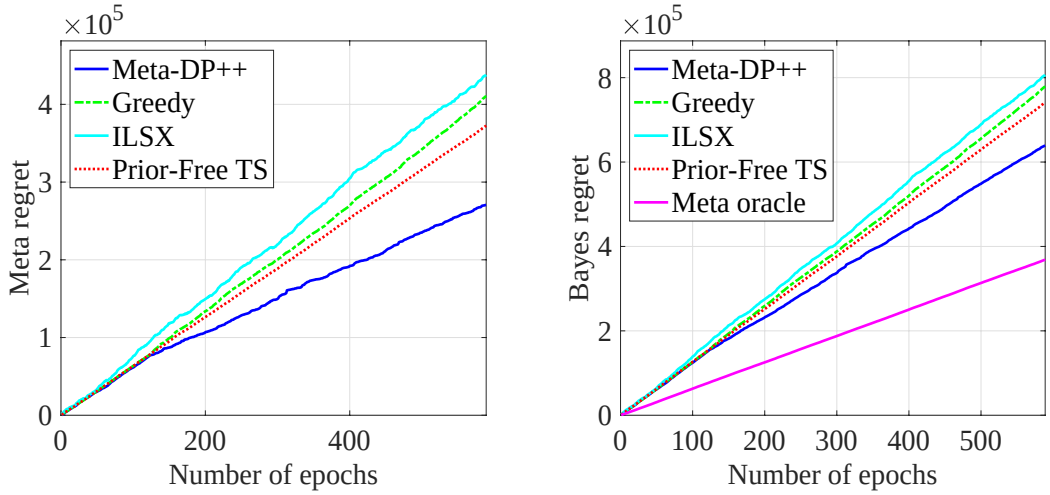
Features: We use the feature selection results from [26], which yields the following features: FICO score, the loan amount approved, prime rate, and the competitor's rate.

Remark 22. *We use three categorical features (state, term of loan, and car type) to define $N = 589$ products, and we additionally have 4 numerical features per product. In contrast, the ILSX algorithm [26] sets $N = 1$ and encodes this information as product features; this results in a feature vector of dimension $4 + 50 + 4 + 3 = \Theta(d + N)$, since each possible value of the categorical feature will be represented as a 1-hot encoding. The resulting meta regret of ILSX will therefore still grow superlinearly in N (unlike our proposed algorithms). Moreover, their demand model is less expressive compared to ours since it does not allow for different price elasticities by state/term/car type (see our earlier Remark 2 for discussion).*

Remark 23. *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 §4.3.4).*

Setup: Following the approach of [145] and [26], 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 $[0, 30]$.

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.,



(a) Cumulative meta regret and Bayes regret for `Meta-DP++` and benchmark algorithms

Figure 4-4: Computational results on a real dataset on online auto loans.

[127]). [40] 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 use the entire dataset to estimate each product's demand parameter, and then fit a multivariate Gaussian prior over the empirical distribution of product demand parameters — our meta oracle uses this prior. However, our regret is evaluated with respect to the true data (*i.e.*, our meta oracle may perform poorly in Bayes regret if the prior is far from a multivariate Gaussian). Thus, this experiment can provide a check on whether our algorithms (which seek to mimic the meta oracle) are robust to model misspecification of the prior.

Results: We average our results over 100 random permutations of the data. The results are shown in Figure 4-4. We first note that, despite potential misspecification of the prior's model class, the meta oracle (prior-dependent Thompson Sampling) achieves much better Bayes regret (right panel) than all algorithms. This implies that the (potentially misspecified) shared prior across products is informative, and thus leveraging shared structure may be valuable. Then, by design, our `Meta-DP++` algorithm learns this shared structure, incurring meta regret that grows sublinearly in N (left panel). Consistent with our results on synthetic data, we see that the `Meta-DP++` algorithm significantly outperforms the bench-

mark algorithms; this is true 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.

Chapter 5

Calibrating Sales Forecast in a Pandemic Using Competitive Online Non-Parametric Regression

5.1 Problem Formulation

In this section, we introduce the notations and the learning protocol.

5.1.1 Notations

We define $[n]$ to be the set $\{1,2,\ldots,n\}$ for any positive integer n. We denote $\mathbf{1}[\cdot]$ as the indicator function. For $p \in [1,\infty]$, we use $\|x\|_p$ to denote the ℓ_p-norm of a vector $x \in \mathbb{R}^d$. We denote $x \vee y$ and $x \wedge y$ as the maximum and minimum between $x, y \in \mathbb{R}$, respectively. For a set $\mathscr{A} \subseteq \mathbb{R}$, we write $\Delta(\mathscr{A})$ as the simplex over $\mathscr{A}$. We adopt the asymptotic notations $O(\cdot), \Omega(\cdot)$, and $\Theta(\cdot)$ [65]. When logarithmic factors are omitted, we use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$, respectively. With some abuse, these notations are used when we try to avoid the clutter of writing out constants explicitly. Given a finite set of covariates $\mathscr{X} \subset \mathbb{R}$, a function $f : \mathscr{X} \to [0,1]$ is isotonic (non-decreasing) if and only if for any pair $x_1, x_2 \in \mathscr{X}$, $x_1 \leq x_2$ implies $f(x_1) \leq f(x_2)$ [156]. For ease of exposition, we denote $\mathscr{F}$ as the set of all possible

isotonic functions that maps $\mathscr{X}$ to $[0,1]$, *i.e.*,

$$\mathscr{F} = \{f : \mathscr{X} \to [0,1] : f \text{ is isotonic}\}. \tag{5.1}$$

5.1.2 Learning Protocol

At each time step $t \in [T]$, the following events happen in sequence: 1) The DM observes the covariate $x_t \in \mathscr{X} \subset \mathbb{R}$ for this time step. We assume the covariates $x_1, \ldots, x_T$ are sampled from a simulatable joint distribution $\mathscr{D} \in \Delta(\mathscr{X}^T)$. 2) She then predicts the label of x_t as $\hat{y}_t$. 3) The label $y_t \in [0,1]$ is revealed, and the DM suffers the squared loss $\ell(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$. Except for boundedness, we make no assumption on the sequence of y_t's, and they could even be chosen adversarially. Specifically, we denote $p_t \in \Delta([0,1])$ as the (possibly adversarial) distribution from which y_t is sampled from. When making prediction at each time step $t \in [T]$, the DM can employ any *non-anticipatory* policy π_t that only takes the history information $\mathscr{H}_{t-1} = \{x_s, y_s\}_{s=1}^{t-1}$ ($\mathscr{H}_0$ is defined to be the empty set $\emptyset$), the observed covariate x_t, and the joint covariate distribution $\mathscr{D}$ as input, and outputs a distribution $q_t \in \Delta([0,1])$ from which $\hat{y}_t$ is sampled from. We denote $\pi = \{\pi_s\}_{s=1}^T$ as the policy used by the DM. The policy π is evaluated by *regret*, defined as

$$\sum_{t=1}^{T} \ell(\hat{y}_t(\pi_t), y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t). \tag{5.2}$$

Intuitively, regret is the difference between the squared ℓ_2-norm associated with labels generated by the algorithm and labels generated by an adversary and the squared ℓ_2-norm associated with labels generated by the best isotonic (non-decreasing) function in hindsight and the adversarial labels. The objective of the DM is to choose a policy π that will minimize the worst case regret over *all* possible choices of y_t's (or p_t's) by the adversary. Formally, the objective is

$$\mathscr{R}_T(\pi|\mathscr{F}) = \mathop{\mathbb{E}}_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{\hat{y}_1 \sim q_1 y_1 \sim p_1} \ldots \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T \sim q_T y_T \sim p_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t(\pi_t), y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \right]. \tag{5.3}$$

We make two remarks regarding the model.

Remark 24 (Model Robustness)**.** *As observed earlier, our notion of regret follows [152], and is similar to competitive analysis used in online resource allocation. Specifically, we make no statistical assumption on the label's generative process, but the adversary cannot be too far from our best understanding of nature. We also demonstrate in Proposition 33 of Section 5.1.4 that the naive greedy learning policy would incur linear in T regret without making additional assumptions on the label's generative process (e.g., eq. (1.1)). Consequently, our model (and our forthcoming solution) is more robust than directly applying eq. (1.1) because now the process that governs the (covariate, label) observations is only asked to be close to monotonic.*

Remark 25 (Worst case extrapolation)**.** *As shown in the definition of regret in eq. (5.3), when the DM makes extrapolations, i.e., when $x_t \notin [\min_{s\in[t-1]} x_s, \max_{s\in[t-1]} x_s]$, she needs to take precautions for any possible y_t to minimize the worst case regret.*

Remark 26 (Comparisons with Online Linear Regression)**.** *A related model is the online linear regression setting [170], where the only difference is that f has to be linear instead of non-parametric (or isotonic). Since the linear function is a proper subset of the isotonic function class, it follows that the isotonic function class is more expressive than the linear function class. Consequently, our oracle can have a smaller loss when compared with an oracle defined w.r.t. the linear function class, yet minimizing regret against in our case is much more challenging.*

5.1.3 Additional Notations: Data-Dependent Discretization

As discussed in Theorem 3 and Section 4 of [121], it is important to work with the discrete isotonic function class when trying to minimize regret. [121] first make the observation that it is only the orders of x_t's (not there particular values) play a role in this problem, and can thus assume w.lo.g. that $\{x_t\}_{t=1}^T$ is a permutation of $[T]$ since they assume all the x_t's are known ahead. Then, they construct a fixed discrete isotonic function class $\mathscr{F}([T]) = \left\{ f \in \mathscr{F} : \forall t \in [T]\ f(t) = \frac{k_t}{K},\ \text{where } k_t \in \{0, 1, \ldots, K\} \right\}$. Here, K is a positive integer to be

specified. Afterwards, they show that the regret defined w.r.t. $\mathscr{F}([T])$ is at most $T/4K^2$ away from the one defined w.r.t. $\mathscr{F}$, and thus work with the former.

Different than [121], however, we do not have access to the x_t's ahead of time, and thus cannot initiate $\mathscr{F}([T])$. To get aroun this, we work with a covariate-specific discrete isotonic function class. Specifically, given a sequence of T covariates $x'_1,\ldots,x'_T \in \mathscr{X}$, we define the corresponding *data-dependent* discrete isotonic function class as

$$\mathscr{F}(\{x'_t\}_{t=1}^T) = \left\{ f \in \mathscr{F} : \forall t \in [T]\ f(x'_t) = \frac{k_t}{K},\ \text{where } k_t \in \{0,1,\ldots,K\} \right\}. \tag{5.4}$$

It is shown in Theorem 4 of [121] that, for any $\mathscr{H}_T = \{x_t, y_t\}_{t=1}^T$, the cumulative loss of the optimal isotonic function in $\mathscr{F}(\{x_t\}_{t=1}^T)$ is at most $T/4K^2$ larger than that of the optimal isotonic function in $\mathscr{F}$, *i.e.*,

$$\inf_{f\in\mathscr{F}} \sum_{t=1}^T \ell(f(x_t),y_t) \leq \inf_{f\in\mathscr{F}(\{x_t\}_{t=1}^T)} \sum_{t=1}^T \ell(f(x_t),y_t) + \frac{T}{4K^2}. \tag{5.5}$$

5.1.4 Inadequacy of Naive Greedy Learning

In this subsection, we demonstrate that simply following the conventional iterative least squares (ILS) policy (*i.e.*, the naive greedy learning policy) would lead to $\Theta(T)$ regret, and thus formally justified that new learning policy is needed for this setup. Recall that at each time step t, the ILS policy would take all historical observations $\mathscr{H}_{t-1} = \{x_s, y_s\}_{s=1}^{t-1}$ and compute $\hat{f}_t$ using least squares as follows:

$$\hat{f}_t = \arg\min_{f\in\mathscr{F}} \sum_{s=1}^{t-1} \ell(f(x_s),y_s). \tag{5.6}$$

Afterwards, it predicts the label at time step t as $\hat{y}_t = \hat{f}_t(x_t)$. In the following proposition, we show that without imposing a monotonic generative process on y_t's, the ILS policy's regret would scale linearly in T even if it does not need to make any non-parametric extrapolations.

Proposition 33. *For any $T \geq 14$, there exists an assignment of $y_1, \ldots, y_T \in [0,1]$ such that even if the DM knows the values of $x_1, \ldots, x_T$ and y_1, y_2 ahead of time and all $x_t \in [x_1, x_2]$ $\forall t \geq 3$ (i.e., no extrapolation is needed), the regret of the ILS policy is at least $(T-1)/8$.*

Proof. Proof Sketch. The compete proof is provided in Section D.1 of the appendix. We let $x_1 = 1, x_2 = T, x_t = t-1$ $\forall t \in [3,T]$, and

$$y_t = \begin{cases} 1 & \text{if } x_t \text{ is odd and } x_t < T; \\ 0 & \text{if } x_t \text{ is even and } x_t < T; \\ \frac{1}{2} & \text{if } x_t = T. \end{cases}$$

See Fig. 5-1 for an illustration. Then, the oracle can employ a constant function $f(x_t) =$

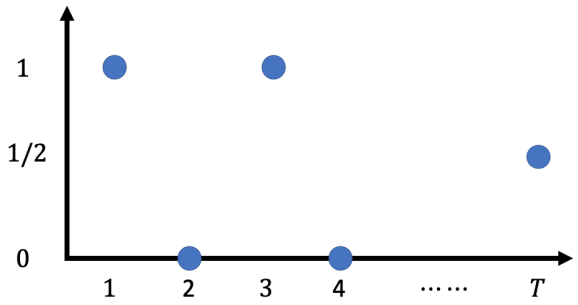
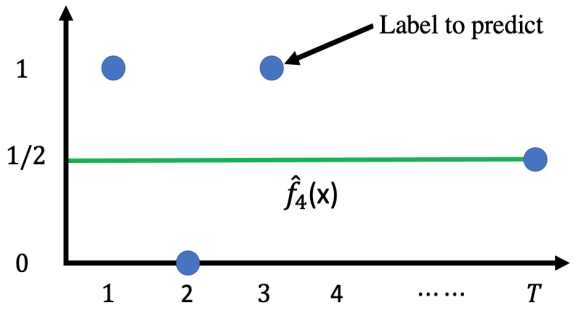


Figure 5-1: Illustration of the true labels

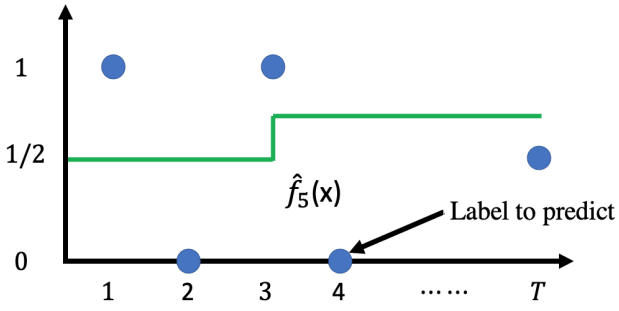
$1/2$ and achieve a cumulative loss at most $\inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq = (T-1)/4$. Now, if the DM follows the ILS policy, it is easy to verify (following the properties of isotonic regression [121]) that when x_t is odd $\hat{f}_t(x) = 1/2$ and hence, $\ell(\hat{f}_t(x_t), y_t) = 1/4$ (left panel of Fig. 5-2); while when x_t is even

$$\hat{f}_t(x) = \begin{cases} \frac{1}{2} & \text{if } x \leq x_{t-1}; \\ \frac{3}{4} & \text{oterhwise}, \end{cases}$$

and hence, $\ell(\hat{f}_t(x_t), y_t) = 9/16$ (right panel of Fig. 5-2). Altogether, $\sum_{t=1}^{T} \ell(\hat{f}_t(x_t), y_t) \geq 12(T-1)/32$. The statement thus follows. □



(a) When x_t is odd, $y_t = 1$ and ILS predicts $\hat{y}_t = 1/2$



(b) When x_t is even, $y_t = 0$ and ILS predicts $\hat{y}_t = 3/4$

Figure 5-2: Oscillating behavior of $\hat{f}_t(\cdot)$

Remark 27 (Instatbility of ILS policy without Eq. (1.1))**.** *In absence of eq.* (1.1)*, the counterexample in Proposition 33 demonstrates that the ILS policy would oscillate between the oracle's choice (i.e., when x_t is odd) and a clearly sub-optimal choice (i.e., when x_t is even) and hence incur a regret that scales with T.*

5.2 Simulating Exponential Weights Policy

[121] designed a computationally-efficient policy to overcomes the computational barrier of general online non-parametric regression methods [152, 89], when all the covariates are known beforehand. In particular, the policy 1) Views $\{x_1, \ldots, x_T\}$ as $[T]$, and constructs the discrete isotonic function class $\mathscr{F}([T])$. 2) Implement a dynamic programming accelerated version of the exponential weights forecast over $\mathscr{F}([T])$ to attain the optimal regret bound. In our setting, however, we only have access to the generative process, but not the realized values, of the covariates, and hence cannot implement step 1). Nevertheless, inspired by [121], we propose a novel Simulating Exponential Weights (`SEW`) policy that additionally

incorporates the simulated future covariates for our problem.

5.2.1 Design Details and Efficient Implementation

For each time step t, after observing the covariate x_t, the SEW policy first samples the covariates $x'_{t+1},\dots,x'_T$ from the joint covariate distribution $\mathscr{D}$. The DM then views $x_1,\dots,x_t$ and $x'_{t+1},\dots,x'_T$ as the covariates she would encounter throughout the course of prediction, and constructs the data-dependent discrete isotonic function class $\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)$ with $K = \left\lceil T^{1/3}/[4(\log_e(T+1))^{1/3}]\right\rceil$ (to be justified in Theorem 34) as described in eq. (5.4) of Section 5.1.3. Afterwards, she makes the prediction according to the exponential weights algorithm [131] over $\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)$. In particular, for each isotonic function $f \in \mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)$, this exponential weight algorithm assigns a weight, which is inversely proportional to the exponential of f's total loss in the previous $t-1$ time steps. Then, it computes the prediction by a weighted average over all the f's evaluated on x_t, *i.e.*,

$$\hat{y}_t = \frac{\sum_{f\in\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)} f(x_t)\exp\left(-\sum_{s=1}^{t-1}(f(x_s)-y_s)^2/2\right)}{\sum_{f\in\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)} \exp\left(-\sum_{s=1}^{t-1}(f(x_s)-y_s)^2/2\right)}. \tag{5.7}$$

Efficient computation: A direct computation of $\hat{y}_t$ via eq. (5.7) is inefficient as it requires an enumeration of every f in $\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)$, whose cardinality is of order $\Theta(T^K)$ (see Theorem 4 of [121]). To mitigate this computational obstacle, we adopt the dynamic programming acceleration technique introduced in [121]. For ease of exposition, we denote $\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)$ by $\mathscr{F}_t$. At each time step t, we can sort the observed covariates $x_1,\dots,x_t$ and the randomly sampled covariates $x'_{t+1},\dots,x'_T$ in ascending order as $z_1,z_2,\dots,z_T$. W.l.o.g., we assume all the z_s's are mutually different. We define for every $s\in[T]$ and every $k\in\{0,\dots,K\}$, $u_s^k = \exp\left(-\mathbf{1}\left[\left(z_s \in \{x_j\}_{j=1}^t\right)\right]\cdot\left(k/K - y_{\sigma(s)}\right)^2/2\right)$, where $\sigma(s)$ is the corresponding subscript of the x_q that is equal to z_s if $z_s\in\{x_j\}_{j=1}^t$, *i.e.*, $\sigma(s)=q$. Suppose x_t is the i^{th} smallest in all the covariates, *i.e.*, $x_t = z_i$, then starting from $w_0^k = 1$ for every k, we recursively compute for all $k=0,\dots,K$, $w_{s+1}^k = \sum_{j=0}^k u_s^j w_s^j$, and

Algorithm 8 SEW policy

Input: Time horizon T, joint covariate distribution $\mathscr{D}$.
Initialize: $K \leftarrow \left\lceil T^{1/3}/[4(\log_e(T+1))^{1/3}] \right\rceil$.
for $t = 1, \ldots, T$ **do**
  Observe context x_t and sample $(x'_{t+1}, \ldots, x'_T)$ from $\mathscr{D}$.
  Rank $\{x_1, \ldots, x_t, x'_{t+1}, \ldots, x'_T\}$ in ascending order as $\{z_{\sigma(1)}, \ldots, z_{\sigma(T)}\}$, where $\sigma(s)$ is the subscript of the s^{th} smallest covariate in $x_1, \ldots, x_t, x'_{t+1}, x'_T$. Let $i \in [T]$ be such that $x_t = z_{\sigma(i)}$.
  $u_s^k \leftarrow \exp\left(-\mathbf{1}\left[\left(z_s \in \{x_j\}_{j=1}^t\right)\right] \cdot \frac{\left(k/K - y_{\sigma(s)}\right)^2}{2}\right)$ for all $s \in [T]$ and $k = 0, \ldots, K$.
  $w_0^k \leftarrow 1, v_T^k \leftarrow 1$ for all $k = 0, \ldots, K$.
  for $s = 1, \ldots, i-1$ **do**
    $w_{s+1}^k \leftarrow \sum_{j=0}^k u_s^j w_s^j$ for $k = 0, \ldots, K$.
  end for
  for $s = T, \ldots, i+1$ **do**
    $v_{s-1}^k \leftarrow \sum_{j=k}^K u_s^j v_s^j$ for $k = 0, \ldots, K$.
  end for
  Predict $\hat{y}_t = \frac{\sum_{k=0}^K \frac{k}{K} w_i^k v_i^k}{\sum_{k=0}^K w_i^k v_i^k}$
end for

this process goes from $s = 1, 2, \ldots, i-1$; Also, starting from $u_T^k = 1$ for every k, we recursively compute for all $k \in \{0, 1, \ldots, K\}$, $v_{s-1}^k = \sum_{j=k}^K u_s^j v_s^j$, and this process goes from $s = T, T-1, \ldots, i+1$. Finally, one can compute $\hat{y}_t = \frac{\sum_{k=0}^K \frac{k}{K} w_i^k v_i^k}{\sum_{k=0}^K w_i^k v_i^k}$. We include a more detailed description of this dynamic programming acceleration in Section D.4 of the appendix. The above procedure gives a $O(TK^2)$ per time step algorithm. A pseudo-code implementation of the SEW policy is provided in Algorithm 8.

5.2.2 Regret Bound

We are now ready to present the regret bound of the SEW policy, whose proof is presented in Section D.3 of the appendix.

Theorem 34. *With* $K = \left\lceil T^{1/3}/[4(\log_e(T+1))^{1/3}] \right\rceil$, *the regret of the* SEW *policy is* $\tilde{O}(T^{1/3})$.

Before providing the intuition about the proof of Theorem 34, we make the following remark regarding optimality of the SEW policy.

Remark 28 (Optimality)**.** *Inspecting the lower bound in Theorem 5 of [121] for the case*

when all x_t's are known beforehand, a strictly easier setting than ours, we know that the regret upper bound in Theorem 34 is optimal up to logarithmic terms.

5.2.3 Proof Sketch of Theorem 34

We begin by applying eq. (5.5) to the regret defined in eq. (5.3),

$$\mathscr{R}_T(\pi|\mathscr{F}) \leq \mathscr{R}_T\left(\pi|\mathscr{F}\left(\{x_t\}_{t=1}^T\right)\right) + \frac{T}{4K^2} \tag{5.8}$$

With this, it is enough to restrict our attention to minimize regret w.r.t. the loss-minimizing isotonic function in $\mathscr{F}(\{x_t\}_{t=1}^T)$. For ease of exposition, we refer to the total squared loss of the loss-minimizing isotonic function in $\mathscr{F}(\{x_t\}_{t=1}^T)$,

$$\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \sum_{t=1}^T \ell(f(x_t), y_t),$$

as the *data-dependent benchmark*.

We then pick a sequence of potential functions $\{V_t\}_{t=0}^T$ that satisfy (informally)

1. The data-dependent benchmark has to incur a total squared loss of at least $-V_T$, *i.e.*,

$$\text{Data-dependent benchmark} := \inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \sum_{t=1}^T \ell(f(x_t), y_t) \geq -V_T, \tag{5.9}$$

2. By using the `SEW` policy, the loss of each time step t is at most $V_{t-1} - V_t$, *i.e.*,

$$\text{Loss of the } \texttt{SEW} \text{ policy at time step } t := \ell(\hat{y}_t, y_t) \leq V_{t-1} - V_t. \tag{5.10}$$

Therefore, we have

$$\begin{aligned}\mathscr{R}_T\left(\pi|\mathscr{F}\left(\{x_t\}_{t=1}^T\right)\right) &= \sum_{t=1}^T \text{Loss of the } \texttt{SEW} \text{ policy at time step } t - \text{Data-dependent benchmark} \\ &\leq \sum_{t=1}^T (V_{t-1} - V_t) + V_T = V_0. \end{aligned} \tag{5.11}$$

To derive the potential functions $\{V_t\}_{t=1}^T$, we follow a modified version of the backward induction relaxation framework in [152] (see Section D.2 of the appendix for more details). Intuitively, each V_t is defined to be the softmax upper bound for the opposite number of expected (w.r.t. $x_{t+1},\ldots,x_T$) data-dependent benchmark of first t time steps, *i.e.*,

$$\begin{aligned} V_t :=& \operatorname{softmax}\left(-\mathop{\mathbb{E}}_{x_{t+1},\ldots,x_T}[\text{data-dependent benchmark of first } t \text{ time steps }]\right) \\ =& \operatorname{softmax}\left(-\mathop{\mathbb{E}}_{x_{t+1},\ldots,x_T}\left[\inf_{f\in\mathscr{F}(\{x_t\}_{t=1}^T)}\sum_{t=1}^{t}\ell(f(x_t),y_t)\right]\right) \\ =& 2\mathop{\mathbb{E}}_{x_{t+1},\ldots,x_T}\log_e\left\{\sum_{f\in\mathscr{F}(\{x_t\}_{t=1}^T)}\exp\left[-\frac{\sum_{s=1}^t(f(x_s)-y_s)^2}{2}\right]\right\} \end{aligned} \tag{5.12}$$

and

$$V_0 := V_0 = K\log_e(T+1) = \tilde{O}(\inf_{\pi}\mathscr{R}_T(\pi|\mathscr{F}(\{x_t\}_{t=1}^T))). \tag{5.13}$$

Here, eq (5.12) critically exploits the availability of the generative process of the covariates, and eq. (5.13) (together with eq. (5.11)) implies the optimality of the `SEW` policy.

Finally, combining eq. (5.8) and (5.13), and setting $K = \left\lceil T^{1/3}/[4(\log_e(T+1))^{1/3}]\right\rceil$, we conclude the statement of the theorem. The formal proof, including how to derive the relaxation for our setting as well as how to apply it to show the regret bound of the `SEW` policy, is provided in Section D.3 of the appendix.

Remark 29 (The Importance of Combining Online Learning and Pandemic Modeling)**.** *In [152], the relaxation framework is developed w.r.t. a model where the x_t's are also chosen by an adversary (formally, this is replacing the expectation over x_t's by supremum over x_t's in the regret definition (5.3)). Unfortunately, as [121] pointed out, the DM has to suffer $\Omega(T)$ regret if the x_t's are chosen by the adversary when $\mathscr{F}$ is the isotonic function class. This implies that a direct adoption of the original relaxation framework is not enough to show the regret bound of the `SEW` policy. We overcome this by critically exploiting the fact that the generative process of x_t's (e.g., the SIR epidemic model in the case of AB InBev's) is available, i.e., eq. (5.12), and customize the relaxation framework to fully leverage the*

availability of the generative process.

Remark 30 (Comparison to [121])**.** *Although the design of the* `SEW` *policy follows [121], its regret analysis requires different techniques. This is because different than [121] (and most existing online learning algorithms [53, 46]), which generate predictions based only on historical observations, the* `SEW` *policy additionally makes use of the simulated future covariates. This makes direct application of conventional regret analysis techniques challenging.*

5.3 Numerical Results

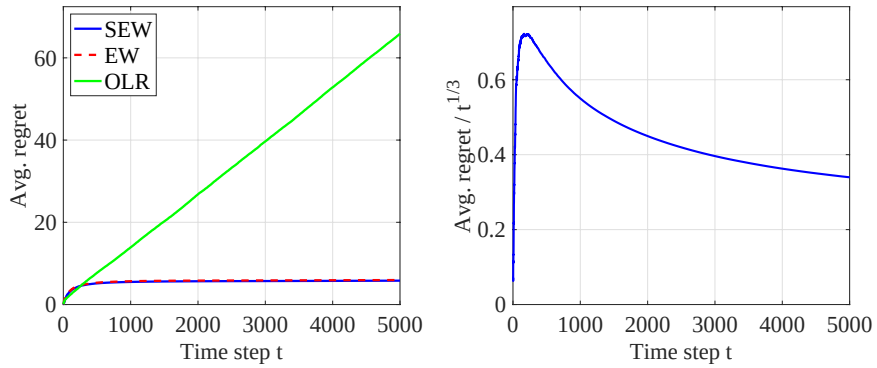
In this section, we conduct numerical simulations with synthetic and AB InBev's datasets to empirically gauge the performance of the `SEW` policy. The first set of simulations, presented in Section 5.3.1, makes use of synthetic datasets to demonstrate that the `SEW` policy outperforms competing algorithms, and show that the regret growth of the `SEW` policy matches our theoretical analysis presented in Theorem 34. We then apply our method to AB InBev's datasets to forecast sales amid the COVID-19 pandemic.

5.3.1 Numerical Simulations with Synthetic Data

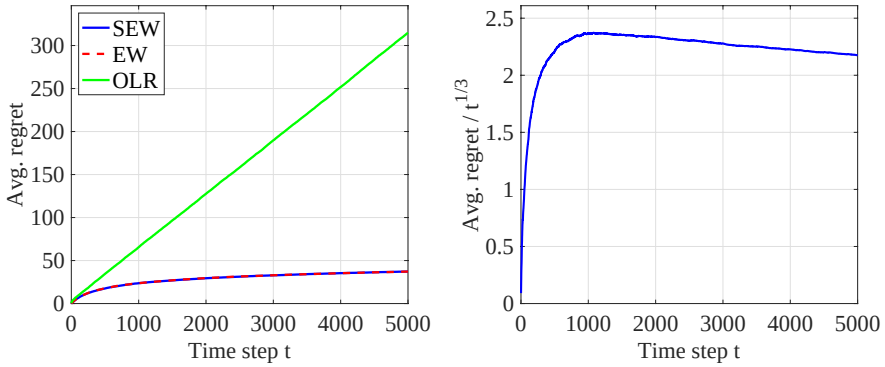
We set $T = 5000$, and assume the covariates x_t's are sampled uniformly from $\mathscr{X} = [0,1]$. The labels y_t's are generated as $y_t = h(x_t)$ for every $t \in [T]$. We consider three different options for $h(\cdot)$:

- *Cubic process:* In this case, $h(x) = x^3$.
- *Stair process:* In this case, $h(x) = \mathbf{1}[x > 0.5]$.
- *Hybrid process:* In this case, $h(x) = \frac{2}{3} \cdot \mathbf{1}[x \in [0.25, 0.5)] + \frac{1}{3} \cdot \mathbf{1}[x \in [0.5, 0.75)] + \mathbf{1}[x \in [0.75, 1]]$.

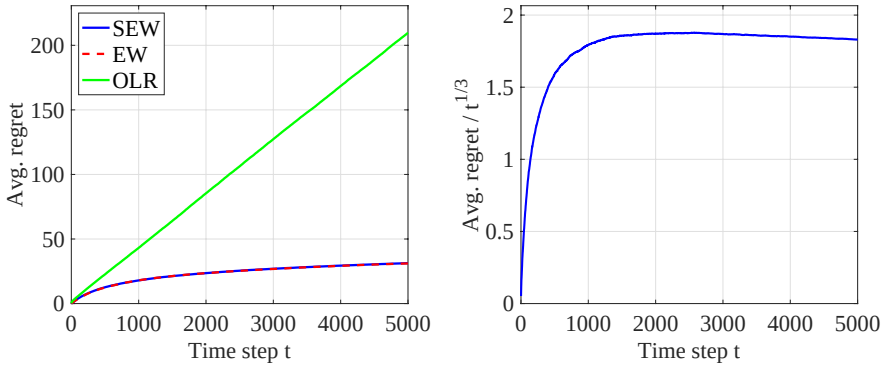
For all the cases, we have $y_t \in [0,1]$ for all $t \in [T]$. For the cubic process and the stair process, $h(\cdot)$'s are non-decreasing. Moreover, the cubic process is continuous, and can be



(a) Cubic process



(b) Stair process



(c) Hybrid process

Figure 5-3: Results for synthetic dataset.

well approximated by a linear function (via Taylor expansion). For the last case, however, $h(\cdot)$ is non-monotonic nor continuous although its overall trend is increasing.

We evaluate the cumulative regret of the `SEW` policy against the online linear regression (OLR) algorithm [170] and the exponential weights (EW) algorithm for the fixed design

case [121]. In the latter case, the EW algorithm has access to the realization of all the x_t's in advance. All the results are averaged over 20 runs.

Results

The results are displayed in Fig. 5-3. In each of the three cases, the left hand side plot depicts the cumulative regret of the three algorithms, while the right hand side plot depicts the cumulative regret of the SEW policy divided by $t^{1/3}$.

Comparing the cumulative regrets of the SEW policy and the EW algorithm, we can see that even though the SEW policy only knows the distribution of x_t's, but not their realizations, its cumulative regret is very close (less than 3% difference) to the cumulative regret of the EW algorithm, which has access to all the realizations of the x_t's.

For the cubic process in Fig. 5-3(a), the SEW policy outperforms the OLR algorithm when $t \geq 400$. This is because the cubic function permits a good linear approximation and the OLR algorithm takes advantage of this. In contrast, the isotonic function class is more expressive, and thus requires more samples to learn. Nevertheless, once a certain sample size threshold is reached, the SEW policy outperforms the OLR algorithm.

For the stair and hybrid processes in Fig. 5-3(b) and Fig. 5-3(c), since no good linear approximation exists, we can see that the regret of the SEW policy is significantly smaller than that of the OLR algorithm throughout.

Finally, the right hand side plots verify that the regret growth of the SEW policy are indeed $T^{1/3}$ in all three cases.

5.3.2 Numerical Simulations with AB InBev's Data

In this section, we apply the SEW policy to the problem of forecasting sales under the COVID-19 pandemic. We use datasets from AB InBev for three geographical regions. We hereafter refer to them as region A, region B, and region C. Each data set contains the baseline sales forecast data and the actual sales data of all AB InBev's products in the respective geographical area. The baseline sales forecast applies AB InBev's offline learning algorithm with historical sales data and other external data, such as social & economic data,

to forecast the sales volumes.

Following the modeling approach in Section 1.4, we sequentially predict the offset in sales (*i.e.*, level of calibration to the baseline sales forecast) caused by the COVID-19 pandemic with the `SEW` policy. We then generate an `SEW` policy calibrated sales forecast by subtracting the predicted offsets from the baseline sales forecast. For both regions A and B, we use the COVID-19 confirmed case numbers, death numbers, and recovered numbers in January 2020 and February 2020 to initialize the respective SIR epidemic model (the SIR epidemic models will then be updated monthly, see Section 5.3.2 for model details), and then evaluate the performances of the `SEW` policy with data beginning from March 2020. For region C, since the pandemic hit this region about a month later, we use the COVID-19 confirmed case numbers, death numbers, and recovered numbers in January, February, and March 2020 to initialize the respective SIR epidemic model, and then evaluate the performances of the `SEW` policy with data beginning from April 2020. We compare our `SEW` policy with the baseline sales forecast and the online linear regression (OLR) algorithm [170] calibrated sales forecast. We measure the performance of each algorithm by weighted mean absolute percentage error (WMAPE) and mean squared error (MSE).

According to AB InBev's needs, we first consider a monthly forecast setting where we predict the sales volumes month by month. To accommodate other potential business applications, we further consider a weekly forecast setting where we predict the sales volumes week after week.

- **Monthly forecast (AB InBev's main focus):** In this case, we predict the sales volumes month by month. Since the number of monthly time periods under the COVID-19 pandemic is small (*e.g.*, ≤ 8 for both regions), the total number of periods in our model will be too small if we set each month as a time step. As a result, we set each day of a month as a time step. Therefore, at the beginning of each month, we predict the daily offsets for this month all at once, and apply the summation of the daily offsets as the monthly offset. Here, the covariate x_t is the COVID-19 case estimate generated by the SIR model at the beginning of each month (for every day t of the month), and is applied in line 4 of Algorithm 8. $\hat{y}_t$ and y_t are the predicted offset and the actual offset of day t, respectively.

- **Weekly forecast:** In this case, we predict the sales volumes week by week. For each week, we need to make the prediction at the beginning of this week. We thus set each week as a time step, the covariate x_t as the mean COVID-19 case estimates of week t. In this process, $\hat{y}_t$ and y_t are the predicted offset and the actual offset of week t, respectively. We apply the SIR epidemic model to generate the COVID-19 case estimates when implementing line 4 of Algorithm 8.

Pandemic Modeling

When generating the COVID-19 case estimates, we work with AB InBev to apply the SIR epidemic model [114]. We use the least squares minimization method to update the parameters, including spread rate β and recovery rate γ, of the SIR epidemic model for both regions month by month.

	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
Spread rate β	0.329	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350
Recovery rate γ	0.050	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

Table 5.1: Monthly updated parameters of SIR epidemic model for region A

	Mar	Apr	May	Jun	Jul (2^{nd} wave)	Aug
Spread rate β	0.280	0.283	0.271	0.270	0.543	0.209
Recovery rate γ	0.074	0.095	0.068	0.067	0.504	0.187

	Sep	Oct	Nov
Spread rate β	0.206	0.952	0.207
Recovery rate γ	0.146	0.872	0.130

Table 5.2: Monthly updated parameters of SIR epidemic model for region B

	Apr	May	Jun	Jul	Aug (2^{nd} wave)	Sep
Spread rate β	0.328	0.239	0.220	0.200	0.200	0.200
Recovery rate γ	0.060	0.051	0.050	0.050	0.106	0.128

	Oct	Nov	Dec (3^{rd} wave)
Spread rate β	0.200	0.200	0.200
Recovery rate γ	0.083	0.050	0.105

Table 5.3: Monthly updated parameters of SIR epidemic model for region C

For region A, we update the SIR epidemic model parameters at the beginning of each month (from March 2020 to July 2020) with historical confirmed COVID-19 case numbers,

death numbers, and recovered numbers from late January 2020 to the end of the previous month. The resulted parameters for region A are reported in Table 5.1.

For region B, we also begin by updating the SIR epidemic model parameters at the beginning of each month with historical confirmed COVID-19 case numbers, death numbers, and recovered numbers from late January 2020 to the end of the previous month. Different than region A, however, the active COVID-19 case numbers in June 2020 indicates that region B was hit by a second wave of the COVID-19 pandemic. To capture the second wave in July 2020 and August 2020, we update the parameters of the SIR epidemic model for July 2020 and August 2020 only with historical data beginning from June 2020. The resulted parameters for region B are reported in Table 5.2.

For region C, since it has experienced three different waves of the COVID-19 pandemic, we follow a similar procedure as region B to compute the SIR forecast parameters, *i.e.*, when a new wave of the pandemic occurs, we discard historical data observed in the prior waves. The resulted parameters for region C are reported in Table 5.3.

Hyper-Parameters

We also introduce a scaling factor and make some minor modifications to the prediction range of the `SEW` policy to better adapt to AB InBev's sales forecast calibration problem.

Scaling factor: Our model and algorithm require that all y_t's belong to the range $[0,1]$. However, the actual offsets in AB InBev's datasets go far beyond this range. To mitigate for this, we additionally apply a (dynamic) scaling factor when implementing the `SEW` policy.

For the monthly forecast setting, we initialize the scaling factor to be the absolute value of the mean of the daily offsets observed from late January 2020 to the end of February 2020. We then sequentially update the scaling factor at the beginning of each month m. Specifically, we first compute the absolute value of the mean offset of all the months from March 2020 to month $m-1$, 2020, and then set the scaling factor to be the floor of the maximum of them. Formally,

$$\text{scaling factor} = \left\lfloor \max \left\{ \left| \frac{\sum_{\text{time step } s \in \text{March}} y_s}{31} \right|, \ldots, \left| \frac{\sum_{\text{time step } s \in \text{month } m-1} y_s}{\#\text{days in month } m-1} \right| \right\} \right\rfloor .$$

For the weekly forecast setting, the scaling factor for each week is the scaling factor of the month to which this week belongs to multiplied by 7 (*i.e.*, there are 7 days in a week).

By doing so, we re-scale the mean of the quantities y_t/scaling facotr so that it lies in $[-1,1]$ most of the time.

Prediction range: To accommodate possibly negative offsets, we further modify lines 6, 7, 9, 12 of Algorithm 8 so that the iterations of k starts from $-K$ instead of 0. The three summations in lines 9 and 14 should also begin from $k=-K$ instead of $k=0$.

Combining both, we replace $y_{\sigma(s)}$ by $y_{\sigma(s)}$/scaling factor in line 6 of Algorithm 8, and $\hat{y}_t$ is computed as scaling factor $\times \left(\sum_{k=-K}^{K} \frac{k}{K} w_i^k v_i^k\right) / \left(\sum_{k=-K}^{K} w_i^k v_i^k\right)$ in line 14 of Algorithm 8.

Mitigation for Impact of Lockdown

From late January 2020 to late March 2020, part of region A has imposed a strict lockdown, where businesses were shut down completely. Since customer behaviors can be drastically different before and after the lockdown is lifted, the lift of lockdown can create non-stationarity for our problem. To combat non-stationarity, we follow an intuitive *restarting strategy* that has been employed by non-stationary online learning in [36, 57]. Specifically, we restart the `SEW` policy and the OLR algorithm at the beginning of April 2020. Under this design, both of them discard all historical data observed during the lockdown period (from late January 2020 to late March 2020). We refer to the restarting version of the `SEW` policy and the OLR algorithm as the `Re-SEW` policy and the Re-OLR algorithm, respectively.

For regions B, since no strict lockdown has been imposed, we don't implement the `Re-SEW` policy and the Re-OLR algorithm for it.

For region C, it has imposed two lockdowns, one from early March 2020 to late May 2020 and the other one beginning in October 2020. Therefore, the `Re-SEW` policy and the Re-OLR algorithm discard all historical data observed during March 2020 to late May 2020 when making predictions from June 2020 to September 2020; while they discard historical data observed from June 2020 to September 2020 (but retaining those observed from March 2020 to late May 2020) when making predictions from October 2020 onwards.

Results for Region A

We first report the results for region A.

Monthly setting: The monthly forecast results are shown in Table 5.4 and Table 5.5.

Table 5.4 implies that the `Re-SEW` policy calibrated sales forecast outperforms the OLR forecast in 6 out of 9 months, the Re-OLR forecast in 8 out of 9 months, and it also outperforms the baseline sales forecast in 6 out of the 9 months.

Table 5.5 implies that the WMAPE of the `Re-SEW` policy calibrated sales forecast is smaller than the WMAPE of the Re-OLR calibrated sales forecast by 53%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the Re-OLR calibrated sales forecast by 88%.

	Mar	Apr	May	Jun	Jul	Aug
Baseline	175.70%	27.09%	-5.46%	-8.58%	**-0.18%**	-5.52%
OLR	**3.17%**	-27.14%	-29.69%	-20.17%	-8.10%	-11.03%
Re-OLR	**3.17%**	-25.79%	4.38%	7.01%	13.03%	3.76%
`SEW` policy	-16.29%	-35.77%	-12.64%	-5.86%	3.90%	-0.70%
`Re-SEW` policy	-16.29%	**-8.81%**	**-4.33%**	**-5.24%**	3.87%	**-0.69%**

	Sep	Oct	Nov
Baseline	6.93%	**2.31%**	-2.21%
OLR	**2.45%**	-4.16%	-9.64%
Re-OLR	24.26%	21.89%	7.57%
`SEW` policy	12.06%	5.08%	**-1.65%**
`Re-SEW` policy	12.06%	5.08%	**-1.65%**

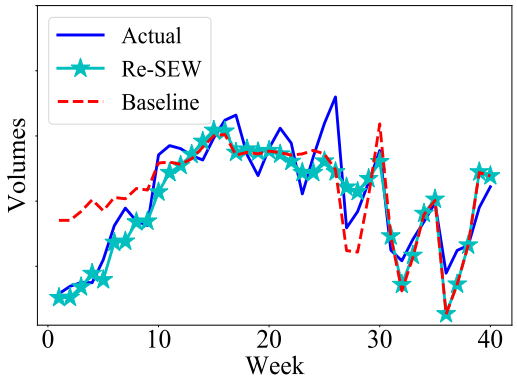
Table 5.4: Percentage forecast errors of different methods for monthly forecast, negative indicates underestimation, the best method of each month is bold (Region A).

	Baseline	OLR	Re-OLR	`SEW` policy	`Re-SEW` policy
WMAPE	12.87%	14.34%	11.69%	9.09%	**5.42%**
MSE	15894705.66	13072148.80	7007389.41	5422899.81	**1449112.15**

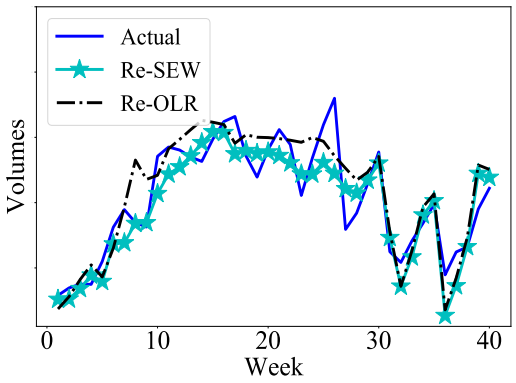
Table 5.5: WMAPE and MSE of different methods for monthly forecast, results of the best method is bold (Region A).

	Baseline	OLR	Re-OLR	`SEW` policy	`Re-SEW` policy
WMAPE	19.78%	19.06%	16.63%	17.89%	**13.84%**
MSE	5497065.12	4528688.89	3595356.34	4228711.84	**2673345.05**

Table 5.6: WMAPE and MSE of different methods for weekly forecast, results of the best method is bold (Region A).



(a) `Re-SEW` policy outperforms the baseline sales forecast in 21 out of 40 weeks.



(b) `Re-SEW` policy outperforms the Re-OLR algorithm in 24 out of 40 weeks.

Figure 5-4: Plot for weekly forecast of region A

The WMAPE of the `Re-SEW` policy calibrated sales forecast is smaller than the WMAPE of the OLR calibrated sales forecast by 62%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the OLR calibrated sales forecast by 88%.

The WMAPE of the `Re-SEW` policy calibrated sales forecast is smaller than the WMAPE of the baseline sales forecast by by 57%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the baseline sales forecast by by 90%.

The WMAPE of the `Re-SEW` policy calibrated sales forecast is smaller than the WMAPE of the `SEW` policy calibrated sales forecast by by 40%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the `SEW` policy calibrated sales forecast by by 73%.

Weekly setting: The weekly forecast results from March to the end of July for region A are depicted in Fig. 5-4 (For brevity, we only plot the results of the Re-SEW policy and the Re-OLR method as their WMAPEs are smaller than the SEW policy and the OLR method, respectively). The WMAPE and MSE of different methods are presented in Table 5.6.

In terms of forecast accuracy, the Re-SEW policy calibrated sales forecast outperforms the Re-OLR calibrated sales forecasts in 24 out of the 40 weeks. It outperforms the baseline sales forecast in 21 out of the 40 weeks.

Table 5.6 implies that the WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the Re-OLR calibrated sales forecast by 16%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the Re-OLR calibrated sales forecast's by 25%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the OLR calibrated sales forecast by 27%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the OLR calibrated sales forecast's by 40%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the baseline sales forecast by 30%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the baseline sales forecast's by 51%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the SEW policy calibrated sales forecast by by 22%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the SEW policy calibrated sales forecast by by 36%.

Remark 31 (Benefits of Restarting)**.** *Comparing the performances of the SEW policy and the Re-SEW policy in the monthly forecast and weekly forecast settings, we can read from the results that the restarting scheme benefits the Re-SEW policy more in the monthly setting. This is because for the monthly forecast setting, the number of updates (i.e., the number of months) the algorithms can make is very small even though the total number of time steps T is larger than the weekly forecast setting. Therefore, it is important for the algorithms to discard irrelevant historical observations in the monthly setting.*

Results for Region B

For region B, since the SIR epidemic model is not very accurate at the beginning of March, the OLR algorithm performs poorly as it is very sensitive to the COVID-19 case estimates. To fix this problem, we propose an OLR++ algorithm, which requires that the level of calibration output by the OLR++ algorithm to be in the range $[-\text{scaling factor}, \text{scaling factor}]$ in March.

Monthly setting: The monthly forecast results are shown in Table 5.7 and Table 5.8.

Table 5.7 implies that the SEW policy calibrated sales forecast outperforms the OLR++ calibrated sales forecast in 7 out of the 9 months and it also outperforms the baseline sales forecast in all 9 months.

Table 5.8 implies that the WMAPE of the SEW policy calibrated sales forecast is smaller than the WMAPE of the OLR++ calibrated sales forecast by 67%. The MSE of the SEW policy calibrated sales forecast is smaller than the MSE of the OLR++ calibrated sales forecast by 95%.

The WMAPE of the SEW policy calibrated sales forecast is smaller than the WMAPE of the baseline sales forecast by by 90%. The MSE of the SEW policy calibrated sales forecast is smaller than the MSE of the baseline sales forecast by by 98%.

Weekly setting: The weekly forecast results from March to the end of August for region B are depicted in Fig. 5-5. The WMAPE and MSE of different methods are presented in Table 5.9.

Table 5.9 implies that the WMAPE of the SEW policy calibrated sales forecast is smaller than the WMAPE of the OLR++ calibrated sales forecast by 21%. The MSE of the SEW policy calibrated sales forecast is smaller than the MSE of the OLR++ calibrated sales forecast's by 51%.

The WMAPE of the SEW policy calibrated sales forecast is smaller than the WMAPE of the baseline sales forecast by 52%. The MSE of the SEW policy calibrated sales forecast is smaller than the MSE of the baseline sales forecast's by 71%.

In terms of forecast accuracy, the SEW policy calibrated sales forecast outperforms the OLR++ calibrated sales forecast in 20 out of the 40 weeks. It outperforms the baseline

	March	April	May	June	July	August
Baseline	47.93%	36.99%	32.99%	26.37%	30.60%	43.82%
OLR++	-73.03%	-3.26%	-3.64%	-3.34%	0.96%	8.58%
SEW policy	**4.99%**	**-0.99%**	**-2.28%**	**-2.66%**	**-0.30%**	**5.27%**

	Sep	Oct	Nov
Baseline	34.03%	50.08%	32.21%
OLR++	**2.12%**	9.97%	**-5.23%**
SEW policy	-2.72%	**3.67%**	-10.63%

Table 5.7: Percentage forecast errors of different methods for monthly forecast, negative indicates underestimation, the best method of each month is bold (Region B).

	Baseline	OLR++	SEW policy
WMAPE	36.48%	10.65%	**3.47%**
MSE	1021912.07	318705.80	**13175.01**

Table 5.8: WMAPE and MSE of different methods for monthly forecast, results of the best method is bold (Region B).

	Baseline	OLR++	SEW policy
WMAPE	36.67%	22.46%	**17.63%**
MSE	326090.48	194756.36	**92519.97**

Table 5.9: WMAPE and MSE of different methods for weekly forecast (Region B).

sales forecast in 32 out of the 40 weeks.

Results for Region C

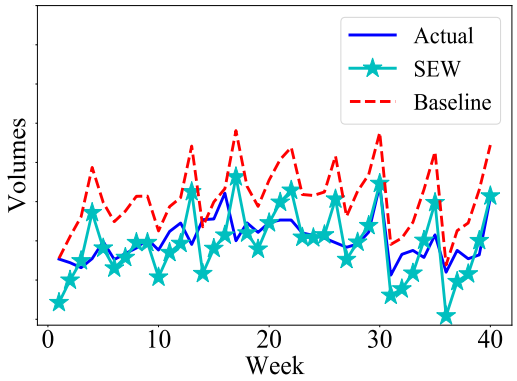
We first report the results for region A.

Monthly setting: The monthly forecast results are shown in Table 5.10 and Table 5.11.

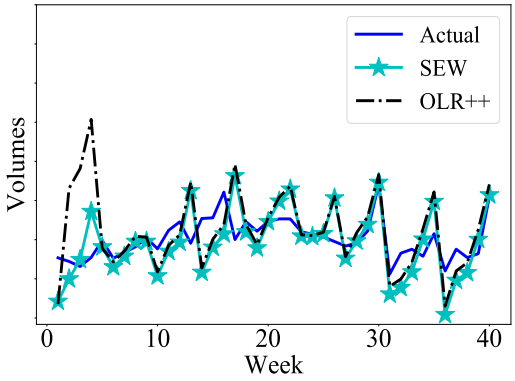
Table 5.10 implies that the Re-SEW policy calibrated sales forecast outperforms the OLR forecast in 8 out of 9 months, the Re-OLR forecast in 8 out of 9 months, and it also outperforms the baseline sales forecast in 6 out of the 9 months.

Table 5.11 implies that the WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the Re-OLR calibrated sales forecast by 63%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the Re-OLR calibrated sales forecast by 87%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the OLR calibrated sales forecast by 62%. The MSE of the Re-SEW policy calibrated



(a) SEW policy outperforms the baseline sales forecast in 32 out of 40 weeks.



(b) SEW policy outperforms the OLR++ algorithm in 20 out of 40 weeks.

Figure 5-5: Plot for weekly forecast of region B

sales forecast is smaller than the MSE of the OLR calibrated sales forecast by 85%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the baseline sales forecast by by 56%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the baseline sales forecast by by 81%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the SEW policy calibrated sales forecast by by 48%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the SEW policy calibrated sales forecast by by 77%.

Weekly setting: The weekly forecast results from March to the end of July for region A are depicted in Fig. 5-6 (For brevity, we only plot the results of the Re-SEW policy and the

	Apr	May	Jun	Jul	Aug	Sep
Baseline	21.68%	24.95%	**2.13%**	-10.03%	-15.66%	**1.28%**
OLR	4.99%	6.73%	-11.29%	-24.70%	-30.55%	9.30%
Re-OLR	4.99%	9.88%	-7.99%	34.14%	30.41%	17.30%
SEW policy	-2.75%	**0.86%**	-15.70%	-10.03%	-15.66%	**1.28%**
Re-SEW policy	**1.90%**	3.60%	-6.13%	**9.15%**	**4.36%**	**1.28%**

	Oct	Nov	Dec
Baseline	**34.58%**	7.78%	56.07%
OLR	44.35%	-12.61%	41.19%
Re-OLR	**5.79%**	-25.71%	50.69%
SEW policy	34.57%	**7.78%**	56.07%
Re-SEW policy	6.62%	-22.35%	**17.45%**

Table 5.10: Percentage forecast errors of different methods for monthly forecast, negative indicates underestimation, the best method of each month is bold (Region C).

	Baseline	OLR	Re-OLR	SEW policy	Re-SEW policy
WMAPE	16.98%	19.68%	20.13%	14.27%	**7.30%**
MSE	29494.01	38682.73	43011.96	24558.33	**5456.78**

Table 5.11: WMAPE and MSE of different methods for monthly forecast, results of the best method is bold (Region C).

	Baseline	OLR	Re-OLR	SEW policy	Re-SEW policy
WMAPE	23.35%	21.44%	20.19%	20.78%	**13.89%**
MSE	11359.11	12386.56	8815.76	12737.73	**5604.29**

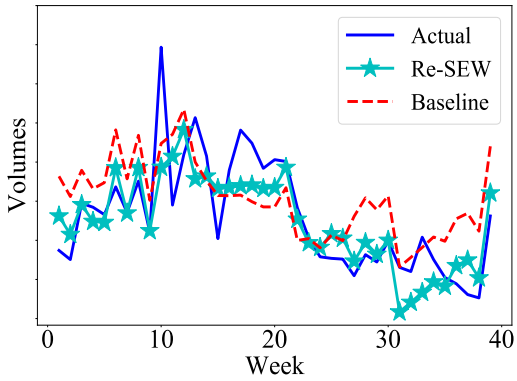
Table 5.12: WMAPE and MSE of different methods for weekly forecast, results of the best method is bold (Region C).

Re-OLR method as their WMAPEs are smaller than the SEW policy and the OLR method, respectively). The WMAPE and MSE of different methods are presented in Table 5.12.

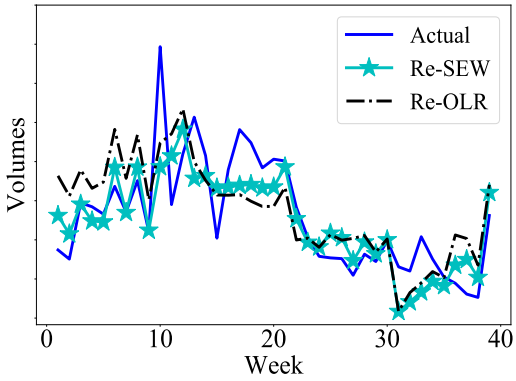
In terms of forecast accuracy, the Re-SEW policy calibrated sales forecast outperforms the Re-OLR calibrated sales forecasts in 28 out of the 39 weeks. It outperforms the baseline sales forecast in 29 out of the 39 weeks.

Table 5.12 implies that the WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE of the Re-OLR calibrated sales forecast by 31%. The MSE of the Re-SEW policy calibrated sales forecast is smaller than the MSE of the Re-OLR calibrated sales forecast's by 36%.

The WMAPE of the Re-SEW policy calibrated sales forecast is smaller than the WMAPE



(a) `Re-SEW` policy outperforms the baseline sales forecast in 29 out of 39 weeks.



(b) `Re-SEW` policy outperforms the Re-OLR algorithm in 28 out of 39 weeks.

Figure 5-6: Plot for weekly forecast of region C

of the OLR calibrated sales forecast by 25%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the OLR calibrated sales forecast's by 54%.

The WMAPE of the `Re-SEW` policy calibrated sales forecast is smaller than the WMAPE of the baseline sales forecast by 30%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the baseline sales forecast's by 51%.

The WMAPE of the `Re-SEW` policy calibrated sales forecast is smaller than the WMAPE of the `SEW` policy calibrated sales forecast by by 40%. The MSE of the `Re-SEW` policy calibrated sales forecast is smaller than the MSE of the `SEW` policy calibrated sales forecast by by 50%.

Chapter 6

Concluding Remarks

In this thesis, we considered data-driven operations in several different dynamically changing environments.

In Chapter 2, we developed general data-driven decision-making algorithms with state-of-the-art dynamic regret bounds for non-stationary bandit settings. We characterized the minimax dynamic regret lower bound and presented a tuned Sliding Window Upper-Confidence-Bound algorithm with matching dynamic regret. We further proposed the parameter-free bandit-over-bandit framework that automatically adapts to the unknown non-stationarity. Finally, we conducted extensive numerical experiments on both synthetic and real-world data to validate our theoretical results.

In Chapter 3, we study the problem of un-discounted reinforcement learning in a gradually changing environment. In this setting, the parameters, *i.e.*, the reward and state transition distributions, can be different from time to time as long as the total changes are bounded by some variation budgets, respectively. We first incorporate the sliding window estimator and the novel confidence widening technique into the UCRL2 algorithm to propose a `SWUCRL2-CW` algorithm with low dynamic regret when the variation budgets are known. We then design a parameter-free `BORL` algorithm that allows us to enjoy the same dynamic regret bound as the `SWUCRL2-CW` algorithm without knowing the variation budgets. The main ingredient of the proposed algorithms is the novel confidence widening technique, which injects extra optimism into the design of learning algorithms, and thus

ensure low dynamic regret bounds. This is in contrast to the widely held believe that optimistic exploration algorithms for (stationary and non-stationary) stochastic online learning settings should employ the lowest possible level of optimism. To extend this finding, we also use the problem of single-item inventory control with fixed cost as an example to demonstrate how one can leverage special structures in the state transition distributions to attain low dynamic regret bound without widening the confidence region.

In Chapter 4, 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 [168, 124]. 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.*, [33, 14]. 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 trials.

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 [105, 132]. 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.*,[162, 31]. 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.

In Chapter 5, together with AB InBev, we consider the problem of sales forecasts calibration due to the impact of the COVID-19 pandemic. Combining tools from online learning and pandemic modeling, we develop a data-driven online non-parametric regression method that takes the current and simulated future active COVID-19 case numbers as input, and outputs the level of calibration of AB InBev's baseline sales forecast. Without making any statistical assumptions on the labels, we propose a computationally-efficient Simulating Exponential Weights (`SEW`) policy for the online non-parametric regression setting. We show that the `SEW` policy achieves the minimax-optimal regret bound. We also demonstrate the empirical performances of the `SEW` policy on both synthetic and AB InBev's datasets of different geographical regions. The AB InBev' numerical experiments show that our method is capable of reducing the forecasting errors in terms of WMAPE and MSE by by more than 50% in the monthly forecast (AB InBev's main focus) and 15% in the weekly forecast.

Bibliography

[1] AB InBev. Annual report 2019. 2019.

[2] Yasin Abbasi-Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013.

[3] Yasin Abbasi-Yadkori, David Pál, and Csaba. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances Neural Information Processing Systems 25 (NIPS)*, 2011.

[4] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[5] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[6] Jacob Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandits, with and without censored feedback. *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

[7] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Proceedings of Annual Conference on Learning Theory (COLT)*, 2017.

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

[9] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

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

[11] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 1184–1194. Curran Associates, Inc., 2017.

[12] Shipra Agrawal and Randy Jia. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2019.

[13] Robin Allesiardo, Raphael Feraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. In *International Journal of Data Science and Analytics*, 2017.

[14] Arielle Anderer, Hamsa Bastani, and John Silberholz. Adaptive clinical trial designs with surrogates: When should we bother? *Available at SSRN 3397464*, 2019.

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

[16] Raman Arora, Ofer Dekel, and Ambuj Tewari. Deterministic mdps with adversarial rewards and bandit feedback. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 93–101, 2012.

[17] J.Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Annual Conference on Learning Theory (COLT)*, 2009.

[18] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. In *SIAM Journal on Computing, 2002, Vol. 32, No. 1 : pp. 48–77*, 2002.

[19] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. In *Journal of Machine Learning Research, 3:397–422, 2002.*, 2002.

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

[21] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning, 47, 235–256*, 2002.

[22] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert Schapire. The non-stochastic multi-armed bandit problem. In *SIAM Journal on Computing*, 2013.

[23] Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, 2019.

[24] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. In *Annals of Mathematical Statistics*, 1955.

[25] Santiago Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 2019.

[26] Gah-Yi Ban and N Bora Keskin. Personalized dynamic pricing with machine learning. 2017.

[27] Gah-Yi Ban and N. Bora Keskin. Personalized dynamic pricing with machine learning. 2018.

[28] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. In *Operations Research*, 2018.

[29] Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 35–42, 2009.

[30] Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 2020.

[31] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 2020.

[32] Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Learning across experiments. In *https://arxiv.org/abs/1902.10918*, 2019.

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

[34] Dimitri Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2017.

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

[36] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed bandit with non-stationary rewards. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

[37] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. In *Operations Research, 2015, 63 (5), 1227–1244*, 2015.

[38] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. In *Forthcomming in Stochastic Systems*, 2018.

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

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

[41] Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. In *https://arxiv.org/abs/1902.01575*, 2019.

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

[43] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[45] Lisa Brown. A-b inbev finalizes $100b billion acquisition of sabmiller, creating world's largest beer company. In *Chicago Tribune*, 2017.

[46] S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning, 2012, Vol. 5, No. 1: pp. 1–122, 2012.

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

[48] Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[49] Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

[50] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[51] Adrian Rivera Cardoso, He Wang, and Huan Xu. Large scale markov decision processes with changing rewards. 2019.

[52] Felipe Caro, A. Gürhan Kök, and Victor Martınez de Albeniz. The future of retail operations. In *Forthcoming at Manufacturing & Services Operations Management*, 2020.

[53] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

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

[55] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

[56] Weidong Chen, Cong Shi, and Izak Duenyas. Optimal learning algorithms for stochastic inventory systems with random capacities. In *SSRN Preprint 3287560*, 2019.

[57] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. In *Proceedings of Conference on Learning Theory (COLT)*, 2019.

[58] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under non-stationarity. In *arXiv:1903.01461*, 2019.

[59] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[60] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *arXiv:2006.14389 [cs.LG]*, 2019.

[61] C. Chiang, T. Yang, C. Lee, M. Mahdavi, C. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *Proceedings of Conference on Learning Theory (COLT)*, 2012.

[62] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[63] Maxime Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. 2016.

[64] Columbia. Center for pricing and revenue management datasets, 2015.

[65] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to algorithms. In *MIT Press*, 2009.

[66] Ruomeng Cui, Gad Allon, Achal Bassamboo, and Jan A. Van Mieghem. Information sharing in supply chains: An empirical and theoretical valuation. In *Management Science*, 2015.

[67] Ruomeng Cui, Dennis J. Zhang, and Achal Bassamboo. Learning from inventory availability information: Evidence from field experiments on amazon. In *Management Science*, 2018.

[68] Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Conference on Learning Theory (COLT)*, 2008.

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

[70] Arnoud V. den Boer and N. Bora Keskin. Discontinuous demand functions: Estimation and pricing. In *Management Science*, 2020.

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

[72] Travis Dick, András György, and Csaba Szepesvári. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[73] Eyal Even-Dar, Sham M Kakade, , and Yishay Mansour. Experts in a markov decision process. In *Advances in Neural Information Processing Systems 18 (NIPS)*, 2005.

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

[75] Louis Faury, Yoan Russac, Marc Abeille, and Clement Calauzenes. Regret bounds for generalized linear bandits under parameter drift. In *https://arxiv.org/abs/2103.05750*, 2021.

[76] Qi Feng, Sirong Luo, and Dan Zhang. Dynamic inventory-pricing control under backorder: Demand estimation and policy optimization. In *Manufacturing & Service Operations Management*, 2014.

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

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

[79] Sarah Filippi, Olivier Cappe, Aurelien Garivier, and Csaba Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing 23 (NIPS)*, 2010.

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

[81] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *NIPS*, 2018.

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

[83] Arthur Flajolet and Patrick Jaillet. Real-time bidding with side information. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[84] Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[85] Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems 31*, pages 2998–3008. Curran Associates, Inc., 2018.

[86] Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of ucrl2b. *https://rlgammazero.github.io/*, 2019.

[87] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1578–1586. PMLR, 10–15 Jul 2018.

[88] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. In *IEEE/ACM Transactions on Networking*, 2012.

[89] Pierre Gaillard and Sebastien Gerchinovitz. A chaining algorithm for online non-parametric regression. In *Conference on Learning Theory (COLT)*, 2015.

[90] Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv:1805.10066*, 2018.

[91] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of International Conferenc on Algorithmic Learning Theory (ALT)*, 2011.

[92] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, pages 174–188. Springer Berlin Heidelberg, 2011.

[93] Google. The arrival of real-time bidding. 2011.

[94] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

[95] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. CRC Press, 1999.

[96] Neha Gupta, Ole-Christoffer Granmo, and Ashok Agrawala. Thompson sampling for dynamic multi-armed bandits. In *Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops*, 2011.

[97] Nima Hamidi and Mohsen Bayati. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

[98] Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. Optimal no-regret learning in repeated first-price auctions. In *arXiv:2003.09795*, 2020.

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

[100] J. Michael Harrison, N. Bora Keskin, and Assaf Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. In *Management Science*, 2012.

[101] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed bandit, dynamic environments and meta-bandits. In *NIPS workshop of Online Trading between Exploration and Exploitation*, 2006.

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

[103] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. In *Machine Learning*, 2007.

[104] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[105] Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *AISTATS*, pages 375–383, 2014.

[106] Woonghee Tim Huh and Paat Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 2009.

[107] A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan. Online optimization : Competing with dynamic comparators. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[108] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010.

[109] Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *JMLR*, 2019.

[110] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. *arxiv:1912.01192*, 2019.

[111] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv:1902.03736*, 2019.

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

[113] Z. Karnin and O. Anava. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

[114] William Ogilvy Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society A*, 1927.

[115] N. Keskin and A. Zeevi. Chasing demand: Learning and earning in a changing environments. In *Mathematics of Operations Research, 2016, 42(2), 277–307*, 2016.

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

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

[118] Michael Jong Kim and Andrew E.B. Lim. Robust multiarmed bandit problems. *Management Science*, 2016.

[119] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. In *Information and Computation*, 1997.

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

[121] Wojciech Kotlowski, Wouter M. Koolen, and Alan Malek. Online isotonic regression. In *Annual Conference on Learning Theory (COLT)*, 2016.

[122] Wojciech Kotlowski, Wouter M. Koolen, and Alan Malek. Random permutation online isotonic regression. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[123] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*, 2015.

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

[125] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2018.

[126] Alan Laub. *Matrix Analysis for Scientists and Engineers*. Society of Industrial and Applied Mathematics, 2004.

[127] Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of International conference on World wide web (WWW)*, 2010.

[128] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

[129] Xiaocheng Li, Yufeng Zheng, Zhenpeng Zhou, and Zeyu Zheng. Demand prediction, predictive shipping, and product allocation for large-scale e-commerce. In *Available at SSRN: https://ssrn.com/abstract=3277125*, 2019.

[130] Yingying Li, Aoxiao Zhong, Guannan Qu, and Na Li. Online markov decision processes with time-varying transition probabilities and rewards. 2019.

[131] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. In *Information and Computation*, 1994.

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

[133] Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[134] H. Luo, C. Wei, A. Agarwal, and J. Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of Conference on Learning Theory (COLT)*, 2018.

[135] Will Ma. Improvements and generalizations of stochastic knapsack and markovian bandits approximation algorithms. *Mathematics of Operations Research*, 2018.

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

[137] Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Is model-free learning nearly optimal for non-stationary rl? In *arXiv:2010.03161 [cs.LG]*, 2020.

[138] Susan Murphy, Mark van der Laan, James Robins, and CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association (JASA)*, 2001.

[139] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 1804–1812. 2010.

[140] Gergely Neu, Andras Gyorgy, and Csaba Szepesvari. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 805–813. PMLR, 21–23 Apr 2012.

[141] John Von Neumann. On the theory of board games. In *Math. Ann.*, 1928.

[142] Nielsen CPG, FMCG & RETAIL. Key consumer behavior thresholds identified as the coronavirus outbreak evolves. 2020.

[143] Optimizely. Online, 2019. [Last accessed January 21, 2019].

[144] Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

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

[146] Doina Precup, Richard Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *International Conference on Machine Learning (ICML)*, 2000.

[147] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[148] Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. 2016.

[149] Rajat Raina, Andrew Y Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *ICML*, pages 713–720. ACM, 2006.

[150] Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. In *https://arxiv.org/abs/1707.09727*, 2017.

[151] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[152] Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Annual Conference on Learning Theory (COLT)*, 2014.

[153] Philippe Rigollet and Jan-Christian Hütter. *High Dimensional Statistics*. Lecture Notes, 2019.

[154] R. Rigollet and J. Hütter. *High Dimensional Statistics*. Lecture Notes, 2018.

[155] Alessandro Rinaldo. Lecture notes on advanced statistical theory. In *Available at: http://www.stat.cmu.edu/ arinaldo/Teaching/36755/F17/*, 2017.

[156] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. John Wiley & Sons, 1998.

[157] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5478–5486. PMLR, 2019.

[158] Peter E. Rossi, Greg M. Allenby, and Robert McCulloch. *Bayesian Statistics and Marketing*. John Wiley & Sons, Ltd, 2005.

[159] Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Multi-point bandit algorithms for nonstationary online nonconvex optimization. In *https://arxiv.org/abs/1907.12340*, 2019.

[160] Paat Rusmevichientong and John Tsitsiklis. Linearly parameterized bandits. In *Mathematics of Operations Research, 35(2):395–411, 2010*, 2010.

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

[162] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. In *Mathematics of Operations Research*, 2014.

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

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

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

[166] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

[167] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018.

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

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

[170] Vladimir Vovk. Competitive on-line linear regression. In *Advances in Neural Information Processing Systems*, 1997.

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

[172] Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly-linear (sometimes sublinear) running time. In *Mathematics of Operations Research*, 2019.

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

[174] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing 29 (NIPS)*, 2016.

[175] Chen-Yu Wei, Mehdi Jafarnia-Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. *arXiv:1910.07072*, 2019.

[176] Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *arXiv:2102.05406 [cs.LG]*, 2021.

[177] Lai Wei and Vaibhav Srivastava. On abruptly-changing and slowly-varying multi-armed bandit problems. In *Proceedings of Annual American Control Conference (ACC)*, 2018.

[178] World Health Organization (WHO). Coronavirus disease (covid-19) pandemic. 2020.

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

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

[181] Jia Yuan Yu and Shie Mannor. Online learning in markov decision processes with arbitrarily changing rewards and transitions. In *Proceedings of the International Conference on Game Theory for Networks*, 2009.

[182] Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

[183] Hao Yuan, Qi Luo, and Cong Shi. Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. In *SSRN Preprint 3329611*, 2019.

[184] Cun-Hui Zhang. Risk bounds in isotonic regression. In *Annals Statistics*, 2002.

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

[186] Huanan Zhang, Xiuli Chao, and Cong Shi. Closing the gap: A learning algorithm for the lost-sales inventory system with lead times. *Management Science*, 2018.

[187] Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

[188] Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. In *https://arxiv.org/abs/1907.12340*, 2019.

[189] Xiang Zhou, Ningyuan Chen, Xuefeng Gao, and Yi Xiong. Regime switching bandits. *arXiv:2001.09390*, 2020.

[190] Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Delay-adaptive learning in generalized linear contextual bandits. In *ArXiv:2003.05174 [cs.LG]*, 2019.

[191] Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, 2019.

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

[193] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 928–935, 2003.

Appendix A

Proofs for Chapter 2

A.1 Proof of Theorem 1

First, let's review the lower bound of the linear bandit setting, which is related to ours except that the θ_t's do not vary across rounds, and are equal to the same (unknown) θ, *i.e.*, $\forall t \in [T]\ \theta_t = \theta$.

Lemma 35 ([125]). *For any $T_0 \geq \sqrt{d}/2$ and let $D = \left\{x \in \mathbb{R}^d : \|x\| \leq 1\right\}$, then there exists a $\theta \in \left\{\pm\sqrt{d/4T_0}\right\}^d$, such that the worst case regret of any algorithm for linear bandits with unknown parameter θ is $\Omega(d\sqrt{T_0})$.*

Going back to the non-stationary environment, suppose nature divides the whole time horizon into $\lceil T/H \rceil$ blocks of equal length H rounds (the last block can possibly have less than H rounds), and each block is a decoupled linear bandit instance so that the knowledge of previous blocks cannot help the decision within the current block. Following Lemma 35, we restrict the sequence of θ_t's are drawn from the set $\left\{\pm\sqrt{d/4H}\right\}^d$. Moreover, θ_t's remain fixed within a block, and can vary across different blocks, *i.e.*,

$$\forall i \in \left[\left\lceil \frac{T}{H} \right\rceil\right] \forall t_1, t_2 \in [(i-1)H+1, i\cdot H \wedge T] \quad \theta_{t_1} = \theta_{t_2}. \tag{A.1}$$

We argue that even if the DM knows this additional information, it still incur a regret $\Omega(d^{2/3}B_T^{1/3}T^{2/3})$. Note that different blocks are completely decoupled, and information is

thus not passed across blocks. Therefore, the regret of each block is $\Omega\left(d\sqrt{H}\right)$, and the total regret is at least

$$\left(\left\lceil\frac{T}{H}\right\rceil-1\right)\Omega\left(d\sqrt{H}\right)=\Omega\left(dTH^{-\frac{1}{2}}\right). \tag{A.2}$$

Intuitively, if H, the number of length of each block, is smaller, the worst case regret lower bound becomes larger. But too small a block length can result in a violation of the variation budget. So we work on the total variation of θ_t's to see how small can H be. The total variation of the θ_t's can be seen as the total variation across consecutive blocks as θ_t remains unchanged within a single block. Observe that for any pair of $\theta,\theta'\in\left\{\pm\sqrt{d/4H}\right\}^d$, the ℓ_2 difference between θ and θ' is upper bounded as

$$\sqrt{\sum_{i=1}^{d}\frac{4d}{4H}}=\frac{d}{\sqrt{H}} \tag{A.3}$$

and there are at most $\lfloor T/H\rfloor$ changes across the whole time horizon, the total variation is at most

$$B=\frac{T}{H}\cdot\frac{d}{\sqrt{H}}=dTH^{-\frac{3}{2}}. \tag{A.4}$$

By definition, we require that $B\leq B_T$, and this indicates that

$$H\geq(dT)^{\frac{2}{3}}B_T^{-\frac{2}{3}}. \tag{A.5}$$

Taking $H=\left\lceil(dT)^{\frac{2}{3}}B_T^{-\frac{2}{3}}\right\rceil$, the worst case regret is

$$\Omega\left(dT\left((dT)^{\frac{2}{3}}B_T^{-\frac{2}{3}}\right)^{-\frac{1}{2}}\right)=\Omega\left(d^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}}\right). \tag{A.6}$$

Note that in order for $H\leq T$, we require $B_T\geq dT^{-1/2}$. Also, to make $|\langle x,\theta_t\rangle|\leq 1$ for all $t\in[T]$ and $x\in D_t$, we need $\|\theta_t\|\leq 1$, which means $\sqrt{d^2/4H}\leq 1$ or $B_T\leq 8d^{-2}T$.

A.2 Proof of Theorem 2

The difference $\hat{\theta}_t - \theta_t$ has the following expression:

$$\begin{aligned}
&V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \theta_s + \sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s\right) - \theta_t \\
=&V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) + V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s - \lambda\theta_t\right), \qquad \text{(A.7)}
\end{aligned}$$

The first term on the right hand side of eq. (A.7) is the estimation inaccuracy due to the non-stationarity; while the second term is the estimation error due to random noise. We now upper bound the two terms separately. We upper bound the first term under the Euclidean norm.

Lemma 36. *For any $t \in [T]$, we have*

$$\left\| V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t)\right\|_2 \le \sum_{s=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 .$$

Proof. Poof. In the proof, we denote $B(1)$ as the unit Euclidean ball, and $\lambda_{\max}(M)$ as the maximum eigenvalue of a square matrix M. By folklore, we know that $\lambda_{\max}(M) = \max_{z\in B(1)} z^\top M z$. In addition, recall the definition that $V_{t-1} = \lambda I + \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top$ We prove the Lemma as follows:

$$\begin{aligned}
&\left\| V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t)\right\|_2 \\
=&\left\| V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \left[\sum_{p=s}^{t-1}(\theta_p - \theta_{p+1})\right]\right\|_2 \\
=&\left\| V_{t-1}^{-1}\sum_{p=1\vee(t-w)}^{t-1}\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top (\theta_p - \theta_{p+1})\right\|_2 \qquad \text{(A.8)} \\
\le&\sum_{p=1\vee(t-w)}^{t-1}\left\| V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top\right)(\theta_p - \theta_{p+1})\right\|_2 \qquad \text{(A.9)}
\end{aligned}$$

$$\leq \sum_{p=1\vee(t-w)}^{t-1} \lambda_{\max}\left(V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top\right)\right)\left\|\theta_p-\theta_{p+1}\right\|_2 \tag{A.10}$$

$$\leq \sum_{p=1\vee(t-w)}^{t-1} \left\|\theta_p-\theta_{p+1}\right\|_2. \tag{A.11}$$

Equality (A.8) is by the observation that both sides of the equation is summing over the terms $X_s X_s^\top(\theta_p-\theta_{p+1})$ with indexes (s,p) ranging over $\{(s,p): 1\vee(t-w)\leq s\leq p\leq t-1\}$. Inequality (A.9) is by the triangle inequality.

Inequality (A.10) is by the fact that, for any matrix $M\in\mathbb{R}^{d\times d}$ with $\lambda_{\max}(M)\geq 0$ and any vector $y\in\mathbb{R}^d$, we have $\|My\|_2\leq\sqrt{\lambda_{\max}(M^2)}\,\|y\|_2$. Applying the above claim with $M=V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top\right)$ and $y=\theta_p-\theta_{p+1}$ demonstrates inequality (A.10).

Finally, for inequality (A.11), we denote the corresponding basis for each X_s as $\psi_{i(s)}$, *i.e.*, $X_s=z_s\psi_{i(s)}=z_s\Psi e_{i(s)}$, where e_i is the i^{th} standard orthonormal basis. Let $A_1=\sum_{s=1\vee(t-w)}^{t-1} e_{i(s)}e_{i(s)}^\top+\lambda I$ and $A_2=\sum_{s=1\vee(t-w)}^{p} e_{i(s)}e_{i(s)}^\top$, it is evident that $V_{t-1}=\Psi A_1\Psi^\top$ and $\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top=\Psi A_2\Psi^\top$. Therefore, we have

$$\begin{aligned}\lambda_{\max}\left(\left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top\right)V_{t-1}^{-2}\left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top\right)\right) &= \lambda_{\max}\left(\Psi A_2\Psi^\top(\Psi A_1\Psi^\top)^{-2}\Psi A_2\Psi^\top\right)\\ &= \lambda_{\max}\left(\Psi A_2 A_1^{-2}A_2\Psi^\top\right)=\lambda_{\max}\left(A_2A_1^{-2}A_2\right)\leq 1,\end{aligned} \tag{A.12}$$

where we have used the fact that both A_1 and A_2 are diagonal matrix in the last step. Altogether, the Lemma is proved. □

Applying Theorem 2 of [3], we have the following upper bound for the second term in eq. (2.2).

Lemma 37 ([3]). *For any $t\in[T]$ and any $\delta\in[0,1]$, we have*

$$\left\|\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s-\lambda\theta_t\right\|_{V_{t-1}^{-1}}\leq R\sqrt{d\ln\left(\frac{1+wL^2/\lambda}{\delta}\right)}+\sqrt{\lambda}S$$

holds with probability at least $1-\delta$.

Combining the above two lemmas: fixed any $\delta\in[0,1]$, we have that for any $t\in[T]$ and

any $x \in D_t$,

$$\begin{aligned}
\left|x^\top(\hat{\theta}_t - \theta_t)\right| &= \left|x^\top\left(V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right) + x^\top V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s - \lambda\theta_t\right)\right| \\
&\leq \left|x^\top\left(V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right)\right| + \left|x^\top V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s - \lambda\theta_t\right)\right| \qquad \text{(A.13)} \\
&\leq \|x\|_2 \cdot \left\|V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right\|_2 + \|x\|_{V_{t-1}^{-1}}\left\|\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s - \lambda\theta_t\right\|_{V_{t-1}^{-1}} \qquad \text{(A.14)} \\
&\leq L\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s - \theta_{s+1}\|_2 + \beta\|x\|_{V_{t-1}^{-1}}, \qquad \text{(A.15)}
\end{aligned}$$

where inequality (A.13) uses triangle inequality, inequality (A.14) follows from Cauchy-Schwarz inequality, and inequality (A.15) are consequences of Lemmas 36, 37.

A.3 Proof of Theorem 3

In the proof, we choose λ so that $\beta \geq 1$, for example by choosing $\lambda \geq 1/S^2$. By virtue of UCB, the regret in any round $t \in [T]$ is

$$\langle x_t^* - X_t, \theta_t \rangle \leq L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + \langle X_t, \hat{\theta}_t \rangle + \beta \|X_t\|_{V_{t-1}^{-1}} - \langle X_t, \theta_t \rangle \tag{A.16}$$

$$\leq 2L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + 2\beta \|X_t\|_{V_{t-1}^{-1}}. \tag{A.17}$$

Inequality (A.16) is by an application of our `SW-UCB` algorithm established in equation (2.9). Inequality (A.17) is by an application of inequality (A.15), which bounds the difference $|\langle X_t, \hat{\theta}_t - \theta_t \rangle|$ from above. By the assumption $|\langle X, \theta_t \rangle| \leq 1$ in Section 2.1, it is evident that $\langle X_t, \hat{\theta}_t - \theta_t \rangle \leq |\langle X_t, \hat{\theta}_t \rangle| + |\langle X_t, -\theta_t \rangle| \leq 2$, and we have

$$\langle x_t^* - X_t, \theta_t \rangle \leq 2L \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2 + 2\beta \left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1 \right). \tag{A.18}$$

Summing equation (A.18) over $1 \leq t \leq T$, the regret of the `SW-UCB` algorithm is upper bounded as

$$\begin{aligned}
\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB algorithm}\right)\right] =& \mathbf{E}\left[\sum_{t \in [T]} \langle x_t^* - X_t, \theta_t \rangle\right] \\
\leq& 2L \left[\sum_{t=1}^{T} \sum_{s=1 \vee (t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_2\right] + 2\beta \cdot \mathbf{E}\left[\sum_{t=1}^{T} \left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1 \right)\right] \\
=& 2L \left[\sum_{s=1}^{T} \sum_{t=s+1}^{(s+w) \wedge T} \|\theta_s - \theta_{s+1}\|_2\right] + 2\beta \cdot \mathbf{E}\left[\sum_{t=1}^{T} \left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1 \right)\right] \\
\leq& 2LwB_T + 2\beta \cdot \mathbf{E}\left[\sum_{t=1}^{T} \left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1 \right)\right].
\end{aligned} \tag{A.19}$$

What's left is to upper bound the quantity $2\beta \cdot \mathbf{E}\left[\sum_{t \in [T]} \left(1 \wedge \|X_t\|_{V_{t-1}^{-1}}\right)\right]$. Following the trick introduced by the authors of [3], we apply Cauchy-Schwarz inequality to the term

$\sum_{t\in[T]}\left(1\wedge\|X_t\|_{V_{t-1}^{-1}}\right)$.

$$\sum_{t\in[T]}\left(1\wedge\|X_t\|_{V_{t-1}^{-1}}\right)\leq\sqrt{T}\sqrt{\sum_{t\in[T]}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}}. \tag{A.20}$$

By dividing the whole time horizon into consecutive pieces of length w, we have

$$\sqrt{\sum_{t\in[T]}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}}\leq\sqrt{\sum_{i=0}^{\lceil T/w\rceil-1}\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}}. \tag{A.21}$$

While a similar quantity has been analyzed by Lemma 11 of [3], we note that due to the fact that V_t's are accumulated according to the sliding window principle, the key eq. (6) in Lemma 11's proof breaks, and thus the analysis of [3] cannot be applied here. To this end, we state a technical lemma based on the Sherman-Morrison formula.

Lemma 38. *For any $i\leq\lceil T/w\rceil-1$,*

$$\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}\leq\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{\overline{V}_{t-1}^{-1}},$$

where

$$\overline{V}_{t-1}=\sum_{s=i\cdot w+1}^{t-1}X_sX_s^\top+\lambda I. \tag{A.22}$$

Proof. Proof of Lemma 38. For a fixed $i\leq\lceil T/w\rceil-1$,

$$\begin{aligned}\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}&=\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge X_t^\top V_{t-1}^{-1}X_t\\&=\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge X_t^\top\left(\sum_{s=1\vee(t-w)}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}X_t.\end{aligned} \tag{A.23}$$

Note that $i\cdot w+1\geq 1$ and $i\cdot w+1\geq t-w\ \forall t\leq(i+1)w$, we have

$$i\cdot w+1\geq 1\vee(t-w). \tag{A.24}$$

Consider any d-by-d positive definite matrix A and d-dimensional vector y, then by the Sherman-Morrison formula, the matrix

$$B = A^{-1} - \left(A + yy^\top\right)^{-1} = A^{-1} - A^{-1} + \frac{A^{-1}yy^\top A^{-1}}{1 + y^\top A^{-1} y} = \frac{A^{-1}yy^\top A^{-1}}{1 + y^\top A^{-1} y} \tag{A.25}$$

is positive semi-definite. Therefore, for a given t, we can iteratively apply this fact to obtain

$$\begin{aligned}
&X_t^\top \left(\sum_{s=i\cdot w+1}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t \\
=&X_t^\top \left(\sum_{s=i\cdot w}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t + X_t^\top \left(\left(\sum_{s=i\cdot w+1}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} - \left(\sum_{s=i\cdot w}^{t-1} X_s X_s^\top + \lambda I\right)^{-1}\right) X_t \\
=&X_t^\top \left(\sum_{s=i\cdot w}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t \\
&+ X_t^\top \left(\left(\sum_{s=i\cdot w+1}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} - \left(X_{i\cdot w} X_{i\cdot w}^\top + \sum_{s=i\cdot w+1}^{t-1} X_s X_s^\top + \lambda I\right)^{-1}\right) X_t \\
\geq&X_t^\top \left(\sum_{s=i\cdot w}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t \\
&\vdots \\
\geq&X_t^\top \left(\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t .
\end{aligned} \tag{A.26}$$

Plugging inequality (A.26) to (A.23), we have

$$\begin{aligned}
\sum_{t=i\cdot w+1}^{(i+1)w} 1 \wedge \|X_t\|^2_{V_{t-1}^{-1}} &\leq \sum_{t=i\cdot w+1}^{(i+1)w} 1 \wedge X_t^\top \left(\sum_{s=i\cdot w+1}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t \\
&\leq \sum_{t=i\cdot w+1}^{(i+1)w} 1 \wedge \|X_t\|^2_{\overline{V}_{t-1}^{-1}} ,
\end{aligned} \tag{A.27}$$

which concludes the proof. □

From Lemma 38 and eq. (A.21), we know that

$$\begin{aligned}
2\beta \sum_{t\in[T]} \left(1 \wedge \|X_t\|_{V_{t-1}^{-1}}\right) \leq & 2\beta\sqrt{T} \cdot \sqrt{\sum_{i=0}^{\lceil T/w\rceil-1} \sum_{t=i\cdot w+1}^{(i+1)w} 1 \wedge \|X_t\|^2_{\overline{V}_{t-1}^{-1}}} \\
\leq & 2\beta\sqrt{T} \cdot \sqrt{\sum_{i=0}^{\lceil T/w\rceil-1} 2d\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)} \qquad \text{(A.28)} \\
\leq & 2\beta T\sqrt{\frac{2d}{w}\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)}.
\end{aligned}$$

Here, eq. (A.28) follows from Lemma 11 of [3].

Now putting these two parts to eq. (A.19), we have

$$\begin{aligned}
&\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB} \text{ algorithm}\right)\right] \\
\leq & 2LwB_T + 2\beta T\sqrt{\frac{2d}{w}\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)} + 2T\delta \\
= & 2LwB_T + \frac{2T}{\sqrt{w}}\left(R\sqrt{d\ln\left(\frac{1+wL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S\right)\sqrt{2d\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)} + 2T\delta. \qquad \text{(A.29)}
\end{aligned}$$

Now if B_T is known, we can take $w = \Theta\left((dT)^{2/3}B_t^{-2/3}\right)$ and $\delta = 1/T$, we have

$$\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB} \text{ algorithm}\right)\right] = \widetilde{O}\left(d^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}}\right);$$

while if B_T is not unknown, taking $w = \Theta\left((dT)^{2/3}\right)$ and $\delta = 1/T$, we have

$$\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB} \text{ algorithm}\right)\right] = \widetilde{O}\left(d^{\frac{2}{3}}B_T T^{\frac{2}{3}}\right).$$

A.4 Proof of Lemma 4

For any block i, the absolute sum of rewards can be written as

$$\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t,\theta_t\rangle+\eta_t\right|\leq\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}|\langle X_t,\theta_t\rangle|+\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|\leq H\nu+\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|,$$

where we have iteratively applied the triangle inequality as well as the fact that $|\langle X_t,\theta_t\rangle|\leq\nu$ for all t.

Now by property of the R-sub-Gaussian [154], we have the absolute value of the noise term η_t exceeds $2R\sqrt{\ln T}$ for a fixed t with probability at most $1/T^2$ *i.e.*,

$$\Pr\left(\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|\geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right)\leq\frac{2H}{T^2}. \tag{A.30}$$

Applying a simple union bound, we have

$$\begin{aligned}&\Pr\left(\exists i\in\left\lceil\frac{T}{H}\right\rceil:\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|\geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right)\\&\leq\sum_{i=1}^{\lceil T/H\rceil}\Pr\left(\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|\geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right)\leq\frac{2}{T}.\end{aligned} \tag{A.31}$$

Therefore, we have

$$\Pr\left(Q\geq H\nu+2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right)\leq\Pr\left(\exists i\in\left\lceil\frac{T}{H}\right\rceil:\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|\geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right)\leq\frac{2}{T}. \tag{A.32}$$

The statement then follows.

A.5 Proof of Proposition 5

By design of the BOB algorithm, its dynamic regret can be decomposed as the regret of the SW-UCB algorithm with the optimally tuned window size $w_i = w^\dagger$ for each block i plus the loss due to learning the value $w^\dagger$ with the EXP3 algorithm, *i.e.*,

$$\begin{aligned}\mathbf{E}\left[\text{Regret}_T(\texttt{BOB algorithm})\right] =& \mathbf{E}\left[\sum_{t=1}^{T}\langle x_t^*, \theta_t\rangle - \sum_{t=1}^{T}\langle X_t, \theta_t\rangle\right] \\ =& \mathbf{E}\left[\sum_{t=1}^{T}\langle x_t^*, \theta_t\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger}, \theta_t\right\rangle\right] \\ &+ \mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger}, \theta_t\right\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i}, \theta_t\rangle\right].\end{aligned} \tag{A.33}$$

Here, eq. (A.33) holds as the BOB algorithm restarts the SW-UCB algorithm in each block, and for a round t in block i, X_t^w refers to the action selected in round t by the SW-UCB algorithm with window size $w \wedge (t-(i-1)H-1)$ initiated at the beginning of block i.

By Theorem 3, the first expectation in eq. (A.33) can be upper bounded as

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T}\langle x_t^*, \theta_t\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger}, \theta_t\right\rangle\right] =& \mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle x_t^* - X_t^{w^\dagger}, \theta_t\right\rangle\right] \\ =& \sum_{i=1}^{\lceil T/H\rceil}\widetilde{O}\left(w^\dagger B_T(i) + \frac{dH}{\sqrt{w^\dagger}}\right) \\ =& \widetilde{O}\left(w^\dagger B_T + \frac{dT}{\sqrt{w^\dagger}}\right),\end{aligned} \tag{A.34}$$

where

$$B_T(i) = \sum_{t=(i-1)H+1}^{(i\cdot H\wedge t)-1}\|\theta_t - \theta_{t+1}\|_2$$

is the total variation in block i.

We then turn to the second expectation in eq. (A.33). We can easily see that the number of rounds for the EXP3 algorithm is $\lceil T/H\rceil$ and the number of possible values of w_i's is $|J|$. If the maximum absolute sum of reward of any block does not exceed Q, the authors of

[18] gives the following regret bound.

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger},\theta_t\right\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i},\theta_t\rangle \,\middle|\, \forall i\in[\lceil T/H\rceil]\ \sum_{t=(i-1)H+1}^{i\cdot H\wedge T} Y_t \le Q/2\right] \\
=&\widetilde{O}\left(Q\sqrt{\frac{|J|T}{H}}\right). \qquad \text{(A.35)}
\end{aligned}$$

Note that the regret of our problem is at most T, eq. (A.35) can be further upper bounded as

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger},\theta_t\right\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i},\theta_t\rangle\right] \\
\le&\widetilde{O}\left(Q\sqrt{\frac{|J|T}{H}}\right)\times\Pr\left(\forall i\in[\lceil T/H\rceil]\ \sum_{t=(i-1)H+1}^{i\cdot H\wedge T} Y_t\le Q/2\right) \\
&+\mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger},\theta_t\right\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i},\theta_t\rangle \,\middle|\, \exists i\in[\lceil T/H\rceil]\ \sum_{t=(i-1)H+1}^{i\cdot H\wedge T} Y_t\ge Q/2\right] \\
&\quad\times\Pr\left(\exists i\in[\lceil T/H\rceil]\ \sum_{t=(i-1)H+1}^{i\cdot H\wedge T} Y_t\ge Q/2\right) \\
\le&\widetilde{O}\left(\sqrt{H|J|T}\right)+T\cdot\frac{2}{T} \\
=&\widetilde{O}\left(\sqrt{H|J|T}\right). \qquad \text{(A.36)}
\end{aligned}$$

Combining eq. (A.33), (A.34), and (A.36), the statement follows.

A.6 Proof of Theorem 6

With Proposition 5 as well as the choices of H and J in eq. (2.12), the regret of the BOB algorithm is

$$\mathscr{R}_T(\texttt{BOB algorithm}) = \widetilde{O}\left(w^\dagger B_T + \frac{dT}{\sqrt{w^\dagger}} + \sqrt{H|J|T}\right) = \widetilde{O}\left(w^\dagger B_T + \frac{dT}{\sqrt{w^\dagger}} + d^{\frac{1}{2}}T^{\frac{3}{4}}\right). \tag{A.37}$$

Therefore, we have that when $B_T \geq d^{-1/2}T^{1/4}$, the BOB algorithm is able to converge to the optimal window size, *i.e.*, $w^\dagger = w^*$ $(\leq H)$, and the dynamic regret of the BOB algorithm is upper bounded as

$$\mathscr{R}_T(\texttt{BOB algorithm}) = \widetilde{O}\left(d^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}} + d^{\frac{1}{2}}T^{\frac{3}{4}}\right); \tag{A.38}$$

while if $B_T < d^{-1/2}T^{1/4}$, the BOB algorithm converges to the window size $w^\dagger = H$, and the dynamic regret is

$$\mathscr{R}_T(\texttt{BOB algorithm}) = \widetilde{O}\left(dB_T T^{\frac{1}{2}} + d^{\frac{1}{2}}T^{\frac{3}{4}}\right) = \widetilde{O}\left(d^{\frac{1}{2}}T^{\frac{3}{4}}\right). \tag{A.39}$$

Combining the above two cases, we conclude the desired dynamic regret bound.

A.7 Proof of Theorem 7

Similar to eq. (A.7), we can rewrite the difference $\hat{\theta}_t - \theta_t$ as

$$\begin{aligned} & V_{t-1}^* \left(\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \theta_s + \sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s \right) - \theta_t \\ =& V_{t-1}^* \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) + V_{t-1}^* \left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s \right). \end{aligned} \tag{A.40}$$

We then analyze the two terms in eq. (A.40) separately. For the first term,

$$\begin{aligned} \left\| V_{t-1}^* \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_\infty &= \left\| V_{t-1}^* \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \left[\sum_{p=s}^{t-1} (\theta_p - \theta_{p+1}) \right] \right\|_\infty \\ &= \left\| \sum_{p=1\vee(t-w)}^{t-1} \left[V_{t-1}^* \sum_{s=1\vee(t-w)}^{p} X_s X_s^\top (\theta_p - \theta_{p+1}) \right] \right\|_\infty \\ &\leq \sum_{p=1\vee(t-w)}^{t-1} \left\| V_{t-1}^* \sum_{s=1\vee(t-w)}^{p} X_s X_s^\top (\theta_p - \theta_{p+1}) \right\|_\infty \\ &\leq \sum_{p=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_\infty . \end{aligned} \tag{A.41}$$

Here, almost all the steps follow exactly the same arguments as those of eq. (A.8)-(A.11), except that in inequality (A.41), we make the direct observation that

$$V_{t-1}^* = \begin{pmatrix} \frac{\mathbf{1}[N_{t-1}(1)>0]}{N_{t-1}(1)} & 0 & \dots & \dots & \dots & 0 \\ 0 & \frac{\mathbf{1}[N_{t-1}(2)>0]}{N_{t-1}(2)} & 0 & \dots & \dots & 0 \\ 0 & 0 & \ddots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{\mathbf{1}[N_{t-1}(d-1)>0]}{N_{t-1}(d-1)} & 0 \\ 0 & 0 & 0 & \dots & 0 & \frac{\mathbf{1}[N_{t-1}(d)>0]}{N_{t-1}(d)} \end{pmatrix} \tag{A.42}$$

and

$$\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top = \begin{pmatrix} N_p'(1) & 0 & \dots & \dots & \dots & 0 \\ 0 & N_p'(2) & 0 & \dots & \dots & 0 \\ 0 & 0 & \ddots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & N_p'(d-1) & 0 \\ 0 & 0 & 0 & \dots & 0 & N_p'(d) \end{pmatrix}, \tag{A.43}$$

where $N_p'(i)$ is the number of times that action e_i is selected during rounds $1 \vee (t-w), \dots, p$ for all $i \in [d]$. As $p \le t-1$, we have $N_p'(i) \le N_{t-1}(i)$ for all $i \in [d]$. Now, $V_{t-1}^* \sum_{s=1\vee(t-w)}^{p} X_s X_s^\top$ is a diagonal matrix with all diagonal entries less than 1, and hence the argument.

For the second term of eq. (A.40), we consider for any fixed $i \in [d]$,

$$\begin{aligned} \left| e_i^\top V_{t-1}^* \left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s \right) \right| &= \frac{\mathbf{1}[N_{t-1}(i) > 0]}{N_{t-1}(i)} \left| e_i^\top \left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s \right) \right| \\ &= \frac{\mathbf{1}[N_{t-1}(i) > 0] \left(\sum_{s=1\vee(t-w)}^{t-1} \mathbf{1}[I_s = i] \eta_s \right)}{N_{t-1}(i)}, \end{aligned} \tag{A.44}$$

where the first step again use the definition of V_{t-1}^* in eq. (A.42). Now if $N_{t-1}(i) = 0$, eq. (A.44) equals to 0; while if $N_{t-1}(i) > 0$, we can apply the Corollary 1.7 of [154] to obtain that

$$\Pr\left(\left| \frac{\mathbf{1}[N_{t-1}(i) > 0] \left(\sum_{s=1\vee(t-w)}^{t-1} \mathbf{1}[I_s = i] \eta_s \right)}{N_{t-1}(i)} \right| \le R \sqrt{\frac{2 \ln(2dT^2)}{N_{t-1}(i)}} \right) \ge 1 - \frac{1}{dT^2}. \tag{A.45}$$

Hence, with probability at least $1 - 1/dT^2$, for any fixed $t \in [T]$ and any fixed $i \in [d]$,

$$\begin{aligned} \left| e_i^\top (\hat{\theta}_t - \theta_t) \right| &= \left| e_i^\top \left(V_{t-1}^* \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \right) + e_i^\top V_{t-1}^* \left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s - \lambda \theta_t \right) \right| \\ &\le \left| e_i^\top \left(V_{t-1}^* \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \right) \right| + \left| e_i^\top V_{t-1}^* \left(\sum_{s=1\vee(t-w)}^{t-1} \eta_s X_s - \lambda \theta_t \right) \right| \end{aligned} \tag{A.46}$$

$$\leq \|e_i\|_1 \cdot \left\| V_{t-1}^* \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_\infty + R\sqrt{\frac{2\ln(2dT^2)}{N_{t-1}(i)}} \tag{A.47}$$

$$\leq \sum_{s=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_\infty + R\sqrt{\frac{2\ln(2dT^2)}{N_{t-1}(i)}}, \tag{A.48}$$

where inequality (A.46) applies the triangle inequality, inequality (A.47) follows from the Holder's inequality as well as inequality (A.44) and (A.45), and inequality (A.48) follows from inequality (A.41).

The statement of the theorem now follows immediately by applying union bound over the decision set and the time horizon as well as the simple observation $\|e_i\|_{V_{t-1}^*} = \sqrt{1/N_{t-1}(i)}$.

A.8 Proof of Theorem 10

From the proof of Proposition 1 in [79], we know that for all $x \in D$

$$\left|\mu\left(\langle x,\theta_t\rangle\right)-\mu\left(\left\langle x,\hat{\theta}_t\right\rangle\right)\right| \leq k_\mu \left| x^\top G_{t-1}^{-1}\left[\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\left\langle X_s,\hat{\theta}_t\right\rangle\right)\right)X_s\right]\right|, \tag{A.49}$$

where

$$G_{t-1}=\int_0^1\left[\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \mu\left(\left\langle X_s, s_0\theta_t+(1-s_0)\hat{\theta}_t\right\rangle\right)\right] ds_0$$

By virtue of the maximum quasi-likelihood estimation, *i.e.*, eq. (2.25) we have

$$\sum_{s=1\vee(t-w)}^{t-1}\mu\left(\left\langle X_s,\hat{\theta}_t\right\rangle\right)X_s=\sum_{s=1\vee(t-w)}^{t-1}Y_s X_s=\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_s\rangle\right)+\eta_s\right)X_s, \tag{A.50}$$

and (A.49) is

$$\begin{aligned}
&k_\mu\left|x^\top G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\langle X_s,\theta_s\rangle\right)-\eta_s\right)X_s\right| \\
=&k_\mu\left|x^\top G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\langle X_s,\theta_s\rangle\right)\right)X_s-x^\top G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s\right| \\
\leq&k_\mu\left|x^\top G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\langle X_s,\theta_s\rangle\right)\right)X_s\right|+k_\mu\left|x^\top G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s\right| \qquad \text{(A.51)} \\
\leq&k_\mu\left|x^\top G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\langle X_s,\theta_s\rangle\right)\right)X_s\right|+\beta\|x\|_{V_{t-1}^{-1}} \qquad \text{(A.52)} \\
\leq&k_\mu\|x\|_2\left\|G_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\langle X_s,\theta_s\rangle\right)\right)X_s\right\|_2+\beta\|x\|_{V_{t-1}^{-1}} \qquad \text{(A.53)} \\
\leq&\frac{k_\mu L}{c_\mu}\left\|V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1}\left(\mu\left(\langle X_s,\theta_t\rangle\right)-\mu\left(\langle X_s,\theta_s\rangle\right)\right)X_s\right\|_2+\beta\|x\|_{V_{t-1}^{-1}}.
\end{aligned}$$

Here, inequality (A.51) is a consequence of the triangle inequality, inequality (A.52) again follows from Proposition 1 of [79], inequality (A.53) is the Cauchy-Schwarz inequality,

and the last step uses the fact that $G_{t-1} \succeq c_\mu V_{t-1}$. For the firs quantity, we have

$$
\begin{aligned}
&\left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{t-1} \left(\mu\left(\langle X_s, \theta_t\rangle\right) - \mu\left(\langle X_s, \theta_s\rangle\right)\right) X_s \right\|_2 \\
=&\left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{t-1} X_s \sum_{p=s}^{t-1} \left(\mu\left(\langle X_s, \theta_{p+1}\rangle\right) - \mu\left(\langle X_s, \theta_p\rangle\right)\right) \right\|_2 \\
=&\left\| V_{t-1}^{-1} \sum_{p=1\vee(t-w)}^{t-1} \sum_{s=1\vee(t-w)}^{p} X_s \left(\mu\left(\langle X_s, \theta_{p+1}\rangle\right) - \mu\left(\langle X_s, \theta_p\rangle\right)\right) \right\|_2 \\
\leq& \sum_{p=1\vee(t-w)}^{t-1} \left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{p} X_s \left(\mu\left(\langle X_s, \theta_{p+1}\rangle\right) - \mu\left(\langle X_s, \theta_p\rangle\right)\right) \right\|_2 \qquad \text{(A.54)} \\
=& \sum_{p=1\vee(t-w)}^{t-1} \left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{p} X_s \dot{\mu}\left(\langle X_s, \tilde{\theta}_p\rangle\right) X_s^\top \left(\theta_{p+1} - \theta_p\right) \right\|_2 \qquad \text{(A.55)} \\
=& \sum_{p=1\vee(t-w)}^{t-1} \left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{p} \dot{\mu}\left(\langle X_s, \tilde{\theta}_p\rangle\right) X_s X_s^\top \left(\theta_{p+1} - \theta_p\right) \right\|_2 \\
=& \sum_{p=1\vee(t-w)}^{t-1} \lambda_{\max}\left(V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{p} \dot{\mu}\left(\langle X_s, \tilde{\theta}_p\rangle\right) X_s X_s^\top \right) \left\| \left(\theta_{p+1} - \theta_p\right) \right\|_2 \qquad \text{(A.56)} \\
\leq& k_\mu \sum_{p=1\vee(t-w)}^{t-1} \lambda_{\max}\left(V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) \left\| \left(\theta_{p+1} - \theta_p\right) \right\|_2 \\
\leq& k_\mu \sum_{p=1\vee(t-w)}^{t-1} \left\| \left(\theta_{p+1} - \theta_p\right) \right\|_2, \qquad \text{(A.57)}
\end{aligned}
$$

where inequality (A.54) is an immediate consequence of the triangle inequality, eq. (A.55) utilizes the mean value theorem (with $\tilde{\theta}_p$ being some certain linear combination of θ_p and θ_{p+1} for all p), and inequalities (A.56) and (A.57) follow from the same steps as the proof of Lemma 36 in Section A.2.

A.9 Proof of Theorem 13

We start with a regret lower bound result from [38] on drifting K-armed bandits:

Theorem 39 ([38])**.** *Consider the drifting K-armed bandit problem, where $K \geq 2$, with $T \geq 1$ rounds. For any $B_T \in [1/K, T/K]$, there exists a finite class of reward distributions $\tilde{\mathscr{P}} = \{\tilde{P}^{(\ell)}\}_{\ell=1}^{L}$, where $\tilde{P}^{(\ell)} = \{\tilde{P}_{t,k}^{(\ell)}\}_{t\in[T],k\in[K]}$, that satisfy the following:*

- *Each $\tilde{P}_{t,k}^{(\ell)}$ represents the reward distribution of arm k in round t under distribution $\tilde{P}^{(\ell)}$. For each ell, t, k, the distribution $\tilde{P}_{t,k}^{(\ell)}$ is a Bernoulli distribution, with the mean denoted $\tilde{\theta}_{t,k}^{(\ell)}$.*

- *For every $\ell \in [L]$, the following variational budget inequality holds:*

$$\sum_{t=1}^{T-1} \max_{k\in[K]} \left\{ \left| \tilde{\theta}_{t+1}^{(\ell)}(k) - \tilde{\theta}_{t}^{(\ell)}(k) \right| \right\} \leq B_T.$$

- *For any non-anticipatory policy $\tilde{\pi}$, there exists $\ell \in [L]$ under which the dynamic regret is lower bounded:*

$$\sum_{t=1}^{T} \left\{ \max_{k\in[K]} \tilde{\theta}_{t}^{(\ell)}(k) - \mathbb{E}[\tilde{\theta}_{t}^{(\ell)}(I_t)] \right\} \geq \frac{1}{4\sqrt{2}} (KB_T)^{1/3} T^{2/3}.$$

  We denote the choice of arm under policy $\tilde{\pi}$ in round t as I_t, and the expectation is taken over the randomness in the choice of I_t, which is caused by the previous outcomes and the policy's internal randomness.

We prove the Theorem by modifying the class of instances $\mathscr{P}$ to suit the setting of drifting combinatorial semi-bandits. The modification follows the style of Kveton et al. [123]. Let d, m be two integers, where d is divisible by m W.L.O.G.. We define the ground set $E = [d]$. In addition, we define the action set $\mathscr{E}_t = \{a_1, \ldots, a_{d/m}\} \subset \{0,1\}^d$, which contains d/m combinatorial arms and does not vary with t. Each combinatorial arm a_i belongs to $\{0,1\}^d$. For each $1 \leq i \leq d/m$, we define $a_i(j) = 1$ if $(i-1)m+1 \leq j \leq i \cdot m$, and $a_i(j) = 0$ for other j.

Consider Theorem 39 when $K = d/m \geq 2$, and let $\tilde{\mathscr{P}} = \{\tilde{P}^{(\ell)}\}_{\ell=1}^{L}$ be the class of reward distributions for the regret lower bound. For each $\tilde{P}_\ell = \{\tilde{P}_{t,k}^{(\ell)}\}_{t\in[T],k\in[K]}$ (which is on the $K = d/m$-armed bandit instance), we construct another reward distribution $P_\ell = \{P_{t,j}^{(\ell)}\}_{t\in[T],j\in[d]}$ that is defined on the combinatorial semi-bandit instance. For each $j \in [d]$, we identify the index $i \in [d/m]$ such that $(i-1)m+1 \leq j \leq i\cdot m$, and define $P_{t,j}^{(\ell)}$ to be the same distribution as $\tilde{P}_{t,i}^{(\ell)}$. That is, $P_{t,j}^{(\ell)}$ is a Bernoulli distribution with mean $\theta_t(j) = \tilde{\theta}_t(i)$, where $i = \lceil j/m \rceil$. By the second property in Theorem 39, it is straightforward to check that B_T is also a variation budget for $P^{(\ell)}$ for each ℓ, that is,

$$\sum_{t=1}^{T-1} \max_{j\in[d]} \left\{ \left| \theta_{t+1}^{(\ell)}(j) - \theta_t^{(\ell)}(j) \right| \right\} \leq B_T.$$

For each $1 \leq i \leq d/m$, the random rewards $W_t((i-1)m+1), \ldots, W_t(i\cdot m)$ for the items in combinatorial arm i are identical Bernoulli random variables. That is, they simultaneously realize as all ones or all zeros.

Finally, to complete the proof, we relate the dynamic regret of any non-anticipatory policy π on the drifting combinatorial semi-bandit instance to that of some non-anticipatory policy $\tilde{\pi}$ on the drifting K-armed instance. For the combinatorial bandit instance, a non-anticipatory policy π is in fact a sequence of mappings $\{\pi_t\}_{t=1}^{\infty}$, where π_t maps the historical information $H_{t-1} = \{X_s, \{W_s(i)\}_{i\in X_s}\}_{s=1}^{t-1}$ from time 1 to $t-1$ and a random seed U to the combinatorial arm X_t to pull in time t, or more mathematically $\pi_t(H_{t-1}, U) = X_t$. Likewise is true for any non-anticipatory policy $\tilde{\pi}$ for a K-armed instance.

Given a non-anticipatory policy π for the combinatorial semi-bandit instance, we construct another non-anticipatory policy $\tilde{\pi}$ for the K-armed bandit instance that mimics the behaviour of π. Suppose that $\pi_t(H,U) = X_j$ for a realization of the history $H = \{X_s, \{W_s(i)\}_{i\in X_s}\}_{s=1}^{t-1}$ and random seed U. To construct $\tilde{\pi}$, we map the H to the historical information $\tilde{H}$ for the K-armed bandit instance, where $\tilde{H} = \{\tilde{X}_s, \tilde{W}_s\}_{s=1}^{t-1}$ is defined as follows: $\tilde{X}_s = i$ iff $X_s = a_i$, and $\tilde{W}_s = \frac{1}{m}\sum_{i\in[d]} X_s(i)W_s(i)$. It is clear that $\tilde{W}_s \in \{0,1\}$ for each s, by our assumption on the correlations among $\{W_t(i)\}_{i\in[d]}$. Finally, we define $\tilde{\pi}_t(\tilde{H}, U) = i$ if and only if $\pi_t(H,U) = a_i$. It is evident from our construction that π_t is well-defined, in the sense that it maps to a unique arm for every possible realization of $\tilde{H}, U$. Importantly, for any $1 \leq \ell \leq L$, we know

that

$$\text{Expected reward of } \pi \text{ under } P^{(\ell)} = m \times \text{Expected reward of } \tilde{\pi} \text{ under } \tilde{P}^{(\ell)},$$
$$\text{Optimal expected reward under } P^{(\ell)} = m \times \text{Optimal expected reward under } \tilde{P}^{(\ell)},$$

or more mathematically we have $\sum_{t=1}^{T} \max_{a_i \in \mathscr{O}_t} \sum_{j:a_i(j)=1} \theta_t^{(\ell)}(j) = m \times \sum_{t=1}^{T} \max_{k \in [K]} \tilde{\theta}_t^{(\ell)}(k)$. Consequently, by the third property of Theorem 39, we know that for any non-anticipatory policy π, there is an index ℓ such that the dynamic regret of π under $P^{(\ell)}$ is at least $m \times (\frac{1}{4\sqrt{2}}(\frac{d}{m}B_T)^{1/3}T^{2/3})$, which proves the theorem.

A.10 Proof of Theorem 14

Define

$$\bar{\theta}_{t,i} = \frac{\sum_{s=1\vee(t-w)}^{t-1} \theta_s(i) \cdot \mathbf{1}[X_s(i)=1]}{\max\{N_{t-1}(i),1\}}.$$

First, we claim that, with probability at least $1-\delta$, for all $i \in [d], t \in T$ it holds that

$$\left|\bar{\theta}_{t,i} - \hat{\theta}_{t,i}\right| \le 2R\sqrt{\frac{\log(2dT/\delta)}{\max\{N_{t-1}(i),1\}}} \le 4R\sqrt{\frac{\log(2dT/\delta)}{N_{t-1}(i)+1}}. \tag{A.58}$$

The Claim is proved by applying the following inequality for each item $i \in [d]$. Let $\Upsilon_1, \ldots, \Upsilon_T$ be i.i.d R-sub-Gaussian random variables with mean zero. For any $\delta \in (0,1)$, we have

$$\Pr\left(\left|\frac{1}{t-q+1}\sum_{s=q}^{t}\Upsilon_s\right| \le 2R\sqrt{\frac{\log(2dT/\delta)}{t-q+1}} \text{ for all } 1 \le q \le t \le T\right) \ge 1-\frac{\delta}{d}, \tag{A.59}$$

by Corollary 1.7 of Rigollet and Hütter [154] and a union bound over all (q,t) with $1 \le q \le t \le T$ (We can alternatively use Lemma 6 in Abbasi-Yadkori et al. [3] for a slightly worse bound, but holds for more general η_t).

Next, observe that for each i,t, for certain we have

$$\begin{aligned}
\left|\bar{\theta}_{t,i} - \theta_{t,i}\right| &\le \frac{1}{\max\{N_{t-1}(i),1\}} \sum_{s=1\vee(t-w)}^{t-1} \mathbf{1}[X_s(i)=1] \cdot |\theta_s(i) - \theta_t(i)| \\
&\le \frac{1}{\max\{N_{t-1}(i),1\}} \sum_{s=1\vee(t-w)}^{t-1} \mathbf{1}[X_s(i)=1] \cdot \left(\sum_{q=s}^{t-1} \left|\theta_q(i) - \theta_{q+1}(i)\right|\right) \\
&\le \sum_{s=1\vee(t-w)}^{t-1} |\theta_s(i) - \theta_{s+1}(i)| \le \sum_{s=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_\infty .
\end{aligned} \tag{A.60}$$

A.11 Proof of Theorem 15

Recall our notations on $N_{t-1}(i)$ and $\hat{\theta}_{t,i}$ (Note that $\mathbf{1}[X_s(i)=1] = X_s(i)$):

$$N_{t-1}(i) = \sum_{s=1\vee(t-w)}^{t-1} \mathbf{1}[X_s(i)=1],$$
$$\hat{\theta}_{t,i} = \frac{\sum_{s=1\vee(t-w)}^{t-1} W_s(i)\cdot \mathbf{1}[X_s(i)=1]}{\max\{N_{t-1}(i),1\}}. \tag{A.61}$$

First, we claim that, with probability at least $1-\delta$, it holds that

$$\left|\hat{\theta}_{t,i} - \theta_{t,i}\right| \le 4R\sqrt{\frac{\log(2dT/\delta)}{N_{t-1}(i)+1}} + \sum_{s=1\vee(t-w)}^{t-1} \|\theta_s - \theta_{s+1}\|_\infty.$$

Consequently, the following UCB holds for each t with probability at least $1-\delta$:

$$\begin{aligned}
\theta_t^\top X_t &\le \max_{x\in\mathscr{E}_t}\left\{\theta_t^\top x\right\} \\
&\le \max_{x\in\mathscr{E}_t}\left\{\sum_{i\in E}\left[\hat{\theta}_{t,i} + 4R\sqrt{\frac{\log(2dT/\delta)}{N_{t-1}(i)+1}} + \sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_\infty\right]x(i)\right\} \\
&= \sum_{i\in E}\left[\hat{\theta}_{t,i} + 4R\sqrt{\frac{\log(2dT/\delta)}{N_{t-1}(i)+1}} + \sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_\infty\right]X_t(i).
\end{aligned} \tag{A.62}$$

By summing (A.62) across t, we can bound the dynamic regret with probability at least $1-\delta$ as

$$\begin{aligned}
&\mathscr{R}_T(\texttt{SW-UCB} \text{ algorithm for combinatorial semi-bandits}) \\
&\le \underbrace{\sum_{t=1}^{T}\sum_{i\in E} 4R\sqrt{\frac{\log(2dT/\delta)}{N_{t-1}(i)+1}}\cdot\mathbf{1}[X_t(i)=1]}_{(\dagger_{\mathrm{SCB}})} + \underbrace{m\sum_{t=1}^{T}\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_\infty}_{(\ddagger_{\mathrm{SCB}})}.
\end{aligned} \tag{A.63}$$

To complete the proof on the regret bound, we bound each $(\dagger_{\mathrm{SCB}}, \ddagger_{\mathrm{SCB}})$ from above.

Analysing $(\dagger_{\mathrm{SCB}})$. Let's first define the notation $\bar{N}_{i,t} = \sum_{s=1+\lfloor t/w\rfloor\cdot w}^{t-1}\mathbf{1}[X_s(i)=1]$. We can understand $\bar{N}_{i,t}$ as follows, similarly to the derivation in the proof of Lemma 37. On

one hand, the parameter $N_{i,t}$ counts the occurrences of $X_s(i) = 1$ in the w previous rounds (or $t-1$ previous rounds if $t \le w$). On the other hand, for the parameter $\bar{N}_{i,t}$, we first divide the horizon into consecutive blocks of w rounds (with the last block having $T - \lfloor T/w \rfloor \cdot w$ rounds). Then, for a round t, we look at the block that t belongs to, and the parameter $\bar{N}_{i,t}$ counts the occurrences of $X_s(i) = 1$ for $s < t$ in that block. Certainly, we have $\bar{N}_{i,t} \le N_{i,t}$.

We next use $\bar{N}_{i,t}$ to proceed with the bound:

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{i \in E} \sqrt{\frac{\mathbf{1}[X_t(i)=1]}{N_{i,t}+1}} &\le \sum_{t=1}^{T} \sum_{i \in E} \sqrt{\frac{\mathbf{1}[X_t(i)=1]}{\bar{N}_{i,t}+1}} \\
&= \sum_{j=1}^{\lceil T/w \rceil} \sum_{i \in E} \sum_{t=(j-1)w+1}^{j \cdot w \wedge T} \sqrt{\frac{\mathbf{1}[X_t(i)=1]}{\bar{N}_{i,t}+1}} \\
&\le \sum_{j=1}^{\lceil T/w \rceil} \sum_{i \in E} \sum_{t=(j-1)w+1}^{j \cdot w \wedge T} \sqrt{\frac{\mathbf{1}[X_t(i)=1]}{\max\{\bar{N}_{i,t},1\}}} \\
&\le \sum_{j=1}^{\lceil T/w \rceil} \sum_{i \in E} \left\{ 1 + 2\sqrt{\bar{N}_{i,j \cdot w \wedge T}} \right\} \qquad \text{(A.64)} \\
&\le \sum_{j=1}^{\lceil T/w \rceil} \left\{ d + 2\sqrt{dmw} \right\} \qquad \text{(A.65)} \\
&\le \sum_{j=1}^{\lceil T/w \rceil} 3\sqrt{dmw} \le \frac{6\sqrt{dm}T}{\sqrt{w}}. \qquad \text{(A.66)}
\end{aligned}$$

Step (A.64) is by the observation that, when we enumerate the non-zero summands $\sqrt{\frac{\mathbf{1}[X_t(i)=1]}{\max\{\bar{N}_{i,t},1\}}}$ from $t = (i-1)w+1$ to $t = i \cdot w \wedge T$, the enumerated terms are $1/\sqrt{1}, 1/\sqrt{1}, 1/\sqrt{2}, 1/\sqrt{3}, \ldots,$ $1/\sqrt{\max\{\bar{N}_{i,j \cdot w \wedge T}, 1\}}$. The sum of these terms is upper bounded as $1 + 2\sqrt{\bar{N}_{i,j \cdot w \wedge T}}$. Step (A.65) is by the following calculation:

$$\sum_{i \in E} \sqrt{\bar{N}_{i,j \cdot w \wedge T}} \le \sqrt{d \cdot \sum_{i \in E} \bar{N}_{i,j \cdot w \wedge T}} = \sqrt{d \cdot \sum_{i \in E} \sum_{t=(j-1)w+1}^{j \cdot w \wedge T} \mathbf{1}[X_t(i)=1]} \le \sqrt{dmw}.$$

Finally, step (A.66) is by the Theorem's assumption that $(d/m) \le w \le T$.

Analysing $(\ddagger_{\mathbf{SCB}})$**.** We note that

$$m\sum_{t=1}^{T}\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s-\theta_{s+1}\|_\infty = m\sum_{s=1}^{T-1}\sum_{t=s+1}^{T\wedge(s+w)}\|\theta_s-\theta_{s+1}\|_\infty \leq mwB_T. \tag{A.67}$$

A.12 Proof of Theorem 16

Similar to the proof of Proposition 5, the dynamic regret of the BOB algorithm can be decomposed as the regret of the SW-UCB algorithm with the optimally tuned window size $w_i = w^\dagger$ $(\geq d/m)$ for each block i plus the loss due to learning the value $w^\dagger$ with the EXP3 algorithm, *i.e.*,

$$\begin{aligned}\mathbf{E}\left[\mathrm{Regret}_T(\texttt{BOB algorithm})\right] =& \mathbf{E}\left[\sum_{t=1}^{T}\langle x_t^*, \theta_t\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger}, \theta_t\right\rangle\right] \\ &+ \mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger}, \theta_t\right\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i}, \theta_t\rangle\right].\end{aligned} \tag{A.68}$$

Here, eq. (A.68) holds as the BOB algorithm restarts the SW-UCB algorithm in each block, and for a round t in block i, X_t^w refers to the action selected in round t by the SW-UCB algorithm with window size $w \wedge (t-(i-1)H-1)$ initiated at the beginning of block i.

By Theorem 15, the first expectation in eq. (A.68) can be upper bounded as

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T}\langle x_t^*, \theta_t\rangle - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger}, \theta_t\right\rangle\right] =& \mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle x_t^* - X_t^{w^\dagger}, \theta_t\right\rangle\right] \\ =& \sum_{i=1}^{\lceil T/H\rceil}\widetilde{O}\left(w^\dagger m B_T(i) + \frac{\sqrt{dm}H}{\sqrt{w^\dagger}}\right) \\ =& \widetilde{O}\left(w^\dagger B_T + \frac{\sqrt{dm}T}{\sqrt{w^\dagger}}\right),\end{aligned} \tag{A.69}$$

where

$$B_T(i) = \sum_{t=(i-1)H+1}^{(i\cdot H\wedge t)-1}\|\theta_t - \theta_{t+1}\|_\infty$$

is the total variation in block i.

We then turn to the second expectation in eq. (A.68). We can easily see that the number of rounds for the EXP3 algorithm is $\lceil T/H\rceil$ and the number of possible values of w_i's is $|J|$. If the maximum absolute sum of reward of any block does not exceed Q, the authors of

[18] gives the following regret bound.

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger},\theta_t\right\rangle-\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i},\theta_t\rangle\middle|\forall i\in[\lceil T/H\rceil]\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}Y_t\le Q/2\right]\\
=&\widetilde{O}\left(Q\sqrt{\frac{|J|T}{H}}\right). \qquad \text{(A.70)}
\end{aligned}$$

Note that the regret of our problem is at most T, eq. (A.70) can be further upper bounded as

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger},\theta_t\right\rangle-\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t^{w_i},\theta_t\rangle\right]\\
\le&\widetilde{O}\left(Q\sqrt{\frac{|J|T}{H}}\right)\times\Pr\left(\forall i\in[\lceil T/H\rceil]\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}Y_t\le Q/2\right)\\
&+\mathbf{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left\langle X_t^{w^\dagger},\theta_t\right\rangle-\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t(w_i),\theta_t\rangle\middle|\exists i\in[\lceil T/H\rceil]\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}Y_t\ge Q/2\right]\\
&\quad\times\Pr\left(\exists i\in[\lceil T/H\rceil]\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}Y_t\ge Q/2\right)\\
\le&\widetilde{O}\left(m\sqrt{H|J|T}\right)+T\cdot\frac{2}{T}\\
=&\widetilde{O}\left(m\sqrt{H|J|T}\right). \qquad \text{(A.71)}
\end{aligned}$$

Combining eq. (A.68), (A.69), and (A.71), we have for any $w^\dagger\in J$ and $w^\dagger\ge d/m$,

$$\begin{aligned}
\mathbf{E}\left[\text{Regret}_T(\texttt{BOB algorithm})\right]=&\widetilde{O}\left(w^\dagger mB_T(i)+\frac{\sqrt{dm}H}{\sqrt{w^\dagger}}+m\sqrt{H|J|T}\right)\\
=&\widetilde{O}\left(w^\dagger mB_T+\frac{\sqrt{dm}T}{\sqrt{w^\dagger}}+d^{\frac{1}{4}}m^{\frac{3}{4}}T^{\frac{3}{4}}\right).
\end{aligned}$$

where we have plugged in the choices of H and J in eq. (2.28). Therefore, we have that when $B_T\ge d^{-1/4}m^{1/4}T^{1/4}$, the BOB algorithm is able to converge to the optimal window size *i.e.*, $w^\dagger=w^*$ ($\le H$), and the dynamic regret of the BOB algorithm is upper bounded as

$$\mathscr{R}_T(\texttt{BOB algorithm})=\widetilde{O}\left(d^{\frac{1}{3}}m^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}}+d^{\frac{1}{4}}m^{\frac{3}{4}}T^{\frac{3}{4}}\right)=\widetilde{O}\left(d^{\frac{1}{3}}m^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}}\right); \qquad \text{(A.72)}$$

while if $B_T < d^{-1/4}m^{1/4}T^{1/4}$, the BOB algorithm converges to the window size $w^\dagger = H$, and the dynamic regret is

$$\mathscr{R}_T(\texttt{BOB algorithm}) = \widetilde{O}\left(d^{\frac{1}{2}}m^{\frac{1}{2}}B_T T^{\frac{1}{2}} + d^{\frac{1}{2}}T^{\frac{3}{4}}\right) = \widetilde{O}\left(d^{\frac{1}{4}}m^{\frac{3}{4}}T^{\frac{3}{4}}\right). \tag{A.73}$$

Combining the above two cases, we conclude the desired dynamic regret bound.

A.13 Supplementary Details for Section 3.6

When B_T is known , we select w^{opt} that minimizes the explicit regret bound in (A.29), resulting in

$$w^{\text{opt}} = \left\lceil \frac{\bar{w}}{B_T^{2/3}} \right\rceil, \text{ where } \bar{w} = \frac{d^{1/3}T^{2/3}}{2^{1/3}L^{2/3}}\left(R\sqrt{d\ln(T+T^2L^2/\lambda)} + \sqrt{\lambda}S\right)^{2/3}\log^{1/3}\left(1+\frac{TL^2}{d\lambda^2}\right). \tag{A.74}$$

When B_T is not known, we select $w^{\text{obl}} = \lceil \bar{w} \rceil$, which is independent of B_T.

Appendix B

Proofs for Chapter 3

B.1 Supplementary Details about MDPs

B.1.1 Linear Program Formulations

The optimal long term reward ρ_t^* is equal to the optimal value of the linear program $\mathsf{P}(r_t, p_t)$ [147]. For a reward vector r and a transition distribution p, we define

$$\begin{aligned} \mathsf{P}(r,p): \max \quad & \sum_{s\in\mathscr{S}, a\in\mathscr{A}_s} r(s,a)x(s,a) && \text{(B.1)} \\ \text{s.t.} \quad & \sum_{a\in\mathscr{A}_s} x(s,a) = \sum_{s'\in\mathscr{S}, a'\in\mathscr{A}_{s'}} p(s|s',a')x(s',a') && \forall s\in\mathscr{S} \\ & \sum_{s\in\mathscr{S}, a\in\mathscr{A}_s} x(s,a) = 1 \\ & x(s,a) \geq 0 && \forall s\in\mathscr{S}, a\in\mathscr{A}_s \end{aligned}$$

Throughout our analysis, it is useful to consider the following dual formulation $\mathsf{D}(r,p)$ of the optimization problem $\mathsf{P}(r,p)$:

$$\begin{aligned} \mathsf{D}(r,p): \min \quad & \rho && \text{(B.2)} \\ \text{s.t.} \quad & \rho + \gamma(s) \geq r(s,a) + \sum_{s'\in\mathscr{S}} p(s'|s,a)\gamma(s') && \forall s\in\mathscr{S}, a\in\mathscr{A}_s \end{aligned}$$

$$\phi, \gamma(s) \text{ free} \qquad\qquad \forall s \in \mathscr{S}.$$

The following Lemma shows that any feasible solution to $\mathsf{D}(r,p)$ is essentially bounded if the underlying MDP is communicating, which will be crucial in the subsequent analysis.

Lemma 40. *Let (ρ,γ) be a feasible solution to the dual problem $\mathsf{D}(r,p)$, where $(\mathscr{S},\mathscr{A},p)$ consititute a communicating MDP with diameter D. We have*

$$\max_{s,s'\in\mathscr{S}} \left\{\gamma(s)-\gamma(s')\right\} \le 2D.$$

The Lemma is extracted from Section 4.3.1 of [108], and it is more general than [125], which requires (ρ,γ) to be optimal instead of just feasible.

B.1.2 Proof of Proposition 18

We begin with invoking Lemma 40, which guarantees that for each t there is an optimal solution (ρ_t^*,γ_t^*) of $\mathsf{D}(r_t,p_t)$ that satisfies $0 \le \gamma_t^*(s) \le 2D_{\max}$ for all $s \in \mathscr{S}$. Recall for each t:

$$B_{r,t} = \max_{s\in\mathscr{S},a\in\mathscr{A}_s} |r_{t+1}(s,a)-r_t(s,a)|, \quad B_{p,t} = \max_{s\in\mathscr{S},a\in\mathscr{A}_s} \|p_{t+1}(\cdot|s,a)-p_t(\cdot|s,a)\|_1. \tag{B.3}$$

Consider two time indexes $t \le \tau$. We first claim the following two inequalities:

$$\rho_\tau^* \ge \rho_t^* - \sum_{q=t}^{\tau-1} \left(B_{r,q}+2D_{\max}B_{p,q}\right) \tag{B.4}$$

$$\rho_t^* \ge r_\tau(s_\tau,a_\tau) + \left[\sum_{s'\in\mathscr{S}} p_\tau(s'|s_\tau,a_\tau)\gamma_t^*(s') - \gamma_t^*(s_\tau)\right] - \sum_{q=t}^{\tau-1} \left(B_{r,q}+2D_{\max}B_{p,q}\right). \tag{B.5}$$

The proofs of inequalities (B.4, B.5) are deferred to the end. Now, combining (B.4, B.5) gives

$$\rho_\tau^* \ge r_\tau(s_\tau,a_\tau) + \left[\sum_{s'\in\mathscr{S}} p_\tau(s'|s_\tau,a_\tau)\gamma_t^*(s') - \gamma_t^*(s_\tau)\right] - 2\sum_{q=t}^{\tau-1} \left(B_{r,q}+2D_{\max}B_{p,q}\right). \tag{B.6}$$

Let positive integer $W \leq T$ be a window size, which is specified later. Summing (B.6) over $\tau = t, \ldots, t+W-1$ and taking expectation over $\{(s_\tau, a_\tau)\}_{\tau=t}^{t+W-1}$ yield

$$\sum_{\tau=t}^{t-W+1} \rho_\tau^* \geq \mathbb{E}\left[\sum_{\tau=t}^{t-W+1} r_\tau(s_\tau, a_\tau)\right] + \mathbb{E}\left[\sum_{\tau=t}^{t-W} p_\tau(s'|s_\tau, a_\tau)\gamma_t^*(s') - \gamma_t^*(s_{\tau+1})\right] \tag{B.7}$$

$$+\mathbb{E}\left[\sum_{s'\in\mathscr{S}} p_{t-W+1}(s'|s_{t-W+1}, a_{t-W+1})\gamma_t^*(s') - \gamma_t^*(s_t)\right] - 2\sum_{\tau=t}^{t-W+1}\sum_{q=t}^{\tau-1}\left(B_{r,q} + 2D_{\max}B_{p,q}\right) \tag{B.8}$$

$$\geq \mathbb{E}\left[\sum_{\tau=t}^{t-W+1} r_\tau(s_\tau, a_\tau)\right] - 2D_{\max} - 2W\sum_{q=t}^{t+W-1}\left(B_{r,q} + 2D_{\max}B_{p,q}\right). \tag{B.9}$$

To arrive at (B.9), note that the second expectation in (B.7), which is a telescoping sum, is equal to 0, since $s_{\tau+1}$ is distributed as $p(\cdot|s_\tau, a_\tau)$. In addition, we trivially lower bound the first expectation in (B.8) by $-2D_{\max}$ by applying Lemma 40. Next, consider partitioning the horizon of T steps into intervals of W time steps, where last interval could have less than W time steps. That is, the first interval is $\{1, \ldots, W\}$, the second is $\{W+1, \ldots, 2W\}$, and so on. Applying the bound (B.9) on each interval and summing the resulting bounds together give

$$\begin{aligned}\sum_{t=1}^{T} \rho_t^* &\geq \mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t, a_t)\right] - 2\lceil\frac{T}{W}\rceil D_{\max} - 2W\sum_{t=1}^{T}(B_{r,t} + 2D_{\max}B_{p,t}) \\ &\geq \mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t, a_t)\right] - \frac{4TD_{\max}}{W} - 2W(B_r + 2D_{\max}B_p).\end{aligned} \tag{B.10}$$

Now we distinguish two cases:

- **Case 1.** $T \geq B_r/D_{\max} + 2B_p$: In this case, we can choose W to be any integer in the interval $[\sqrt{T/(B_r + 2D_{\max}B_p)}, 2\sqrt{T/(B_r + 2D_{\max}B_p)}]$, and have

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t^{\Pi^*}, a_t^{\Pi^*})\right] - \sum_{t=1}^{T} \rho_t^* \leq 4\sqrt{D_{\max}(B_r + 2D_{\max}B_p)T};$$

- **Case 2.** $T < B_r/D_{\max} + 2B_p$: In this case, one can trivially upper bound

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t,a_t)\right] - \sum_{t=1}^{T}\rho_t^* \le \mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t,a_t)\right] \le B_r + 2B_p.$$

Combining the two cases yields the desired inequality in the Proposition.

Finally, we go back to proving inequalities (B.4,B.5). These inequalities are clearly true when $t = \tau$, so we focus on the case $t < \tau$.

Proving inequality (B.4). It suffices to show that the solution $(\rho_\tau^* + \sum_{q=t}^{\tau-1}(B_{r,q} + 2D_{\max}B_{p,q}), \gamma_\tau^*)$ is feasible to the linear program $\mathsf{D}(r_t, p_t)$. To see the feasibility, it suffices to check the constraint of $\mathsf{D}(r_t, p_t)$ for each state-action pair s, a:

$$\begin{aligned}
&\rho_\tau^* + \sum_{q=t}^{\tau-1}(B_{r,q} + 2D_{\max}B_{p,q}) \\
&\ge \left[r_\tau(s,a) + \sum_{q=t}^{\tau-1} B_{r,q}\right] + \left[-\gamma_\tau^*(s) + \sum_{s'\in\mathscr{S}} p_\tau(s'|s,a)\gamma_\tau^*(s') + \sum_{q=t}^{\tau-1} 2D_{\max}B_{p,q}\right].
\end{aligned}$$

The feasibility is proved by noting that

$$|r_\tau(s,a) - r_t(s,a)| \le \sum_{q=t}^{\tau-1} B_{r,q}, \tag{B.11}$$

$$\begin{aligned}
\left|\sum_{s'\in\mathscr{S}} p_\tau(s'|s,a)\gamma_\tau^*(s') - \sum_{s'\in\mathscr{S}} p_t(s'|s,a)\gamma_\tau^*(s')\right| &\le \|p_\tau(\cdot|s,a) - p_t(\cdot|s,a)\|_1 \, \|\gamma_\tau^*\|_\infty \\
&\le \sum_{q=t}^{\tau-1} B_{p,q}(2D_{\max}).
\end{aligned} \tag{B.12}$$

Proving inequality (B.5). We have

$$\begin{aligned}
\rho_t^* &\ge r_t(s_\tau,a_\tau) + \sum_{s'\in\mathscr{S}} p_t(s'|s_\tau,a_\tau)\gamma_t^*(s') - \gamma_t^*(s_\tau) \\
&\ge r_\tau(s_\tau,a_\tau) + \sum_{s'\in\mathscr{S}} p_t(s'|s_\tau,a_\tau)\gamma_t^*(s') - \gamma_t^*(s_\tau) - \sum_{s=t}^{\tau-1} B_{r,s} \qquad \text{(B.13)} \\
&\ge r_\tau(s_\tau,a_\tau) + \sum_{s'\in\mathscr{S}} p_\tau(s'|s_\tau,a_\tau)\gamma_t^*(s') - \gamma_t^*(s_\tau) - \sum_{s=t}^{\tau-1} B_{r,s} - 2D_{\max}\sum_{s=t}^{\tau-1} B_{p,s}, \qquad \text{(B.14)}
\end{aligned}$$

where steps (B.13, B.14) are by inequalities (B.11, B.12). Altogether, the Proposition is proved. ■

B.1.3 Extended Value Iteration (EVI) by [108]

Algorithm 9 EVI$(H^r, H^p; \varepsilon)$, mostly extracted from [108]

Initialize VI record $u_0 \in \mathbb{R}^{\mathscr{S}}$ as $u_0(s) = 0$ for all $s \in \mathscr{S}$.
for $i = 0, 1, \ldots$ **do**
  For each $s \in \mathscr{S}$, compute VI record $u_{i+1}(s) = \max_{a \in \mathscr{A}_s} \tilde{\Upsilon}_i(s,a)$, where

$$\tilde{\Upsilon}_i(s,a) = \max_{\dot{r}(s,a) \in H^r(s,a)} \{\dot{r}(s,a)\} + \max_{\dot{p} \in H^p(s,a)} \left\{ \sum_{s' \in \mathscr{S}} u_i(s') \dot{p}(s') \right\}.$$

  Define stationary policy $\tilde{\pi} : \mathscr{S} \to \mathscr{A}_s$ as $\tilde{\pi}(s) = \operatorname{argmax}_{a \in \mathscr{A}_s} \tilde{\Upsilon}_i(s,a)$.
  Define optimistic reward $\tilde{r} = \{\tilde{r}(s,a)\}_{s,a}$ with $\tilde{r}(s,a) \in \operatorname{argmax}_{\dot{r}(s,a) \in H^r(s,a)} \{\dot{r}(s,a)\}$.
  Define optimistic distribution $\tilde{p} = \{\tilde{p}(\cdot|s,a)\}_{s,a}$ with $\tilde{p}(\cdot|s,a) \in \operatorname{argmax}_{\dot{p} \in H^p(s,a)} \{\sum_{s' \in \mathscr{S}} u_i(s') \dot{p}(s')\}$.
  Define optimistic dual variables $\tilde{\rho} = \max_{s \in \mathscr{S}} \{u_{i+1}(s) - u_i(s)\}$, $\tilde{\gamma}(s) = u_i(s) - \min_{s \in \mathscr{S}} u_i(s)$.
  if $\max_{s \in \mathscr{S}} \{u_{i+1}(s) - u_i(s)\} - \min_{s \in \mathscr{S}} \{u_{i+1}(s) - u_i(s)\} \leq \varepsilon$ **then**
    Break the **for** loop.
  end if
end for
Return policy $\tilde{\pi}$.
Auxiliary output: optimistic reward and state transition distributions $(\tilde{r}, \tilde{p})$, optimistic dual variables $(\tilde{\rho}, \tilde{\gamma})$.

We provide the pseudo-codes of EVI$(H_r, H_p; \varepsilon)$ proposed by [108] in Algorithm 9. By [108], the algorithm converges in finite time when the confidence region H_p contains a transition distribution p such that $(\mathscr{S}, \mathscr{A}, p)$ constitutes a communicating MDP. The output $(\tilde{\pi}, \tilde{r}, \tilde{p}, \tilde{\rho}, \tilde{\gamma})$ of the EVI$(H_r, H_p; \varepsilon)$ satisfies the following two properties [108].

Property 1. *The dual variables $(\tilde{\rho}, \tilde{\gamma})$ are optimistic, i.e.,*

$$\tilde{\rho} + \tilde{\gamma}(s) \geq \max_{\dot{r}(s,a) \in H_r(s,a)} \{\dot{r}(s,a)\} + \sum_{s' \in \mathscr{S}} \tilde{\gamma}(s') \max_{\dot{p} \in H_p(s,a)} \{\dot{p}(s'|s,a)\}.$$

Property 2. *For each state $s \in \mathscr{S}$, we have*

$$\tilde{r}(s,\tilde{\pi}(s)) \geq \tilde{\rho} + \tilde{\gamma}(s) - \sum_{s' \in \mathscr{S}} \tilde{p}(s'|s,\tilde{\pi}(s))\tilde{\gamma}(s') - \varepsilon.$$

Property 1 ensures the feasibility of the output dual variables $(\tilde{\rho},\tilde{\gamma})$, with respect to the dual program $\mathsf{D}(\dot{r},\dot{p})$ for any $\dot{r},\dot{p}$ in the confidence regions H_r, H_p. The feasibility facilitates the bounding of $\max_{s\in\mathscr{S}} \tilde{\gamma}(s)$, which turns out to be useful for bounding the regret arise from switching among different stationary policies. To illustrate, suppose that H_p is so large that it contains a transition distribution $\dot{p}$ under which $(\mathscr{S},\mathscr{A},\dot{p})$ has diameter D. By Lemma 40 in Section B.1.1, we have $0 \leq \max_{s\in\mathscr{S}} \tilde{\gamma}(s) \leq 2D$.

Property 2 ensures the near-optimality of the dual variables $(\tilde{\rho},\tilde{\gamma})$ to the $(\tilde{r},\tilde{p})$ optimistically chosen from H_r, H_p. More precisely, the deterministic policy $\tilde{\pi}$ near-optimal for the MDP with time homogeneous reward function $\tilde{r}$ and time homogeneous transition distribution $\tilde{p}$, under which the policy $\tilde{\pi}$ achieves a long term average reward is at least $\tilde{\rho}^* - \varepsilon$.

B.2 Proof of Proposition 19

We distinguish two cases:

Case 1. $D_{\max}^2 B_p \geq B_r$: Following the piecewise stationary lower bound construction for the non-stationary bandit setting [36] and the non-stationary RL setting [137], we consider the following stationary MDP $\mathscr{M}$ as specified in the proof of Theorem 5 of [108] for a total of T' time periods, where there are a total of $S/2+1$ states $\{s_0, s_{1,1}, \ldots, s_{\lfloor S/2\rfloor,1}\}$, $\lfloor S/2\rfloor \times \lfloor (A-1)/2\rfloor$ actions for s_0, and $\lfloor (A-1)/2\rfloor$ actions for all $s_{q,1}$. We denote $\mathscr{A}_{s_0} = \mathscr{A}_{s_0}(1) \cup \ldots \cup \mathscr{A}_{s_0}(\lfloor S/2\rfloor)$, where $\mathscr{A}_{s_0}(q)$ with $|\mathscr{A}_{s_0}(q)| = \lfloor (A-1)/2\rfloor$ is the collection of actions that transition from s_0 to $s_{q,1}$. For any actions $a \in \mathscr{A}_{s_0}(q)$ and $a' \in \mathscr{A}_{s_{q,1}}$, we set the rewards to $r(s_0,a) = 0$ and $r(s_{q,1},a') = 1$ $\forall q \in [\lfloor S/2\rfloor]$ deterministically. We also let $p(s_{q,1}|s_0,a) = p(s_0|s_{q,1},a') = 4/D_{\max}, p(s_0|s_0,a) = p(s_{q,1}|s_{q,1},a') = 1-4/D_{\max}$ for all a and all q except for one q^* and $a^* \in \mathscr{A}_{s_0}(q^*)$ such that $p(s_{q,1}|s_0,a) = 4/D_{\max} + \sqrt{S(A-1)/(25T'D_{\max})}$. Here, q^* is first chosen uniformly random among all $q \in [\lfloor S/2\rfloor]$ and then a^* is then chosen

uniformly from $\mathscr{A}_{s_0}(q^*)$.

Lemma 41. *Denoting* $\nu_1 = 4/D_{\max}$ *and* $\nu_2 = \sqrt{S(A-1)/(25T'D_{\max})}$, *the optimal reward of* $\mathscr{M}$ *over a total of* T' *time period is at least* $\frac{\nu_1+\nu_2}{2\nu_1+\nu_2}(T'-1)$ *for any starting state.*

The proof of this lemma is provided in Section B.2.1.

Lemma 42. *For any algorithm, any* $S, A \geq 10, D_{\max} \geq 20\log_A S$, *and* $T' \geq D_{\max}AS$, *the regret of this algorithm, which aims at collecting reward by interacting with* $\mathscr{M}$, *is at least* $\Omega(\sqrt{D_{\max}SAT'})$.

This lemma can be easily shown by combining Lemma 41 and Theorem 5 of [108]. Now we consider partitioning T, the entire time horizon, into epochs of length T'. This results in a total of $\lceil T/T' \rceil$ epochs, each with T' steps (except possibly for the last one). For each epoch, a new pair of (q^*, a^*) is sampled uniformly random. Then, even if the DM knows this additional piece of information, she still has to suffer $\Omega(\sqrt{D_{\max}SAT'})$ (dynamic) regret per epoch according to Lemma 42. This is because the epochs are completely independent. Therefore, the total dynamic regret is of order at least

$$\Omega(T\sqrt{D_{\max}SA/T'}). \tag{B.15}$$

Now, each change of epoch would incur a $2\sqrt{S(A-1)/(25T'D_{\max})}$ consumption of the variation budget B_p, which implies $2(\lceil T/T' \rceil - 1)\sqrt{S(A-1)/(25T'D_{\max})} \leq B_p$. Consequently, $T' \geq 5^{-2/3}D_{\max}^{-1/3}S^{1/3}(A-1)^{1/3}B_p^{-2/3}T^{2/3}$. Taking the least possible value of T', (B.15) becomes $\Omega(D_{\max}^{2/3}B_p^{1/3}S^{1/3}A^{1/3}T^{2/3})$.

Case 2. $B_r \geq D_{\max}^2 B_p$: For this case, we try to map the problem to a multi-armed bandits problem with SA actions. We consider exactly the same state and action space as the previous case. The state transition probability between s_0 and all $s_{q,1}$'s are set deterministically to 1. The reward of any action related to s_0 is 0. Now, we can follow the same argument as [36] to set the reward of the actions related to $s_{q,1}$'s. This gives us a lower bound of $\Omega(B_r^{1/3}S^{1/3}A^{1/3}T^{2/3})$ for any $B_r \in [S^{-1}A^{-1}, S^{-1}A^{-1}T]$.

B.2.1 Proof of Lemma 41

In computing the optimal reward, we only need to consider s_0 and s_1 $(= s_{q^*,1})$. Let $V_t(s_0)$ and $V_t(s_1)$ be the optimal reward of going from time step t to T'. One can easily formulate the following recursive equations

$$V_{T'+1}(s_0) = V_{T'+1}(s_1) = 0, \tag{B.16}$$

$$V_t(s_0) = (1-\nu_1-\nu_2)V_{t+1}(s_0) + (\nu_1+\nu_2)V_{t+1}(s_1), \tag{B.17}$$

$$V_t(s_1) = 1 + \nu_1 V_{t+1}(s_0) + (1-\nu_1)V_{t+1}(s_1). \tag{B.18}$$

From (B.17) and (B.18), one can easily derive that

$$\begin{aligned} V_{t-2}(s_0) &= (2-2\nu_1-\nu_2)V_{t-1}(s_0) + (2\nu_1+\nu_2-1)V_t(s_0) + \nu_1+\nu_2, \\ V_{t-2}(s_1) &= (2-2\nu_1-\nu_2)V_{t-1}(s_1) + (2\nu_1+\nu_2-1)V_t(s_1) + \nu_1+\nu_2. \end{aligned}$$

Re-arranging the terms

$$\begin{aligned} V_{t-2}(s_0) + (2\nu_1+\nu_2-1)V_{t-1}(s_0) &= V_{t-1}(s_0) + (2\nu_1+\nu_2-1)V_t(s_0) + \nu_1+\nu_2, \\ V_{t-2}(s_1) + (2\nu_1+\nu_2-1)V_{t-1}(s_1) &= V_{t-1}(s_1) + (2\nu_1+\nu_2-1)V_t(s_1) + \nu_1+\nu_2. \end{aligned}$$

Taking the telescoping sum from $t=3$ to $T'+1$, we have

$$\begin{aligned} V_1(s_0) + (2\nu_1+\nu_2-1)V_2(s_0) &= V_{T'}(s_0) + (2\nu_1+\nu_2-1)V_{T'+1}(s_0) + (T'-1)(\nu_1+\nu_2), \\ V_1(s_1) + (2\nu_1+\nu_2-1)V_2(s_1) &= V_{T'}(s_1) + (2\nu_1+\nu_2-1)V_{T'+1}(s_1) + (T'-1)(\nu_1+\nu_2). \end{aligned}$$

Through direction computation, one could easily verify that $V_{T'}(s_0) = V_{T'+1}(s_0) = V_{T'+1}(s_1) = 0$ and $V_{T'}(s_1) = 1$, which gives us

$$\begin{aligned} V_1(s_0) + (2\nu_1+\nu_2-1)V_2(s_0) &= (T'-1)(\nu_1+\nu_2), \\ V_1(s_1) + (2\nu_1+\nu_2-1)V_2(s_1) &= 1 + (T'-1)(\nu_1+\nu_2). \end{aligned}$$

Note that the reward of $\mathscr{M}$ is non-negative and thus $V_1(s_0) \geq V_2(s_0),\ V_1(s_1) \geq V_2(s_1)$. Therefore,

$$V_1(s_0) \geq \frac{\nu_1+\nu_2}{2\nu_1+\nu_2}(T'-1), \quad V_1(s_1) \geq \frac{\nu_1+\nu_2}{2\nu_1+\nu_2}(T'-1)+\frac{1}{2\nu_1+\nu_2},$$

which concludes the proof.

B.3 Proof of Proposition 20

The sequence $p_1,\ldots,p_W$ alternates between the following 2 instances p^1,p^2. Now, define the common state space $\mathscr{S}=\{1,2\}$ and action collection $\mathscr{A}=\{\mathscr{A}_1,\mathscr{A}_2\}$, where $\mathscr{A}_1=\{a_1,a_2\}$, $\{\mathscr{A}_2\}=\{b_1,b_2\}$. We assume all the state transitions are deterministic, and a graphical illustration is presented in Fig. B-1. Clearly, we see that both instances have diameter 1.

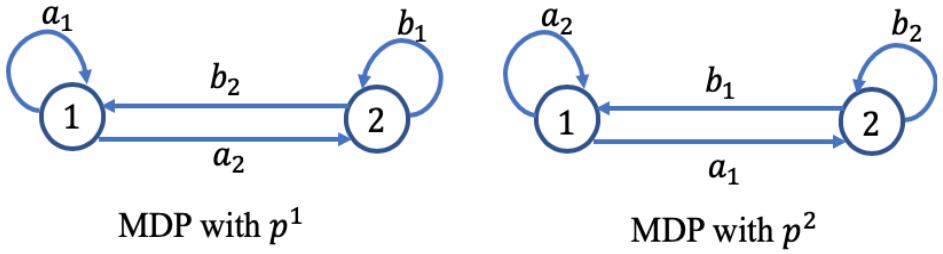


Figure B-1: Example MDPs. Since the transitions are deterministic, the probabilities are omitted.

Now, consider the following two deterministic and stationary policies $\pi^1:\pi^1(1)=a_1,\pi^1(2)=b_2$, and $\pi^2:\pi^2(1)=a_2,\pi^2(2)=b_1$. Since the MDP is deterministic, we have $\hat{p}_{W+1}=\bar{p}_{W+1}$.

In the following, we construct a trajectory where the DM alternates between policies π^1,π^2 during time $\{1,\ldots,W\}$ while the underlying transition distribution alternates between p^1,p^2. In the construction, the DM is almost always at the self-loop at state 1 (or 2) throughout the horizon, no matter what action a_1,a_2 (or b_1,b_2) she takes. Consequently, it will trick the DM into thinking that $\hat{p}_{W+1}(1|1,a_i)\approx 1$ for each $i\in\{1,2\}$, and likewise $\hat{p}_{W+1}(2|2,b_i)\approx 1$ for each $i\in\{1,2\}$. Altogether, this will lead the DM to conclude that

$(\mathscr{S}, \mathscr{A}, \hat{p}_{W+1})$ constitute a high diameter MDP, since the probability of transiting from state 1 to 2 (and 2 to 1) are close to 0.

The construction is detailed as follows. Let $W = 4\tau$. In addition, let the state transition distributions be

$$p_1 = \ldots = p_\tau = p^1, \quad p_{\tau+1} = \ldots = p_{2\tau} = p^2, \quad p_{2\tau+1} = \ldots = p_{3\tau} = p^1, \quad p_{3\tau+1} = \ldots = p_{4\tau} = p^2.$$

The DM starts at state 1. She follows policy π^1 from time 1 to time 2τ, and then policy π^2 from $2\tau+1$ to 4τ.

Under the specified MDP models and policies, it can be readily verified that the DM takes action a_1 from time 1 to $\tau+1$, action b_2 from time $\tau+2$ to 2τ, action b_1 from time $2\tau+1$ to $3\tau+1$, and action a_2 from time $3\tau+2$ to 4τ. As a result, the DM is at state 1 from time 1 to $\tau+1$, state 2 from time $\tau+2$ to $3\tau+1$, and eventually state 1 from time $3\tau+2$ to 4τ as depicted in Fig. B-2. We thus have:

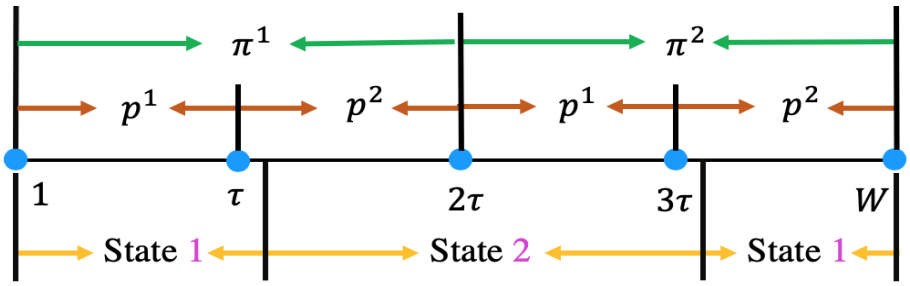


Figure B-2: Illustration of the latent MDPs, policies, and state visits.

$$\hat{p}_{W+1}(1|1,a_1) = \frac{\tau}{\tau+1}, \quad \hat{p}_{W+1}(2|1,a_1) = \frac{1}{\tau+1}, \quad \hat{p}_{W+1}(1|1,a_2) = 1, \quad \hat{p}_{W+1}(2|1,a_2) = 0$$
$$\hat{p}_{W+1}(2|2,b_1) = \frac{\tau}{\tau+1}, \quad \hat{p}_{W+1}(1|2,b_1) = \frac{1}{\tau+1}, \quad \hat{p}_{W+1}(2|2,b_2) = 1, \quad \hat{p}_{W+1}(1|2,b_2) = 0,$$

and It can be readily verified that the diameter of $(\mathscr{S}, \mathscr{A}, \hat{p}_{W+1})$ is $\tau+1 = \Theta(W)$. Finally, for the confidence region $H_{p,W+1}(0) = \{H_{p,W+1}(s,a;0)\}_{s,a}$ constructed without confidence widening, for any $\tilde{p} \in H_{p,W+1}(0)$ we have

$$\tilde{p}(2|1,a_1) = \tilde{p}(1|2,b_1) = O\left(\sqrt{\frac{\log W}{\tau+1}}\right), \quad \tilde{p}(2|1,a_2) = \tilde{p}(1|2,b_2) = O\left(\sqrt{\frac{\log W}{\tau-1}}\right)$$

respectively. Since the stochastic confidence radii $\Theta\left(\sqrt{\frac{\log W}{\tau+1}}\right)$ and $\Theta\left(\sqrt{\frac{\log W}{\tau-1}}\right)$ dominate the sample mean $\frac{1}{\tau+1}$ and 0. Therefore, for any $\tilde{p} \in H_{p,W+1}(0)$, the diameter of the MDP constructed by $(\mathscr{S}, \mathscr{A}, \tilde{p})$ is at least $\Omega\left(\sqrt{\frac{W}{\log W}}\right)$.

B.4 Proof of Theorem 21

Recall

$$\mathscr{E}_r = \{\bar{r}_t(s,a) \in H_{r,t}(s,a)\ \forall s,a,t\}, \quad \mathscr{E}_p = \{\bar{p}_t(\cdot|s,a) \in H_{p,t}(s,a;0)\ \forall s,a,t\},$$

we first show that the events $\mathscr{E}_r$ and $\mathscr{E}_p$ hold with probability at least $1-\delta$.

Lemma 43. *We have* $\Pr[\mathscr{E}_r] \geq 1-\delta/2$, $\Pr[\mathscr{E}_p] \geq 1-\delta/2$.

The proof of Lemma 43 is provided in Section B.5. We then define the following variation measure for each t in an episode m:

$$\mathsf{var}_{r,t} = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} B_{r,q}, \quad \mathsf{var}_{p,t} = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} B_{p,q}.$$

With these notations, we provide an upper bound on the difference between ρ_t^* and `SWUCRL2-CW` algorithm's reward at a time step t of episode m when $H_{p,\tau(m)}(\eta)$ contains a state transition distribution with small diameter.

Proposition 44. *Consider an episode m, condition on events $\mathscr{E}_r, \mathscr{E}_p$, and suppose that there exists a state transition distribution $\mathring{p} \in H_{p,\tau(m)}(\eta)$ such that the diameter of $(\mathscr{S}, \mathscr{A}, \mathring{p})$ at most D. Then, for every $t \in \{\tau(m), \ldots, \tau(m+1)-1\}$ in episode m, we have*

$$\rho_t^* - r_t(s_t,a_t) \leq \left[\sum_{s' \in \mathscr{S}} p_t(s'|s_t,a_t)\tilde{\gamma}_{\tau(m)}(s')\right] - \tilde{\gamma}_{\tau(m)}(s_t) \tag{B.19}$$

$$+\frac{1}{\sqrt{\tau(m)}} + [2\mathsf{var}_{r,t} + 4D(\mathsf{var}_{p,t} + \eta)] + \left[2\mathsf{rad}\text{-}_{r,\tau(m)}(s_t,a_t) + 4D \cdot \mathsf{rad}\text{-}_{p,\tau(m)}(s,a)\right]. \tag{B.20}$$

The proof of Proposition 44 is provided in Section B.6.

To facilitate our discussion, we denote $M(T)$ as the total number of episodes. By abusing the notation we, let $\tau(M(T)+1)-1=T$. Episode $M(T)$, containing the final round T, is interrupted and the algorithm is forced to terminate as the end of time T is reached. We can now rewrite the difference between the quantity $\sum_{t=1}^{T}\rho_t^*$ and the expected cumulative reward of the `SWUCRL2-CW` algorithm as the sum of difference from each episode:

$$\sum_{t=1}^{T}(\rho_t^*-r_t(s_t,a_t))=\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}(\rho_t^*-r_t(s_t,a_t)) \tag{B.21}$$

To proceed, we define the set

$$U=\{m\in[M(T)]:p_{\tau(m)}(\cdot|s,a)\in H_{p,\tau(m)}(s,a;\eta)\ \forall(s,a)\in\mathscr{S}\times\mathscr{A}_s\}.$$

For each episode $m\in[M(T)]$, we distinguish two cases:

- **Case 1.** $m\in U$: Under this situation, we apply Proposition 44 to bound the difference during the episode, using the fact that $p_{\tau(m)}$ satisfies the assumptions of the proposition with $D=D_{\tau(m)}\leq D_{\max}$.
- **Case 2.** $m\in[M(T)]\setminus U$: In this case, we trivially upper bound the difference of each round in episode m by 1.

For case 1, we bound the difference during episode m by summing the error terms in (B.19, B.20) across the rounds $t\in[\tau(m),\tau(m+1)-1]$ in the episode. The term (B.19) accounts for the error by switching policies. In (B.20), the terms $\mathsf{rad}\text{-}_{r,\tau(m)},\mathsf{rad}\text{-}_{p,\tau(m)}$ accounts for the estimation errors due to stochastic variations, and the term $\mathsf{var}_{r,t},\mathsf{var}_{p,t}$ accounts for the estimation error due to non-stationarity.

For case 2, we need an upper bound on $\sum_{m\in[M(T)]\setminus U}\sum_{t=\tau(m)}^{\tau(m+1)-1}1$, the total number of rounds that belong to an episode in $[M(T)]\setminus U$. The analysis is challenging, since the length of each episode may vary, and one can only guarantee that the length is $\leq W$. A first attempt could be to upper bound as $\sum_{m\in[M(T)]\setminus U}\sum_{t=\tau(m)}^{\tau(m+1)-1}1\leq W\sum_{m\in[M(T)]\setminus U}1$, but the resulting bound appears too loose to provide any meaningful regret bound. Indeed, there could be double counting, as the starting time steps for a pair of episodes in case 2 might

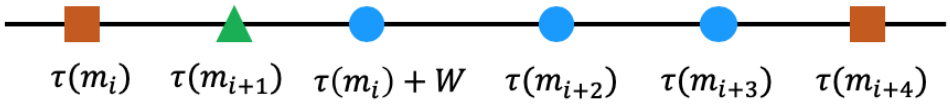


Figure B-3: Both episodes m_i and m_{i+4} belong to Q_T (and thus $\tilde{Q}_T$) because $p_{\tau(m_i)} \notin H_{p,\tau(m_i)}(\eta)$ and $p_{\tau(m_{i+4})} \notin H_{p,\tau(m_{i+4})}(\eta)$. m_{i+1} is added to $\tilde{Q}_T$ (but not Q_T) because $\tau(m_{i+1}) - \tau(m_i) \in [0,W]$. m_{i+2} and m_{i+3} belong to neither of Q_T nor $\tilde{Q}_T$ as $p_{\tau(m_{i+2})} \in H_{p,\tau(m_{i+2})}(\eta)$ and $p_{\tau(m_{i+3})} \in H_{p,\tau(m_{i+3})}(\eta)$.

not even be W rounds apart!

To avoid the trap of double counting, we consider a set $Q_T \subseteq [M(T)] \setminus U$ where the start times of the episodes are sufficiently far apart, and relate the cardinality of Q_T to $\sum_{m\in[M(T)]\setminus U} \sum_{t=\tau(m)}^{\tau(m+1)-1} 1$. The set $Q_T \subseteq [M(T)]$ is constructed sequentially, by examining all episodes $m = 1,\ldots,M(T)$ in the time order. At the start, we initialize $Q_T = \emptyset$. For each $m = 1,\ldots,M(T)$, we perform the following. If episode m satisfies both criteria:

1. There exists some $s \in \mathscr{S}$ and $a \in \mathscr{A}_s$ such that $p_{\tau(m)}(\cdot|s,a) \notin H_{p,\tau(m)}(s,a;\eta)$;
2. For every $m' \in Q_T$, $\tau(m) - \tau(m') > W$,

then we add m into Q_T. Afterwards, we move to the next episode index $m+1$. The process terminates once we arrive at episode $M(T)+1$. The construction ensures that, for each episode $m \in [M(T)]$, if $\tau(m) - \tau(m') \notin [0,W]$ for all $m' \in Q_T$, then $\forall s \in \mathscr{S}\ \forall a \in \mathscr{A}_s\ p_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a)$; otherwise, m would have been added into Q_T.

We further construct a set $\tilde{Q}_T$ to include all elements in Q_T and every episode index m such that there exists $m' \in Q_T$ with $\tau(m) - \tau(m') \in [0,W]$. By doing so, we can prove that every episode $m \in [M(T)] \setminus \tilde{Q}_T$ satisfies $p_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a)\ \forall s \in \mathscr{S}\ \forall a \in \mathscr{A}_s$. The procedures for building $\tilde{Q}_T$ (initialized to Q_T) are described as follows: for every episode index $m \in [M(T)]$, if there exists $m' \in Q_T$, such that $\tau(m) - \tau(m') \in [0,W]$, then we add m to $\tilde{Q}_T$. Formally,

$$\tilde{Q}_T = Q_T \cup \left\{m \in [M(T)] : \exists m' \in Q_T\ \tau(m) - \tau(m') \in [0,W]\right\}.$$

We can formalize the properties of Q_T and $\tilde{Q}_T$ as follows.

Lemma 45. *Conditioned on $\mathscr{E}_p$, $|Q_T| \le B_p/\eta$.*

Lemma 46. *For any episode $m \notin \tilde{Q}_T$, we have $p_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a;\eta)$ for all $s \in \mathscr{S}$ and $a \in \mathscr{A}_s$.*

The proofs of Lemmas 45 and 46 are presented in Sections B.7 and B.8, respectively.

Together with eqn. (B.21) and Proposition 18, we can further decompose the dynamic regret of the `SWUCRL2-CW` algorithm as

$$
\begin{aligned}
&\text{Dyn-Reg}_T(\texttt{SWUCRL2-CW}) \\
=&\sum_{t=1}^{T}\Big(\mathbb{E}[r_t(s_t^{\Pi^*},a_t^{\Pi^*})]-\mathbb{E}[r_t(s_t^{\Pi},a_t^{\Pi})]\Big) \\
=&\sum_{t=1}^{T}\mathbb{E}[r_t(s_t^{\Pi^*},a_t^{\Pi^*})]-\sum_{t=1}^{T}\rho_t^*+\sum_{t=1}^{T}\rho_t^*-\sum_{t=1}^{T}\mathbb{E}[r_t(s_t^{\Pi},a_t^{\Pi})] \\
=&\sum_{t=1}^{T}\mathbb{E}[r_t(s_t^{\Pi^*},a_t^{\Pi^*})]-\sum_{t=1}^{T}\rho_t^*+\sum_{m\in\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}(\rho_t^*-r_t(s_t,a_t)) \\
&+\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}(\rho_t^*-r_t(s_t,a_t)) \\
\le&4\sqrt{D_{\max}(B_r+2D_{\max}B_p)T}+(B_r+2B_p) \\
&+\sum_{m\in\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}(\rho_t^*-r_t(s_t,a_t)) && (\spadesuit) \\
&+\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left\{\left[\sum_{s'\in\mathscr{S}}p_t(s'|s_t,a_t)\tilde{\gamma}_{\tau(m)}(s')-\tilde{\gamma}_{\tau(m)}(s_t)\right]+\frac{1}{\sqrt{\tau(m)}}\right\} && (\clubsuit) \\
&+\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}(2\mathsf{var}_{r,t}+4D_{\max}\cdot\mathsf{var}_{p,t}+2D_{\max}\eta) && (\blacklozenge) \\
&+\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left[2\mathsf{rad}\text{-}_{r,\tau(m)}(s_t,a_t)+4D_{\max}\cdot\mathsf{rad}\text{-}_{p,\tau(m)}(s_t,a_t)\right], && (\blacktriangledown)
\end{aligned}
$$

where the last step makes use of Lemma 46 and Proposition 44. We accomplish the promised dynamic regret bound by the following four Lemmas that bound the dynamic regret terms (♠, ♣, ♦, ♥).

Lemma 47. *Conditioned on $\mathscr{E}_p$, we have*

$$(\spadesuit) = O\left(\frac{B_p W}{\eta}\right).$$

Lemma 48. *Conditioned on events $\mathscr{E}_r, \mathscr{E}_p$, we have with probability at least $1 - O(\delta)$ that*

$$(\clubsuit) = \tilde{O}(D_{max}[M(T) + \sqrt{T}]) = \tilde{O}\left(D_{max}\left[\frac{SAT}{W} + \sqrt{T}\right]\right).$$

Lemma 49. *With certainty,*

$$(\blacklozenge) = O\left((B_r + D_{max}B_p)W + D_{max}T\eta\right).$$

Lemma 50. *With certainty, we have*

$$(\heartsuit) = \tilde{O}\left(\frac{D_{max}S\sqrt{A}T}{\sqrt{W}}\right).$$

The proofs of Lemmas 47, 48, 49, and 50 are presented in Sections B.9, B.10, B.11, and B.12, respectively. Putting all these pieces together, we have the dynamic regret of the `SWUCRL2-CW` algorithm is upper bounded as

$$\tilde{O}\left(\frac{B_p W}{\eta} + B_r W + D_{\max}\left[B_p W + \frac{S\sqrt{A}T}{\sqrt{W}} + T\eta + \frac{SAT}{W} + \sqrt{T}\right] + \sqrt{D_{\max}(B_r + 2D_{\max}B_p)T}\right),$$

and by setting W and η accordingly, we can conclude the proof.

B.5 Proof of Lemma 43

We employ the self-normalizing concentration inequallity [3]. The following inequality is extracted from Theorem 1 in [3], restricted to the case when $d = 1$.

Proposition 51 ([3])**.** *Let $\{\mathscr{F}_q\}_{q=1}^T$ be a filtration. Let $\{\xi_q\}_{q=1}^T$ be a real-valued stochastic process, such that ξ_q is $\mathscr{F}_q$-measurable, and ξ_q is conditionally R-sub-Gaussian, i.e. for all $\lambda \geq 0$, it holds that $\mathbb{E}[\exp(\lambda\xi_q)|\mathscr{F}_{q-1}] \leq \exp(\lambda^2 R^2/2)$. Let $\{Y_q\}_{q=1}^T$ be a non-negative*

real-valued stochastic process such that Y_q is $\mathscr{F}_{q-1}$-measurable. For any $\delta' \in (0,1)$, it holds that

$$\Pr\left(\frac{\sum_{q=1}^{t}\xi_q Y_q}{\max\{1,\sum_{q=1}^{t}Y_q^2\}} \le 2R\sqrt{\frac{\log(T/\delta')}{\max\{1,\sum_{q=1}^{t}Y_q^2\}}} \text{ for all } t\in[T]\right) \ge 1-\delta'.$$

In particular, if $\{Y_q\}_{q=1}^{T}$ be a $\{0,1\}$-valued stochastic process, then for any $\delta' \in (0,1)$, it holds that

$$\Pr\left(\frac{\sum_{q=1}^{t}\xi_q Y_q}{\max\{1,\sum_{q=1}^{t}Y_q\}} \le 2R\sqrt{\frac{\log(T/\delta')}{\max\{1,\sum_{q=1}^{t}Y_q\}}} \text{ for all } t\in[T]\right) \ge 1-\delta'. \tag{B.22}$$

The Lemma is proved by applying Proposition 51 with suiatable choices of

$$\mathscr{F}_{q=1}^{T}, \{\xi_q\}_{q=1}^{T}, \{Y_q\}_{q=1}^{T}, \delta$$

. We divide the proof into two parts.

B.5.1 Proving $\Pr[\mathscr{E}_r] \ge 1-\delta/2$

It suffices to prove that, for any fixed $s\in\mathscr{S}, a\in\mathscr{A}_s, t\in[T]$, it holds that

$$\begin{aligned}
&\Pr\left(\left|\hat{r}_(s,a)-\bar{r}_t(s,a)\right| \le \mathsf{rad}\text{-}_{r,t}(s,a)\right)\\
=&\Pr\left(\left|\frac{1}{N_t^+(s,a)}\sum_{q=(\tau(m)-W)\vee 1}^{t-1}\left[R_q(s,a)-r_q(s,a)\right]\cdot\mathbf{1}(s_q=s,a_q=a)\right| \le 2\sqrt{\frac{\log(2SAT^2/\delta)}{N_t^+(s,a)}}\right)\\
\ge&1-\frac{\delta}{2SAT}.
\end{aligned} \tag{B.23}$$

since then $\Pr[\mathscr{E}_r]\ge 1-\delta/2$ follows from the union bound over all $s\in\mathscr{S}, a\in\mathscr{A}_s, t\in[T]$. Now, the trajectory of the online algorithm is expressed as $\{s_q,a_q,R_q\}_{q=1}^{T}$. Inequality (B.23) directly follows from Proposition 51, with $\{\mathscr{F}_q\}_{q=1}^{T}, \{\xi_q\}_{q=1}^{T}, \{Y_q\}_{q=1}^{T}, \delta$ defined as

$$\begin{aligned}
\mathscr{F}_q &= \{(s_\ell,a_\ell,R_\ell)\}_{\ell=1}^{q}\cup\{(s_{q+1},a_{q+1})\},\\
\xi_q &= R_q(s,a)-r_q(s,a),
\end{aligned}$$

$$Y_q = \mathbf{1}\left(s_q = s, a_q = a, ((t-W)\vee 1) \le q \le t-1\right),$$
$$\delta' = \frac{\delta}{2SAT}.$$

Each ξ_q is conditionally 2-sub-Gaussian, since $-1 \le \xi_q \le 1$ with certainty. Altogether, the required inequality is shown.

B.5.2 Proving $\Pr[\mathscr{E}_p] \ge 1-\delta/2$

We start by noting that, for two probability distributions $p, \{p(s)\}_{s\in\mathscr{S}}, p' = \{p'(s)\}_{s\in\mathscr{S}}$ on $\mathscr{S}$, it holds that

$$\left\|p - p'\right\|_1 = \max_{\theta\in\{-1,1\}^{\mathscr{S}}} \theta(s)\cdot(p(s)-p'(s)).$$

Consequently, to show $\Pr[\mathscr{E}_p] \ge 1-\delta/2$, it suffices to show that, for any fixed $s\in\mathscr{S}, a\in\mathscr{A}_s, t\in[T], \theta\in\{-1,1\}^{\mathscr{S}}$, it holds that

$$\begin{aligned}
&\Pr\left(\sum_{s'\in\mathscr{S}} \theta(s)\cdot\left(\hat{p}_t(s'|s,a) - \bar{p}_t(s'|s,a)\right) \le \mathsf{rad}\text{-}_{p,t}(s,a)\right)\\
\le&\Pr\left(\frac{1}{N_t^+(s,a)}\sum_{q=(\tau(m)-W)\vee 1}^{t-1}\left[\sum_{s'\in\mathscr{S}}\theta(s')\mathbf{1}(s_q=s,a_q=a,s_{q+1}=s')\right]\right.\\
&\qquad\left.-\left[\sum_{s'\in\mathscr{S}}\theta(s')p_q(s'|s,a)\cdot\mathbf{1}(s_q=s,a_q=a)\right] \le 2\sqrt{\frac{\log(2SAT^2 2^S/\delta)}{N_t^+(s,a)}}\right)\\
\ge&1-\frac{\delta}{2SAT2^S},
\end{aligned} \tag{B.24}$$

since then the required inequality follows from a union bound over all $s\in\mathscr{S}, a\in\mathscr{A}_s, t\in[T], \theta\in\{-1.1\}^{\mathscr{S}}$. Similar to the casea of $\mathscr{E}_r$, (B.24) follows from Proposition 51, with $\{\mathscr{F}_q\}_{q=1}^T, \{\xi_q\}_{q=1}^T, \{Y_q\}_{q=1}^T, \delta$ defined as

$$\begin{aligned}
\mathscr{F}_q &= \{(s_\ell, a_\ell)\}_{\ell=1}^{q+1},\\
\xi_q &= \left[\sum_{s'\in\mathscr{S}}\theta(s')\mathbf{1}(s_q=s,a_q=a,s_{q+1}=s')\right] - \left[\sum_{s'\in\mathscr{S}}\theta(s')p_q(s'|s,a)\right],\\
Y_q &= \mathbf{1}\left(s_q=s,a_q=a,((t-W)\vee 1)\le q\le t-1\right),
\end{aligned}$$

$$\delta' = \frac{\delta}{2SAT2^S}.$$

Each ξ_q is conditionally 2-sub-Gaussian, since $-1 \le \xi_q \le 1$ with certainty. Altogether, the required inequality is shown.

B.6 Proof of Proposition 44

In this section, we prove Proposition 44. Throughout the section, we impose the assumptions stated by the Proposition. That is, the events $\mathscr{E}_r, \mathscr{E}_p$ hold, and there exists $\mathring{p}$ with (1) $\mathring{p} \in H_{p,\tau(m)}(\eta)$, (2) $(\mathscr{S}, \mathscr{A}, \mathring{p})$ has diameter at most D. We begin by recalling the following notations:

$$B_{r,t} = \max_{s\in\mathscr{S}, a\in\mathscr{A}_s} |r_{t+1}(s,a) - r_t(s,a)|, \quad B_{p,t} = \max_{s\in\mathscr{S}, a\in\mathscr{A}_s} \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1,$$
$$\mathsf{var}_{r,t} = \sum_{q=\tau(m)-W}^{t-1} B_{r,q}, \quad \mathsf{var}_{p,t} = \sum_{q=\tau(m)-W}^{t-1} B_{p,q}.$$

We then need the following auxiliary lemmas

Lemma 52. *Let t be in episode m. For every state-action pair (s,a), we have*

$$\left|r_t(s,a) - \bar{r}_{\tau(m)}(s,a)\right| \le \mathsf{var}_{r,t}, \quad \left\|p_t(\cdot|s,a) - \bar{p}_{\tau(m)}(\cdot|s,a)\right\|_1 \le \mathsf{var}_{p,t}$$

Lemma 53. *Let t be in episode m. We have*

$$\tilde{\rho}_{\tau(m)} \ge \rho_t^* - \mathsf{var}_{r,t} - 2D \cdot \mathsf{var}_{p,t}.$$

Lemma 54. *Let t be in episode m. For every state-action pair (s,a), we have*

$$\left|\sum_{s'\in\mathscr{S}} \tilde{p}_{\tau(m)}(s'|s,a)\tilde{\gamma}_{\tau(m)}(s') - \sum_{s'\in\mathscr{S}} p_t(s'|s,a)\tilde{\gamma}_{\tau(m)}(s')\right| \le 2D\left[\mathsf{var}_{p,t} + 2\mathsf{rad}\text{-}_{p,\tau(m)}(s,a) + \eta\right].$$

Lemmas 52, 53, 54 are proved in Sections B.6.1, B.6.2, and B.6.3, respectively.

B.6.1 Proof of Lemma 52

We first provide the bound for rewards:

$$\begin{aligned}\left|r_t(s,a)-\bar{r}_{\tau(m)}(s,a)\right| &\leq \left|r_t(s,a)-r_{\tau(m)}(s,a)\right|+\left|r_{\tau(m)}(s,a)-\bar{r}_{\tau(m)}(s,a)\right| \\ &\leq \sum_{q=\tau(m)}^{t-1}\left|r_{q+1}(s,a)-r_q(s,a)\right|+\frac{1}{W}\sum_{w=1}^{W}\left|r_{\tau(m)}(s,a)-r_{\tau(m)-w}(s,a)\right|.\end{aligned}$$

By the definition of $B_{r,q}$, we have

$$\sum_{q=\tau(m)}^{t-1}\left|r_{q+1}(s,a)-r_q(s,a)\right| \leq \sum_{q=\tau(m)}^{t-1} B_{r,q},$$

and

$$\begin{aligned}\frac{1}{W}\sum_{w=1}^{W}\left|r_{\tau(m)}(s,a)-r_{\tau(m)-w}(s,a)\right| &\leq \frac{1}{W}\sum_{w=1}^{W}\sum_{i=1}^{w}\left|r_{\tau(m)-i+1}(s,a)-r_{\tau(m)-i}(s,a)\right| \\ &\leq \frac{1}{W}\sum_{w=1}^{W}\sum_{i=1}^{W}\left|r_{\tau(m)-i+1}(s,a)-r_{\tau(m)-i}(s,a)\right| \\ &= \sum_{i=1}^{W}\left|r_{\tau(m)-i+1}(s,a)-r_{\tau(m)-i}(s,a)\right| \leq \sum_{i=1}^{W} B_{r,\tau(m)-i}.\end{aligned}$$

Next, we provide a similar analysis on the transition distribution.

$$\begin{aligned}\left\|p_t(s,a)-\bar{p}_{\tau(m)}(s,a)\right\|_1 &\leq \left\|p_t(s,a)-p_{\tau(m)}(s,a)\right\|_1+\left\|p_{\tau(m)}(s,a)-\bar{p}_{\tau(m)}(s,a)\right\|_1 \\ &\leq \sum_{q=\tau(m)}^{t-1}\left\|p_{q+1}(s,a)-p_q(s,a)\right\|_1+\frac{1}{W}\sum_{w=1}^{W}\left\|p_{\tau(m)}(s,a)-p_{\tau(m)-w}(s,a)\right\|_1.\end{aligned}$$

By the definition of $B_{p,q}$, we have

$$\sum_{q=\tau(m)}^{t-1}\left\|p_{q+1}(s,a)-p_q(s,a)\right\|_1 \leq \sum_{q=\tau(m)}^{t-1} B_{p,q},$$

and

$$\begin{aligned}\frac{1}{W}\sum_{w=1}^{W}\left\|p_{\tau(m)}(s,a)-p_{\tau(m)-w}(s,a)\right\|_1 &\le \frac{1}{W}\sum_{w=1}^{W}\sum_{i=1}^{w}\left\|p_{\tau(m)-i+1}(s,a)-p_{\tau(m)-i}(s,a)\right\|_1\\ &\le \frac{1}{W}\sum_{w=1}^{W}\sum_{i=1}^{W}\left\|p_{\tau(m)-i+1}(s,a)-p_{\tau(m)-i}(s,a)\right\|_1\\ &= \sum_{i=1}^{W}\left\|p_{\tau(m)-i+1}(s,a)-p_{\tau(m)-i}(s,a)\right\|_1 \le \sum_{i=1}^{W} B_{p,\tau(m)-i}.\end{aligned}$$

Altogether, the lemma is shown.

B.6.2 Proof of Lemma 53

We first demonstrate two immediate consequences about the dual solution $(\tilde{\rho}_{\tau(m)}, \tilde{\gamma}_{\tau(m)})$ by the Proposition's assumptions:

$$0 \le \tilde{\gamma}_{\tau(m)}(s) \le 2D \qquad \text{for all } s \in \mathscr{S}, \tag{B.25}$$

$$\tilde{\rho}_{\tau(m)} + \tilde{\gamma}_{\tau(m)}(s) \ge \bar{r}_{\tau(m)}(s,a) + \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s')\bar{p}_{\tau(m)}(s'|s,a) \qquad \text{for all } s \in \mathscr{S}, a \in \mathscr{A}_s. \tag{B.26}$$

To see inequality (B.25), first observe that

$$\tilde{\rho}_{\tau(m)} + \tilde{\gamma}_{\tau(m)}(s) \ge \max_{\dot{r}(s,a) \in H_{r,\tau(m)}(s,a)} \{\dot{r}(s,a)\} + \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') \max_{\dot{p} \in H_{p,\tau(m)}(s,a;\eta)} \{\dot{p}(s'|s,a)\} \tag{B.27}$$

$$\ge \max_{\dot{r}(s,a) \in H_{r,\tau(m)}(s,a)} \{\dot{r}(s,a)\} + \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s')\mathring{p}(s'|s,a). \tag{B.28}$$

Step (B.27) is by Property 1 of the output from EVI, which is applied with confidence regions $H_{r,\tau(m)}, H_{p,\tau(m)}(\eta)$. Step (B.28) is because of the assumption that $\mathring{p} \in H_{p,\tau(m)}(\eta)$. Altogether, the solution $(\tilde{\rho}_{\tau(m)}, \tilde{\gamma}_{\tau(m)})$ is feasible to $\mathsf{D}(\dot{r}, \mathring{p})$ for any $\dot{r} \in H_{r,\tau(m)}$. Now, by Lemma 40, we have $\max_{s,s' \in \mathscr{S}} |\tilde{\gamma}_{\tau(m)}(s) - \tilde{\gamma}_{\tau(m)}(s')| \le 2D$. Finally, inequality (B.25) follows from the fact that the bias vector $\tilde{\gamma}_{\tau(m)}$ returned by EVI is component-wise non-

negative, and there exists $s \in \mathscr{S}$ such that $\tilde{\gamma}_{\tau(m)} = 0$.

To see inequality (B.26), observe that

$$\tilde{\rho}_{\tau(m)} + \tilde{\gamma}_{\tau(m)}(s) \geq \max_{\dot{r}(s,a) \in H_{r,\tau(m)}(s,a)} \{\dot{r}(s,a)\} + \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') \max_{\dot{p} \in H_{p,\tau(m)}(s,a;\eta)} \{\dot{p}(s'|s,a)\} \tag{B.29}$$

$$\geq \bar{r}_{\tau(m)}(s,a) + \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') \bar{p}_{\tau(m)}(s'|s,a). \tag{B.30}$$

Step (B.29) is again by Property 1 of the output from EVI, and step (B.30) is by the assumptions that $\bar{r}_{\tau(m)} \in H_{r,\tau(m)}$, and $\bar{p}_{\tau(m)} \in H_{p,\tau(m)}(0) \subset H_{p,\tau(m)}(\eta)$.

Now, we claim that $(\tilde{\rho}_{\tau(m)} + \mathsf{var}_{r,t} + 2D \cdot \mathsf{var}_{p,t}, \tilde{\gamma}_{\tau(m)})$ is a feasible solution to the tth period dual problem $\mathsf{D}(r_t, p_t)$, which immediately implies the Lemma. To demonstrate the claim, for every state-action pair (s,a) we have

$$\bar{r}_{\tau(m)}(s,a) \geq r_t(s,a) - \mathsf{var}_{r,t} \tag{B.31}$$

$$\sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') p_{\tau(m)}(s'|s,a) \geq \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') p_t(s'|s,a) - \left\|\tilde{\gamma}_{\tau(m)}\right\|_\infty \left\|p_t(\cdot|s,a) - \bar{p}_{\tau(m)}(\cdot|s,a)\right\|_1$$

$$\geq \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') p_t(s'|s,a) - 2D \cdot \mathsf{var}_{p,t}, . \tag{B.32}$$

Inequality (B.31) is by Lemma 52 on the rewards. Step (B.32) is by inequality (B.25), and by Lemma 52 which shows $\|p_t(\cdot|s,a) - \bar{p}_{\tau(m)}(\cdot|s,a)\|_1 \leq \mathsf{var}_{p,t}$. Altogether, putting (B.31), (B.32) to inequality (B.26), our claim is shown, *i.e.*, for all $s \in \mathscr{S}$ and $a \in \mathscr{A}_s$,

$$\tilde{\rho}_{\tau(m)} + \mathsf{var}_{r,t} + 2D \cdot \mathsf{var}_{p,t} + \tilde{\gamma}_{\tau(m)}(s) \geq r_t(s,a) + \sum_{s' \in \mathscr{S}} \tilde{\gamma}_{\tau(m)}(s') p_t(s'|s,a).$$

Hence, the lemma is proved.

B.6.3 Proof of Lemma 54

We have

$$\left| \sum_{s' \in \mathscr{S}} \left[\tilde{p}_{\tau(m)}(s'|s,a) - p_t(s'|s,a)\right] \tilde{\gamma}_{\tau(m)}(s') \right|$$

$$\leq \underbrace{\left\|\tilde{\gamma}_{\tau(m)}\right\|_\infty}_{(a)} \cdot \left[\underbrace{\left\|\tilde{p}_{\tau(m)}(\cdot|s,a) - \bar{p}_{\tau(m)}(\cdot|s,a)\right\|_1}_{(b)} + \underbrace{\left\|\bar{p}_{\tau(m)}(\cdot|s,a) - p_t(\cdot|s,a)\right\|_1}_{(c)} \right]. \tag{B.33}$$

In step (B.33), we know that

- $(a) \leq 2D$ by inequality (B.25),
- $(b) \leq 2\mathsf{rad}\text{-}_{p,\tau(m)}(s,a) + \eta$, by the facts that $\tilde{p}_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a;\eta)$ and $\bar{p}_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a;0)$,
- $(c) \leq \mathsf{var}_{p,t}$ by Lemma 52 on the bound on p.

Altogether, the Lemma is proved.

B.6.4 Finalizing the Proof

Now, we have

$$\begin{aligned}
& r_t(s_t,a_t) \\
\geq & \bar{r}_{\tau(m)}(s_t,a_t) - \mathsf{var}_{r,t} && \text{(B.34)} \\
\geq & \tilde{r}_{\tau(m)}(s_t,a_t) - \mathsf{var}_{r,t} - 2\cdot\mathsf{rad}\text{-}_{r,\tau(m)}(s_t,a_t) && \text{(B.35)} \\
\geq & \tilde{\rho}_{\tau(m)} + \tilde{\gamma}_{\tau(m)}(s_t) - \left[\sum_{s'\in\mathscr{S}} \tilde{p}_{\tau(m)}(s'|s_t,a_t)\tilde{\gamma}_{\tau(m)}(s')\right] - \frac{1}{\sqrt{\tau(m)}} \\
& \qquad - \mathsf{var}_{r,t} - 2\cdot\mathsf{rad}\text{-}_{r,\tau(m)}(s_t,a_t) && \text{(B.36)} \\
\geq & \rho_t^* + \tilde{\gamma}_{\tau(m)}(s_t) - \left[\sum_{s'\in\mathscr{S}} p_t(s'|s_t,a_t)\tilde{\gamma}_{\tau(m)}(s')\right] - \frac{1}{\sqrt{\tau(m)}} \\
& - 2\left[\mathsf{var}_{r,t} + \mathsf{rad}\text{-}_{r,\tau(m)}(s_t,a_t)\right] - 2D\left[2\cdot\mathsf{var}_{p,t} + 2\cdot\mathsf{rad}\text{-}_{p,\tau(m)}(s,a) + \eta\right]. && \text{(B.37)}
\end{aligned}$$

Step (B.34) is by Lemma 52 on t. Step (B.35) is by conditioning that event $\mathscr{E}_r$ holds. Step (B.36) is by Property 2 for the output of EVI. In step (B.37), we upper bound $\tilde{\rho}_{\tau(m)}$ by Lemma 53 and we upper bound $\sum_{s'\in\mathscr{S}} \tilde{p}_{\tau(m)}(s'|s_t,a_t)\tilde{\gamma}_{\tau(m)}(s')$ by Lemma 54. Rearranging gives the Proposition.

B.7 Proof of Lemma 45

We first claim that, for every episode $m \in Q_T$, there exists some state-action pair (s,a) and some time step $t_m \in [(\tau(m) - W \vee 1), \tau(m) - 1]$ such that

$$\|p_{\tau(m)}(\cdot|s,a) - p_{t_m}(\cdot|s,a)\|_1 > \eta. \tag{B.38}$$

For contradiction sake, suppose the otherwise, that is, $\|p_{\tau(m)}(\cdot|s,a) - p_t(\cdot|s,a)\|_1 \leq \eta$ for every state-action pair s,a and every time step $t \in [(\tau(m) - W \vee 1), \tau(m) - 1]$. For each state-action pair (s,a), consider the following cases on $N_{\tau(m)}(s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{\tau(m)-1} \mathbf{1}(s_q = s, a_q = a)$:

- **Case 1:** $N_{\tau(m)}(s,a) = 0$**.** Then $\hat{p}_{\tau(m)}(\cdot|s,a) = \mathbf{0}^{\mathscr{S}}$, and clearly we have

$$\|p_{\tau(m)}(\cdot|s,a) - \hat{p}_{\tau(m)}(\cdot|s,a)\|_1 = 1 < \mathsf{rad}\text{-}p_{\tau(m)}(s,a) < \mathsf{rad}\text{-}p_{\tau(m)}(s,a) + \eta.$$

- **Case 2:** $N_{\tau(m)}(s,a) > 0$**.** Then we have the following inequalities:

$$\begin{aligned}
&\|p_{\tau(m)}(\cdot|s,a) - \bar{p}_{\tau(m)}(\cdot|s,a)\|_1 \\
=&\left\| \frac{1}{N^+_{\tau(m)}(s,a)} \sum_{q=(\tau(m)-W)\vee 1}^{\tau(m)-1} \left[p_{\tau(m)}(\cdot|s,a) - p_q(\cdot|s,a)\right] \cdot \mathbf{1}(s_q = s, a_q = a) \right\|_1 \qquad \text{(B.39)} \\
\leq&\frac{1}{N^+_{\tau(m)}(s,a)} \sum_{q=(\tau(m)-W)\vee 1}^{\tau(m)-1} \left\|p_{\tau(m)}(\cdot|s,a) - p_q(\cdot|s,a)\right\|_1 \cdot \mathbf{1}(s_q = s, a_q = a) \leq \eta. \qquad \text{(B.40)}
\end{aligned}$$

Step (B.39) is by the definition of $\bar{p}_{\tau(m)}(\cdot|s,a)$, and step (B.40) is by the triangle inequality. Consequently, we have

$$\begin{aligned}
&\|p_{\tau(m)}(\cdot|s,a) - \hat{p}_{\tau(m)}(\cdot|s,a)\|_1 \\
\leq&\|\bar{p}_{\tau(m)}(\cdot|s,a) - \hat{p}_{\tau(m)}(\cdot|s,a)\|_1 + \|p_{\tau(m)}(\cdot|s,a) - \bar{p}_{\tau(m)}(\cdot|s,a)\|_1 \qquad \text{(B.41)} \\
\leq&\mathsf{rad}\text{-}p_{\tau(m)}(s,a) + \eta.
\end{aligned}$$

Step (B.41) is true since we condition on the event $\mathscr{E}_p$,

Combining the two cases, we have shown that $p_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a,\eta)$ for all $s \in \mathscr{S}, a \in \mathscr{A}_s$, contradicting to the fact that $m \in Q_T$. Altogether, our claim on inequality (B.38) is established.

Finally, we provide an upper bound to $|Q_T|$ using (B.38):

$$\begin{aligned} B_p &= \sum_{t\in[T-1]} \max_{s\in\mathscr{S},a\in\mathscr{A}_s} \{\|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1\} \\ &\geq \sum_{m\in Q_T} \sum_{q=t_m}^{\tau(m)-1} \max_{s\in\mathscr{S},a\in\mathscr{A}_s} \left\{\left\|p_{q+1}(\cdot|s,a) - p_q(\cdot|s,a)\right\|_1\right\} \qquad \text{(B.42)} \\ &\geq \sum_{m\in Q_T} \max_{s\in\mathscr{S},a\in\mathscr{A}_s} \left\{\left\|\sum_{q=t_m}^{\tau(m)-1} (p_{q+1}(\cdot|s,a) - p_q(\cdot|s,a))\right\|_1\right\} \qquad \text{(B.43)} \\ &> |Q_T|\eta. \qquad \text{(B.44)} \end{aligned}$$

Step (B.42) follows by the second criterion of the construction of Q_T, which ensures that for distinct $m, m' \in Q_T$, the time intervals $[t_m, \tau(m)]$, $[t_{m'}, \tau(m')]$ are disjoint. Step (B.43) is by applying the triangle inequality, for each $m \in Q_T$, on the state-action pair (s,a) that maximizes the term $\|\sum_{q=t_m}^{\tau(m)-1}(p_{q+1}(\cdot|s,a) - p_q(\cdot|s,a))\|_1 = \|(p_{\tau(m)}(\cdot|s,a) - p_{t_m}(\cdot|s,a))\|_1$. Step (B.44) is by applying the claimed inequality (B.38) on each $m \in Q_T$. Altogether, the Lemma is proved. ■

B.8 Proof of Lemma 46

We prove by contradiction. Suppose there exists an episode $m \notin \tilde{Q}_T$, a state $s \in \mathscr{S}$, and an action $a \in \mathscr{A}_s$ such that $p_{\tau(m)}(\cdot|s,a) \notin H_{p,\tau(m)}(s,a;\eta)$, then m should have been added to Q_T. To see this, we first note that episode m trivially satisfies criterion 1 in the construction of Q_T. Moreover, at the time when m is examined, we know that any m' has been added to Q_T should satisfy $\tau(m) - \tau(m') > W$ as otherwise m would have been added to $\tilde{Q}_T$. Therefore, we have prove $m \in Q_T \subseteq \tilde{Q}_T$, which is clearly a contradiction.

B.9 Proof of Lemma 47

Denote $Q_T = \{m_1, \ldots, m_{|Q_T|}\}$. By construction, for every element $m \in \tilde{Q}_T$, there exists an unique $m' \in Q_T$ such that

$$\tau(m) - \tau(m') \in [0, W]. \tag{B.45}$$

We can thereby partition the elements of $\tilde{Q}_T$ into $|Q_T|$ disjoint subsets $\tilde{Q}_T(m_1), \ldots, \tilde{Q}_T(m_{|Q_T|})$ such that

1. Each element $m \in \tilde{Q}_T$ belongs to exactly one $\tilde{Q}_T(m')$ for some $m' \in Q_T$.
2. Each element $m \in \tilde{Q}_T(m')$ satisfies $\tau(m) - \tau(m') \in [0, W]$.

We bound (♠) from above as

$$\begin{aligned}
&\sum_{m' \in Q_T} \sum_{m \in \tilde{Q}_T(m')} \sum_{t=\tau(m)}^{\tau(m+1)-1} (\rho_t^* - r_t(s_t, a_t)) \\
\leq &\sum_{m' \in Q_T} \sum_{m \in \tilde{Q}_T(m')} \sum_{t=\tau(m)}^{\tau(m+1)-1} 1 && \text{(B.46)} \\
= &\sum_{m' \in Q_T} \sum_{m \in \tilde{Q}_T(m')} (\tau(m+1) - \tau(m)) \\
\leq &\sum_{m' \in Q_T} \left(\max_{m \in \tilde{Q}_T(m')} \tau(m+1) - \tau(m') \right) && \text{(B.47)} \\
= &\sum_{m' \in Q_T} \left[\max_{m \in \tilde{Q}_T(m')} (\tau(m+1) - \tau(m) + \tau(m)) - \tau(m') \right] \\
\leq &\sum_{m' \in Q_T} \left[\max_{m \in \tilde{Q}_T(m')} (\tau(m+1) - \tau(m)) + \max_{m \in \tilde{Q}_T(m')} \tau(m) - \tau(m') \right] \\
\leq &\sum_{m' \in Q_T} [W + W] && \text{(B.48)} \\
= &2W|Q_T| \\
\leq &2B_p W/\eta.
\end{aligned}$$

Here, inequality (B.46) holds by boundedness of rewards, inequality (B.47) follows from

the fact that episodes are mutually disjoint, inequality (B.48) makes the observations that each episode can last for at most W time steps (imposed by the SWUCRL2-CW algorithm) as well as criterion 2 of the construction of $\tilde{Q}_T(m')$'s, and the last step uses Lemma 45.

B.10 Proof of Lemma 48

We first give an upper bound for $M(T)$, the total number of the episodes.

Lemma 55. *Conditioned on events $\mathscr{E}_r, \mathscr{E}_p$, we have $M(T) \leq SA(2+\log_2 W)T/W = \tilde{O}(SAT/W)$ with certainty.*

Proof. First, to demonstrate the bound for $M(T)$, it suffices to show that there are at most $SA(2+\log_2 W)$ many episodes in each of the following cases: (1) between time steps 1 and W, (2) between time steps jW and $(j+1)W$, for any $j \in \{1,\ldots,\lfloor T/W \rfloor - 1\}$, (3) between time steps $\lfloor T/W \rfloor \cdot W$ and T. We focus on case (2), and the edge cases (1, 3) can be analyzed similarly.

Between time steps jW and $(j+1)W$, a new episode $m+1$ is started when the second criterion $v_m(s_t, \tilde{\pi}_m(s_t)) < N^+_{\tau(m)}(s_t, \tilde{\pi}_m(s_t))$ is violated during the current episode m. We first illustrate how the second criterion is invoked for a fixed state-action pair (s,a), and then bound the number of invocations due to (s,a). Now, let's denote $m^1,\ldots,m^L$ as the episode indexes, where $jW \leq \tau(m^1) < \tau(m^2) < \ldots < \tau(m^L) < (j+1)W$, and the second criterion for (s,a) is invoked during m^ℓ for $1 \leq \ell \leq L$. That is, for each $\ell \in \{1,\ldots,L\}$, the DM breaks the **while** loop for episode m^ℓ because $v_{m^\ell}(s,a) = N^+_{\tau(m^\ell)}(s,a)$, leading to the new episode $m^\ell + 1$.

To demonstrate our claimed bound for $M(T)$, we show that $L \leq 2+\log_2 W$ as follows. To ease the notation, let's denote $\psi^\ell = v_{m^\ell}(s,a)$. We first observe that $\psi^1 = N^+_{\tau(m^1)}(s,a) \geq 1$. Next, we note that for $\ell \in \{2,\ldots L\}$, we have $\psi^\ell \geq \sum_{k=1}^{\ell-1} \psi^k$. [1] Indeed, we know that for each ℓ we have $(\tau(m^\ell+1)-1) - \tau(m^1) \leq W$, by our assumption on $m^1,\ldots,m^\ell$. Consequently, the counting sum in $N_{\tau(m^\ell)}(s,a)$, which counts the occurrences of (s,a) in the previous W time steps, must have counted those occurrences corresponding to $\psi^1,\ldots,\psi^{\ell-1}$.

[1] We proceed slightly differently from the stationary case, where the corresponding $N_t(s,a)$ is non-decreasing in t [108], which is clearly not true here due to the use of sliding windows

The worst case sequence of $\psi^1, \psi^2, \ldots, \psi^L$ that yields the largest L is when $\psi^1 = \psi^2 = 1$, $\psi^3 = 2$, and more generally $\psi^\ell = 2^{\ell-2}$ for $\ell \geq 2$. Since $\psi^\ell \leq W$ for all W, we clearly have $L \leq 2 + \log_2 W$, proving our claimed bound on L. Altogether, during the T time steps, there are at most $(SAT(2+\log_2 W))/W$ episodes due to the second criterion and T/W due to the first, leading to our desired bound on $M(T)$. □

Next, we establish the bound for (♣). By Lemma 46, we know that $\tilde{\gamma}_{\tau(m)}(s) \in [0, 2D_{\max}]$ for all $m \in [M(T)] \setminus \tilde{Q}_T$ and s. For each episode $m \in [M(T)] \setminus \tilde{Q}_T$, we have

$$\begin{aligned}
&\sum_{t=\tau(m)}^{\tau(m+1)-1} \left[\sum_{s' \in \mathscr{S}} p_t(s'|s_t, a_t) \tilde{\gamma}_{\tau(m)}(s') - \tilde{\gamma}_{\tau(m)}(s_t) \right] \\
&= \underbrace{-\tilde{\gamma}_{\tau(m)}(s_{\tau(m)}) + \tilde{\gamma}_{\tau(m)}(s_{\tau(m)+1})}_{\leq 2D_{\max}} + \sum_{t=\tau(m)}^{\tau(m+1)-1} \underbrace{\left[\sum_{s' \in \mathscr{S}} p_t(s'|s_t, a_t) \tilde{\gamma}_{\tau(m)}(s') - \tilde{\gamma}_{\tau(m)}(s_{t+1}) \right]}_{=Y_t}.
\end{aligned} \tag{B.49}$$

Summing (B.49) over $m \in [M(T)] \setminus \tilde{Q}_T$ we get

$$\begin{aligned}
&\sum_{m \in [M(T)] \setminus \tilde{Q}_T} \sum_{t=\tau(m)}^{\tau(m+1)-1} \left[\sum_{s' \in \mathscr{S}} p_t(s'|s_t, a_t) \tilde{\gamma}_{\tau(m)}(s') - \tilde{\gamma}_{\tau(m)}(s_t) \right] \\
&\leq 2D_{\max} \cdot \left(M(T) - |\tilde{Q}_T| \right) + \sum_{m \in [M(T)] \setminus \tilde{Q}_T} \sum_{t=\tau(m)}^{\tau(m+1)-1} Y_t.
\end{aligned} \tag{B.50}$$

Define the filtration $\mathscr{H}_{t-1} = \{(s_q, a_q, R_q(s_q, a_q))\}_{q=1}^t$. Now, we know that $\mathbb{E}[Y_t | \mathscr{H}_{t-1}] = 0$, Y_t is $\sigma(\mathscr{H}_t)$-measurable, and $|Y_t| \leq 2D_{\max}$. Therefore, we can apply the Hoeffding inequality [104] to show that

$$\sum_{m \in [M(T)] \setminus \tilde{Q}_T} \sum_{t=\tau(m)}^{\tau(m+1)-1} \left[\sum_{s' \in \mathscr{S}} p_t(s'|s_t, a_t) \tilde{\gamma}_{\tau(m)}(s') - \tilde{\gamma}_{\tau(m)}(s_t) \right] = O\left(D_{\max} \cdot M(T) + D_{\max} \sqrt{T \log \frac{1}{\delta}} \right)$$

with probability $1 - O(\delta)$, where we use the facts that

$$M(T) - |\tilde{Q}_T| \leq M(T)$$

and

$$\sum_{m\in[M(T)]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}1\leq T.$$

Finally, note that

$$\sum_{m\in[M(T)]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}\frac{1}{\sqrt{\tau(m)}}\leq\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}\frac{1}{\sqrt{\tau(m)}}\leq\sum_{i=1}^{\lceil T/W\rceil}\frac{W}{\sqrt{iW}}=O(\sqrt{T}).$$

Hence, the Lemma is proved.

B.11 Proof of Lemma 49

We first note that

$$\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left(2\mathsf{var}_{r,t}+4D_{\max}\cdot\mathsf{var}_{p,t}+2D_{\max}\eta\right)\leq\sum_{t=1}^{T}\left(2\mathsf{var}_{r,t}+4D_{\max}\cdot\mathsf{var}_{p,t}+2D_{\max}\eta\right),$$

since $\mathsf{var}_{r,t}\geq 0$ and $\mathsf{var}_{p,t}\geq 0$ for all t. We can thus work with the latter quantity.

We first bound $\sum_{t=1}^{T}\mathsf{var}_{r,t}$. Now, recall the definition that, for time t in episode m, we have defined $\mathsf{var}_{r,t}=\sum_{q=\tau(m)-W}^{t-1}B_{r,q}$. Clearly, for $iW\leq q<(i+1)W$, the summand $B_{r,q}$ only appears in $\mathsf{var}_{r,t}$ for $iW\leq q<t\leq(i+2)W$, since each episode is contained in $\{i'W,\ldots,(i'+1)W\}$ by our episode termination criteria (t is a multiple of W) of the `SWUCRL2-CW` algorithm. Altogether, we have

$$2\sum_{t=1}^{T}\mathsf{var}_{r,t}\leq 2\sum_{t=1}^{T-1}B_{r,t}W=2B_rW. \tag{B.51}$$

Next, we bound $\sum_{t=1}^{T}\mathsf{var}_{p,t}$. Now, we know that $\tau(m+1)-\tau(m)\leq W$ by our episode termination criteria (t is a multiple of W) of the `SWUCRL2-CW` algorithm. Consequently,

$$4D_{\max}\sum_{t=1}^{T}\mathsf{var}_{p,t}\leq 4D_{\max}\sum_{t=1}^{T-1}B_{p,t}W=4D_{\max}B_pW. \tag{B.52}$$

Finally, combining (B.51, B.52) with $2D_{\max}\sum_{t=1}^{T}\eta$, the Lemma is established.

B.12 Proof of Lemma 50

Due to non-negativity of rad-$r_t(s,a)$'s and rad-$p_t(s,a)$'s, we have

$$\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}\text{rad-}r_{\tau(m)}(s_t,a_t)\le\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}\text{rad-}r_{\tau(m)}(s_t,a_t), \tag{B.53}$$

$$\sum_{m\in[M_T]\setminus\tilde{Q}_T}\sum_{t=\tau(m)}^{\tau(m+1)-1}\text{rad-}p_{\tau(m)}(s_t,a_t)\le\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}\text{rad-}p_{\tau(m)}(s_t,a_t) \tag{B.54}$$

We thus first show that, with certainty,

$$\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}\text{rad-}r_{\tau(m)}(s_t,a_t)=\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}2\sqrt{\frac{2\ln(SAT/\delta)}{N^+_{\tau(m)}(s_t,a_t)}}=\tilde{O}\left(\frac{\sqrt{SA}T}{\sqrt{W}}\right), \tag{B.55}$$

$$\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}\text{rad-}p_{\tau(m)}(s_t,a_t)=\sum_{m=1}^{M(T)}\sum_{t=\tau(m)}^{\tau(m+1)-1}2\sqrt{\frac{2S\log(ATW/\delta)}{N^+_{\tau(m)}(s_t,a_t)}}=\tilde{O}\left(\frac{S\sqrt{A}T}{\sqrt{W}}\right). \tag{B.56}$$

We analyze by considering the dynamics of the algorithm in each consecutive block of W time steps, in a way similar to the proof of Lemma 48. Consider the episodes indexes $m_0,m_1\ldots,m_{\lceil T/W\rceil},m_{\lceil T/W\rceil+1}$, where $\tau(m_0)=1$, and $\tau(m_j)=jW$ for $j\in\{1,\ldots,\lceil T/W\rceil\}$, and $m_{\lceil T/W\rceil+1}=m(T)+1$ (so that $\tau(m_{\lceil T/W\rceil+1}-1)$ is the time index for the last episode in the horizon).

To prove (B.55, B.56), it suffices to show that, for each $j\in\{0,1,\ldots,\lfloor T/W\rfloor\}$, we have

$$\sum_{m=m_j}^{m_{j+1}-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\frac{1}{\sqrt{N^+_{\tau(m)}(s_t,a_t)}}=O\left(\sqrt{SAW}\right). \tag{B.57}$$

Without loss of generality, we assume that $j\in\{1,\ldots,\lfloor T/W\rfloor-1\}$, and the edge cases of $j=0,\lfloor T/W\rfloor$ can be analyzed similarly.

Now, we fix a state-action pair (s,a) and focus on the summands in (B.57):

$$\sum_{m=m_j}^{m_{j+1}-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\frac{\mathbf{1}((s_t,a_t)=(s,a))}{\sqrt{N^+_{\tau(m)}(s_t,a_t)}}=\sum_{m=m_j}^{m_{j+1}-1}\frac{\nu_m(s,a)}{\sqrt{N^+_{\tau(m)}(s,a)}} \tag{B.58}$$

It is important to observe that:

1. It holds that $\nu_{m_j}(s,a) \leq N_{\tau(m_j)}(s,a)$, by the episode termination criteria of the `SWUCRL2-CW` algorithm,

2. For $m \in \{m_j+1,\dots,m_{j+1}-1\}$, we have $\sum_{m'=m_j}^{m-1} \nu_{m'}(s,a) \leq N_{\tau(m)}(s,a)$. Indeed, we know that episodes $m_j,\dots,m_{j+1}-1$ are inside the time interval $\{jW,\dots,(j+1)W\}$. Consequently, the counts of "$(s_t,a_t)=(s,a)$" associated with $\{\nu_{m'}(s,a)\}_{m'=m_j}^{m-1}$ are contained in the W time steps preceding $\tau(m)$, hence the desired inequality.

With these two observations, we have

$$
\begin{aligned}
(\text{B.58}) &\leq \frac{\nu_{m_j}(s,a)}{\sqrt{\max\{\nu_{m_j}(s,a),1\}}} + \sum_{m=m_j+1}^{m_{j+1}-1} \frac{\nu_m(s,a)}{\sqrt{\max\{\sum_{m'=m_j}^{m-1}\nu_{m'}(s,a),1\}}} \\
&\leq \sqrt{\max\{\nu_{m_j}(s,a),1\}} + (\sqrt{2}+1)\sqrt{\max\left\{\sum_{m=m_j}^{m_{j+1}-1}\nu_m(s,a),1\right\}} \qquad \text{(B.59)} \\
&\leq (\sqrt{2}+2)\sqrt{\max\left\{\sum_{t=jW}^{(j+1)W-1}\mathbf{1}((s_t,a_t)=(s,a)),1\right\}}. \qquad \text{(B.60)}
\end{aligned}
$$

Step (B.59) is by Lemma 19 in [108], which bounds the sum in the previous line. Step (B.60) is by the fact that episodes $m_j,\dots,m_{j+1}-1$ partitions the time interval $jW,\dots,(j+1)W-1$, and $\nu_m(s,a)$ counts the occurrences of $(s_t,a_t)=(s,a)$ in episode m. Finally, observe that $(B.58)=0$ if $\nu_m(s,a)=0$ for all $m \in \{m_j,\dots,m_{j+1}-1\}$. Thus, we can refine the bound in (B.60) to

$$
(\text{B.58}) \leq (\sqrt{2}+2)\sqrt{\sum_{t=jW}^{(j+1)W-1}\mathbf{1}((s_t,a_t)=(s,a))}. \qquad \text{(B.61)}
$$

The required inequality (B.57) is finally proved by summing (B.61) over $s \in \mathscr{S} a \in \mathscr{A}$ and applying Cauchy Schwartz:

$$
\sum_{m=m_j}^{m_{j+1}-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\frac{1}{\sqrt{N^+_{\tau(m)}(s_t,a_t)}} = \sum_{s\in\mathscr{S},a\in\mathscr{A}_s}\sum_{m=m_j}^{m_{j+1}-1}\frac{\nu_m(s,a)}{\sqrt{N^+_{\tau(m)}(s,a)}}
$$

$$\leq(\sqrt{2}+2)\sum_{s\in\mathscr{S},a\in\mathscr{A}_s}\sqrt{\sum_{t=jW}^{(j+1)W-1}\mathbf{1}((s_t,a_t)=(s,a))}$$
$$\leq(\sqrt{2}+2)\sqrt{SAW}.$$

Altogether, the Lemma is proved.

B.13 Proof of Theorem 22

To begin, we consider the following regret decomposition, for any choice of $(W^\dagger,\eta^\dagger)\in J$, we have

$$\begin{aligned}
&\text{Dyn-Reg}_T(\texttt{BORL})\\
&=\sum_{i=1}^{\lceil T/H\rceil}\mathbb{E}\left[\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\rho_t^*-\mathsf{R}_i\left(W_i,\eta_i,s_{(i-1)H+1}\right)\right]+\sum_{t=1}^{T}\mathbb{E}[r_t(s_t^{\Pi^*},a_t^{\Pi^*})]-\sum_{t=1}^{T}\rho_t^*\\
&=\sum_{i=1}^{\lceil T/H\rceil}\mathbb{E}\left[\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\rho_t^*-\mathsf{R}_i\left(W^\dagger,\eta^\dagger,s_{(i-1)H+1}\right)\right]\\
&+\sum_{i=1}^{\lceil T/H\rceil}\mathbb{E}\left[\sum_{i=1}^{\lceil T/H\rceil}\mathsf{R}_i\left(W^\dagger,\eta^\dagger,s_{(i-1)H+1}\right)-\mathsf{R}_i\left(W_i,\eta_i,s_{(i-1)H+1}\right)\right]+\sum_{t=1}^{T}\mathbb{E}[r_t(s_t^{\Pi^*},a_t^{\Pi^*})]-\sum_{t=1}^{T}\rho_t^*.
\end{aligned}\tag{B.62}$$

For the first term in eqn. (B.62), we can apply the results from Theorem 21 to each block $i\in\lceil T/H\rceil$, *i.e.*,

$$\begin{aligned}
&\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\left[\rho_t^*-\mathsf{R}\left(W^\dagger,\eta^\dagger,s_{(i-1)H+1}\right)\right]\\
&=\tilde{O}\left(\frac{B_p(i)W^\dagger}{\eta^\dagger}+B_r(i)W^\dagger+D_{\max}\left[B_p(i)W^\dagger+\frac{S\sqrt{A}H}{\sqrt{W^\dagger}}+H\eta^\dagger+\frac{SAH}{W^\dagger}+\sqrt{H}\right]\right),
\end{aligned}\tag{B.63}$$

where we have defined

$$B_r(i)=\left(\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}B_{r,t}\right),\quad B_p(i)=\left(\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}B_{p,t}\right)$$

for brevity. For the second term, it captures the additional rewards of the DM were it uses the fixed parameters $(W^\dagger, \eta^\dagger)$ throughout w.r.t. the trajectory on the starting states of each block by the BORL algorithm, *i.e.*, $s_1, \ldots, s_{(i-1)H+1}, \ldots, s_{(\lceil T/H \rceil - 1)H+1}$, and this is exactly the regret of the EXP3.P algorithm when it is applied to a Δ-arm adaptive adversarial bandit problem with reward from $[0, H]$. Therefore, for any choice of $(W^\dagger, \eta^\dagger)$, we can upper bound this by

$$\tilde{O}\left(H\sqrt{\Delta T/H}\right) = \tilde{O}\left(\sqrt{TH}\right)$$

as $\Delta = O\left(\ln^2 T\right)$. Summing these two, the regret of the BORL algorithm is

$$\tilde{O}\left(\frac{B_p W^\dagger}{\eta^\dagger} + B_r W^\dagger + D_{\max}\left[B_p W^\dagger + \frac{S\sqrt{A}T}{\sqrt{W^\dagger}} + T\eta^\dagger + \frac{SAT}{W^\dagger} + \sqrt{TH}\right] + \sqrt{D_{\max}(B_r + 2D_{\max}B_p)T}\right). \tag{B.64}$$

By virtue of the EXP3.P, the BORL algorithm is able to adapt to any choice of $(W^\dagger, \eta\dagger) \in J$. Note that

$$H \geq W* = \frac{3S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{1}{2}}}{(B_r + B_p + 1)^{\frac{1}{2}}} \geq \frac{3T^{\frac{1}{2}}}{(3T)^{\frac{1}{2}}} \geq 1, \tag{B.65}$$

$$S^{\frac{1}{3}}A^{\frac{1}{4}} \geq \eta^* = \frac{(B_p+1)^{\frac{1}{2}}S^{\frac{1}{3}}A^{\frac{1}{4}}}{(B_r + B_p + 1)^{\frac{1}{4}}T^{\frac{1}{4}}} \geq \frac{S^{\frac{1}{3}}A^{\frac{1}{4}}}{2T^{\frac{1}{2}}} = S^{\frac{1}{3}}A^{\frac{1}{4}}\Phi. \tag{B.66}$$

Therefore, there must exists some $j^\dagger$ and $k^\dagger$ such that

$$H^{j^\dagger/\Delta_W} \leq W^* \leq H^{(j^\dagger+1)/\Delta_W}, \quad S^{\frac{1}{3}}A^{\frac{1}{4}}\Phi^{k^\dagger/\Delta_\eta} \geq \eta^* \geq S^{\frac{1}{3}}A^{\frac{1}{4}}\Phi^{(k^\dagger+1)/\Delta_\eta} \tag{B.67}$$

By adapting $W^\dagger$ to $H^{j^\dagger/\Delta_W}$ and $\eta^\dagger$ to $\Phi^{k^\dagger/\Delta_\eta}$, we further upper bound eqn. (B.64) as

$$\text{Dyn-Reg}_T(\texttt{BORL}) = \tilde{O}\left(D_{\max}(B_r + B_p + 1)^{\frac{1}{4}}S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{3}{4}}\right).$$

where we use $H^{1/\Delta_W} = \exp(1)$ and $\Phi^{1/\Delta_\eta} = \exp(-1)$ in the last step.

B.14 Proof of Proposition 23

We first show the following lemma.

Lemma 56. *Conditioned on $\mathscr{E}_p$, there exists a state transition distribution $p \in H_{p,t}(0)$ such that for every pair of states $s, s' \in \mathscr{S}$,*

$$p(s'|s, a_{(s,s')}) \geq \zeta$$

for every time step $t \in [T]$.

The proof of the lemma is provided in Section B.14.1. We then consider the state transition distribution $p \in H_{p,t}(0)$ specified in Lemma 56. For an arbitrary state $s' \in \mathscr{S}$, we consider the policy π such that $\pi(s) = a_{(s,s')}$ for all $s \in \mathscr{S}$ (see Assumption 6 for the definition of $a_{(s,s')}$). Starting from an arbitrary state $s \in \mathscr{S}$, the policy either hits state s' in the next step, which happens with probability at least ζ, or it transits to another state $s'' \neq s'$, which would then hit state s' in the next step with probability at least ζ. Therefore, the hitting process stochstically dominates the Bernoulli process with success probability ζ, and thus the expected hitting time is at most $1/\zeta$.

B.14.1 Proof of Lemma 56

First, we recall the definition of defintion of confidence region $H_{p,t}(s,a;0)$ in eqn. 3.6,

$$H_{p,t}(s,a;0) = \left\{ \dot{p} \in \Delta^{\mathscr{S}} : \|\dot{p}(\cdot|s,a) - \hat{p}_t(\cdot|s,a)\|_1 \leq \mathsf{rad}\text{-}_{p,t}(s,a) \right\}.$$

For every pair of states $s, s' \in \mathscr{S}$, we construct p by distinguishing the following two cases:

- If $N_t(s, a_{(s,s')}) = 0$, then by definition, $\mathsf{rad}\text{-}_{p,t}(s, a_{(s,s')}) \geq 1$, therefore every probability distribution $\bar{p}$ on $\mathscr{S}$ belongs to $H_{p,t}(s,a;0)$. Setting $p(\cdot|s, a_{(s,s')}) = p_t(\cdot|s, a_{(s,s')})$ for any t satisfies the requirement in the Lemma.

- If $N_t(s, a_{(s,s')}) > 0$, then we know from Assumption 6 and eqn. 3.5 that

$$\bar{p}_t(s'|s, a_{(s,s')}) = \frac{1}{N_t^+(s, a_{(s,s')})} \sum_{q=(\tau(m)-W)\vee 1}^{t-1} p_q(s'|s, a_{(s,s')}) \mathbf{1}(s_q = s, a_q = a_{(s,s')}) \geq \zeta.$$

  By conditioning on $\underset{\surd}{\mathscr{E}}$, we know that $\bar{p}_t(\cdot|s, a_{(s,s')}) \in H_{p,t}(s, a_{(s,s')}; 0)$, and we can thus set $p(\cdot|s, a_{(s,s')}) = \bar{p}_t(\cdot|s, a_{(s,s')})$.

Combining the above cases, the transition probability distribution p satisfies the stated inequality in the Lemma, and we conclude the proof.

B.15 Proof of Proposition 25

We first show that for any time step $t \in [T]$, we have $\rho_t^* - \rho_t^{*\text{pseudo}} = -l \cdot \mathbb{E}[X_t]$. From Section B.1.1, we have ρ_t^* is equal to the optimal value of the following linear program $\mathsf{P}(r_t, p_t)$; while $\rho_t^{*\text{pseudo}}$ is equal to the optimal value of the following linear program $\mathsf{P}(r_t^{\text{pseudo}}, p_t)$. The two linear programs has the same set of constraints, and follows from eqn. (3.12), the only difference is that the objective value of $\mathsf{P}(r_t^{\text{pseudo}}, p_t)$ is $l \cdot \mathbb{E}[X_t]$ more than that of $\mathsf{P}(r_t, p_t)$ (note that the summation of $x(s,a)$ over $s \in \mathscr{S}$ and $a \in \mathscr{A}_s$ is 1 from the second constraint of the linear program (B.1)). Therefore, we have

$$\sum_{t=1}^{T} \left(\rho_t^* - \rho_t^{*\text{pseudo}} \right) = \sum_{t=1}^{T} -l \cdot \mathbb{E}[X_t]. \tag{B.68}$$

Next, conditioned on any demand realizations $X_1, \ldots, X_T$, we can show by induction that the trajectory generated by Π on $\mathscr{M}$ and $\mathscr{M}^{\text{pseudo}}$ are exactly the same as they use the same sequence of state transition distributions. Therefore,

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{E}[r_t(s_t(\mathscr{M}), a_t(\mathscr{M}))|\{X_s\}_{s=1}^t] - \sum_{t=1}^{T} \mathbb{E}[r_t^{\text{pseudo}}(s_t(\mathscr{M}^{\text{pseudo}}), a_t(\mathscr{M}^{\text{pseudo}}))|\{X_s\}_{s=1}^t] \\ =& \sum_{t=1}^{T} -l \cdot X_t. \end{aligned} \tag{B.69}$$

Taking expectation on both sides of eqn. (B.69), and combining this with eqn. (B.68), we

can conclude the statement.

Appendix C

Proofs for Chapter 4

We begin by defining some helpful notation. First, let

$$\texttt{REV}\left(\theta, \hat{\theta}, \Sigma, t\right) = \mathbb{E}\left[\sum_{s=1}^{t} p_{i,s} D_{i,s}(p_{i,s}, x_{i,s})\right],$$

be the expected total revenue over t time steps obtained by running $\text{TS}(\mathcal{N}(\hat{\theta}, \Sigma), 0)$ — the Thompson sampling algorithm in Algorithm 5 with the (possibly incorrect) prior $\mathcal{N}\left(\hat{\theta}, \Sigma\right)$ and exploration parameter $\lambda_e = 0$ — in an epoch with true parameter θ. Second, let

$$\texttt{REV}_*\left(\theta, t\right) = \mathbb{E}\left[\sum_{s=1}^{t} p^*_{i,s} D_{i,s}(p^*_{i,s}, x_{i,s})\right],$$

be the expected total revenue over t time steps obtained by the oracle — recall $p^*_{i,s}$ is the oracle price defined in Eq. (4.3) — in an epoch with true parameter θ.

All norms $\|\cdot\|$ refer to the ℓ_2 norm unless stated otherwise.

C.1 Meta oracle Regret Analysis

We first state the following lemma, whose proof is provided in Section C.1.1.

Lemma 57. *For any epoch $i \in [N]$, the length of the random exploration periods $\mathcal{T}_i$ is*

upper bounded by

$$\mathscr{T}_e = \max\left\{6\log_{e/2}(dNT)/c_1, 2\lambda_e/c_0\right\} \tag{C.1}$$

with probability at least $1-2/(N^3T^2)$. *The constants are given by*

$$c_0 = \frac{\lambda_0}{3}\left[\frac{p_{\max}^2+p_{\min}^2+2}{2} - \sqrt{\left(\frac{p_{\max}^2+p_{\min}^2+2}{2}\right)^2 - (p_{\max}-p_{\min})^2}\right],\ c_1 = \frac{c_0}{(1+p_{\max}^2)x_{\max}^2}.$$

In other words, we incur at most logarithmic regret due to the initial random exploration in Algorithm 5.

Proof. (Proof of Theorem 27) The proof proceeds in three steps. We first show that the regret incurred in the initial random exploration steps is negligible. We then map the remaining regret to a linear bandit formulation, and bound the resulting terms.

First, define the event

$$\mathscr{A} = \{\mathscr{T}_i \le \mathscr{T}_e\ \forall i \in [N]\}. \tag{C.2}$$

By Lemma 57, $\Pr(\neg\mathscr{A}) \le 2/(NT)^2$. We can decompose the regret from Algorithm 5 into exploration and non-exploration periods, conditioned on whether or not $\mathscr{A}$ holds:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\theta_i\sim\mathscr{N}(\theta_*,\Sigma_*)}\left[\mathtt{REV}_*(\theta_i,T) - \sum_{t=1}^{\mathscr{T}_i} p_{i,t}^{\mathrm{TS}} D_{i,t}(p_{i,t}^{\mathrm{TS}},x_{i,t}) - \mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i}^{\mathrm{TS}},T-\mathscr{T}_i\right)\right] \\
&= \mathbb{E}\left[\mathtt{REV}_*(\theta_i,T) - \sum_{t=1}^{\mathscr{T}_i} p_{i,t}^{\mathrm{TS}} D_{i,t}(p_{i,t}^{\mathrm{TS}},x_{i,t}) - \mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i}^{\mathrm{TS}},T-\mathscr{T}_i\right)\middle|\neg\mathscr{A}\right]\Pr(\neg\mathscr{A}) \\
&\quad + \mathbb{E}\left[\mathtt{REV}_*(\theta_i,\mathscr{T}_i) - \sum_{t=1}^{\mathscr{T}_i} p_{i,t}^{\mathrm{TS}} D_{i,t}(p_{i,t}^{\mathrm{TS}},x_{i,t})\middle|\mathscr{A}\right] \\
&\quad + \mathbb{E}\left[\left(\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i) - \mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i}^{\mathrm{TS}},T-\mathscr{T}_i\right)\right)\middle|\mathscr{A}\right] \\
&\le \mathbb{E}\left[\frac{2p_{\max}x_{\max}\sqrt{1+p_{\max}^2}\|\theta_i\|}{2N^2T}\right] + \mathbb{E}\left[2p_{\max}x_{\max}\sqrt{1+p_{\max}^2}\mathscr{T}_e\|\theta_i\|\right] \\
&\quad + \mathbb{E}\left[\left(\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i) - \mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i}^{\mathrm{TS}},T-\mathscr{T}_i\right)\right)\middle|\mathscr{A}\right],
\end{aligned} \tag{C.3}$$

where we have used the facts that $\Pr(\neg\mathscr{A}) \leq 2/(NT)^2$, the worst-case regret achievable in a single time period is $2p_{\max}x_{\max}\sqrt{1+p_{\max}^2}\|\theta_i\|$, and $\mathscr{T}_i \leq \mathscr{T}_e$ on the event $\mathscr{A}$.

The first two terms in Eq. (C.3) are $O(1/(N^2T)) + O(\log(dNT)) = \tilde{O}(1)$. To analyze the third term in Eq. (C.3), we construct a mapping between the dynamic pricing and linear bandit problems, in order to leverage existing results on TS and UCB for linear bandits [162, 3]. In particular, we can map the Bayes regret of an epoch

$$\mathop{\mathbb{E}}_{\theta_i\sim\mathscr{N}(\theta_*,\Sigma_*)}\left[\left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i}, T-\mathscr{T}_i\right)\right)\middle| E\right],$$

to the Bayes regret of the Thompson sampling algorithm [162] for a linear bandit instance as follows. Let the unknown parameter $\theta = \begin{pmatrix}\alpha^\top & \beta^\top\end{pmatrix}^\top$ be drawn from the prior $\mathscr{N}(\theta_*, \Sigma_*)$. Take the decision set to be $A_t = \{(px_{i,t}; p^2x_{i,t}) : p \in [p_{\min}, p_{\max}]\}$, where $x_{i,t}$ is the feature vector drawn i.i.d from the feature distribution. Note that the magnitude of the ℓ_2-norm of an action is at most $p_{\max}\sqrt{1+p_{\max}^2}x_{\max}$ and the noise terms are conditionally $(p_{\max}\sigma)$-subgaussian.

Using this mapping, by Theorem 3 of [3] and Lemma 73 in Appendix C.7, the Bayes regret of an epoch is upper bounded as

$$\begin{aligned}&\mathop{\mathbb{E}}_{\theta_i\sim\mathscr{N}(\theta_*,\Sigma_*)}\left[\left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i}, T-\mathscr{T}_i\right)\right)\middle| E\right]\\ &= \mathbb{E}\left[\widetilde{O}\left(\|\theta\|\sqrt{dT}\left(\|\theta\| + \sqrt{d}\right)\right)\right] \;=\; \mathbb{E}\left[\widetilde{O}\left(\|\theta\|^2\sqrt{dT} + \|\theta\|d\sqrt{T}\right)\right]. \qquad \text{(C.4)}\end{aligned}$$

where Eq. (C.4) follows from the facts that (i) the upper bound on the regret of a linear bandit instance scales linearly with the maximum absolute value of the rewards and, (ii) the absolute value of the expected reward (revenue) for each round is upper bounded as

$$\max_{p\in[p_{\min},p_{\max}]} |\langle m, \theta\rangle| \;\leq\; \max_{p\in[p_{\min},p_{\max}]} \|m\|\|\theta\| \;=\; p_{\max}\sqrt{1+p_{\max}^2}x_{\max}\|\theta\| \;=\; O\left(\|\theta\|\right). \qquad \text{(C.5)}$$

To complete the proof, we must bound $\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\|\theta\|^2\right]$. By the "trace trick", we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\|\theta\|^2\right] &= \operatorname{trace}\left(\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\theta\theta^\top\right]\right) \\
&= \operatorname{trace}\left(\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[(\theta-\theta_*)(\theta-\theta_*)^\top+\theta_*\theta^\top+\theta\theta_*^\top-\theta_*\theta_*^\top\right]\right) \\
&= \operatorname{trace}\left(\Sigma_*+\theta_*\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[\theta^\top\right]+\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}[\theta]\,\theta_*^\top-\theta_*\theta_*^\top\right) \\
&= \operatorname{trace}\left(\Sigma_*+2\theta_*\theta_*^\top-\theta_*\theta_*^\top\right) \\
&= \operatorname{trace}(\Sigma_*)+\operatorname{trace}\left(\|\theta_*\|^2\right) \\
&\le d\overline{\lambda}+S^2 \;=\; O(d)\,, \qquad \text{(C.6)}
\end{aligned}$$

where we have used the definition of the covariance matrix $\Sigma_* = \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}\left[(\theta-\theta_*)(\theta-\theta_*)^\top\right]$, and the last step follows from Assumptions 8 and 10. Moreover, by Cauchy-Schwarz inequality, we have

$$\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\theta_*,\Sigma_*)}[\|\theta\|] \;\le\; \sqrt{\mathbb{E}\left[\|\theta\|^2\right]} \;\le\; \sqrt{d\overline{\lambda}+S^2} \;=\; O\left(\sqrt{d}\right). \qquad \text{(C.7)}$$

Substituting Eqs. (C.6) and (C.7) into Eq. (C.4), we obtain that the third term of Eq. (C.3) is $\tilde{O}(d^{3/2}T^{1/2})$. Noting that the first and second terms of Eq. (C.3) contribute $\tilde{O}(1)$ regret, we can bound the total regret of each epoch as $\tilde{O}(d^{3/2}T^{1/2})$.

Since each epoch is mutually independent, the Bayes regret of Algorithm 5 over all N epochs is simply $N\times\tilde{O}(d^{3/2}T^{1/2}) = \tilde{O}(d^{3/2}NT^{1/2})$. □

C.1.1 Proof of Lemma 57

Recall that $V_{i,t} = \sum_{s=1}^{t}\begin{pmatrix}x_{i,s}^\top & p_{i,s}x_{i,s}^\top\end{pmatrix}^\top\begin{pmatrix}x_{i,s}^\top & p_{i,s}x_{i,s}^\top\end{pmatrix}$ is the Fisher information matrix of epoch i after time step t. Lemma 57 states that $\lambda_{\min}(V_{i,\mathscr{T}_e}) \ge \lambda_e$ with high probability. Since $V_{i,t}$ is a random matrix, we will apply the following matrix Chernoff inequality to lower bound its minimum eigenvalue (Note that $\lambda_{\max}\left(\begin{pmatrix}x_{i,s}^\top & p_{i,s}x_{i,s}^\top\end{pmatrix}^\top\begin{pmatrix}x_{i,s}^\top & p_{i,s}x_{i,s}^\top\end{pmatrix}\right) \le (1+p_{\max}^2)x_{\max}^2$).

Lemma 58 (Theorem 3.1 of [169]). *For any $\zeta \in [0,1)$, any real number u, and any $t \leq \mathscr{T}_i$*

$$\Pr\left(\lambda_{\min}(V_{i,t}) \;\geq\; (1-\zeta)u \text{ and } \lambda_{\min}\left(\mathbb{E}\left[V_{i,t}\right]\right) \;\geq\; u\right)\right) \;\geq\; 1 - d\left(\frac{\exp(-\zeta)}{(1-\zeta)^{1-\zeta}}\right)^{c_1 u/c_0}.$$

The above lemma states that the probability that $\lambda_{\min}(V_{i,t})$ is much less than $\lambda_{\min}\left(\mathbb{E}\left[V_{i,t}\right]\right)$ is small. To apply the above result, we must first lower bound the minimum eigenvalue of $\mathbb{E}\left[V_{i,t}\right]$:

Lemma 59. *For all $t \leq \mathscr{T}_i$, the minimum eigenvalue of $\mathbb{E}\left[V_{i,t}\right]$ is lower bounded as*

$$\lambda_{\min}\left(\mathbb{E}\left[V_{i,t}\right]\right) \geq c_0 t.$$

Proof. (Proof of Lemma 59) From linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}\left[V_{i,t}\right] &= \sum_{\tau \text{ even},\, \tau \leq t} \mathbb{E}\left[\begin{pmatrix} x_{i,\tau} \\ p_{i,t} x_{i,\tau} \end{pmatrix}\begin{pmatrix} x_{i,\tau}^\top & p_{i,\tau} x_{i,\tau}^\top \end{pmatrix}\right] + \sum_{\tau \text{ odd},\, \tau \leq i} \mathbb{E}\left[\begin{pmatrix} x_{i,\tau} \\ p_{i,t} x_{i,\tau} \end{pmatrix}\begin{pmatrix} x_{i,\tau}^\top & p_{i,\tau} x_{i,\tau}^\top \end{pmatrix}\right] \\
&\geq \frac{t}{3}\left(\begin{pmatrix} \mathbb{E}[x_{i,1}x_{i,1}^\top] & p_{\min}\mathbb{E}[x_{i,1}x_{i,1}^\top] \\ p_{\min}\mathbb{E}[x_{i,1}x_{i,1}^\top] & p_{\min}^2\mathbb{E}[x_{i,1}x_{i,1}^\top] \end{pmatrix} + \begin{pmatrix} \mathbb{E}[x_{i,1}x_{i,1}^\top] & p_{\max}\mathbb{E}[x_{i,1}x_{i,1}^\top] \\ p_{\max}\mathbb{E}[x_{i,1}x_{i,1}^\top] & p_{\max}^2\mathbb{E}[x_{i,1}x_{i,1}^\top] \end{pmatrix}\right) \\
&= \frac{t}{3}\begin{pmatrix} 2\mathbb{E}[x_{i,1}x_{i,1}^\top] & (p_{\min}+p_{\max})\,\mathbb{E}[x_{i,1}x_{i,1}^\top] \\ (p_{\min}+p_{\max})\,\mathbb{E}[x_{i,1}x_{i,1}^\top] & \left(p_{\min}^2+p_{\max}^2\right)\mathbb{E}[x_{i,1}x_{i,1}^\top] \end{pmatrix} \\
&= \frac{t}{3}\begin{pmatrix} 2 & (p_{\min}+p_{\max}) \\ (p_{\min}+p_{\max}) & \left(p_{\min}^2+p_{\max}^2\right) \end{pmatrix} \otimes \mathbb{E}[x_{i,1}x_{i,1}^\top].
\end{aligned}$$

We can compute the minimum eigenvalue of $\begin{pmatrix} 2 & (p_{\min}+p_{\max}) \\ (p_{\min}+p_{\max}) & \left(p_{\min}^2+p_{\max}^2\right) \end{pmatrix}$ to be

$$\frac{p_{\max}^2+p_{\min}^2+2}{2} - \sqrt{\left(\frac{p_{\max}^2+p_{\min}^2+2}{2}\right)^2 - (p_{\max}-p_{\min})^2}.$$

Note that the eigenvalues of a symmetric positive semi-definite matrix coincide with its singular values. Thus, we can apply Lemma 74 to obtain that the minimum eigenvalue of

$\mathbb{E}[V_{i,t}]$ is at least

$$\lambda_{\min}(\mathbb{E}[V_{i,t}]) \geq \frac{t}{3} \cdot \lambda_{\min}\begin{pmatrix} 2 & (p_{\min}+p_{\max}) \\ (p_{\min}+p_{\max}) & (p_{\min}^2+p_{\max}^2) \end{pmatrix} \cdot \lambda_{\min}\left(\mathbb{E}[x_{i,1}x_{i,1}^\top]\right)$$
$$\geq \frac{t\lambda_0}{3}\left[\frac{p_{\max}^2+p_{\min}^2+2}{2} - \sqrt{\left(\frac{p_{\max}^2+p_{\min}^2+2}{2}\right)^2 - (p_{\max}-p_{\min})^2}\right],$$

where we have used Assumption 9. □

Proof. (Proof of Lemma 57) Taking $\zeta = 1/2$ in Lemma 58 and substituting the result from Lemma 59, we have

$$\Pr\left(\lambda_{\min}(V_{i,t}) \geq \frac{c_0 t}{2}\right) \geq 1 - 2d\left(\frac{e}{2}\right)^{-\frac{c_1 t}{2}}.$$

Setting $t = \mathscr{T}_e = \max\left\{6\log_{e/2}(dNT)/c_1, \max 2\lambda_e/c_0\right\}$, this implies

$$\Pr\left(\lambda_{\min}(V_{i,\mathscr{T}_e}) \geq \lambda_e\right) \geq 1 - \frac{2}{N^3T^2},$$

and we can conclude the proof. □

C.2 Convergence of Prior Mean Estimate

Lemma 29 shows that, after observing i epochs of length T, our estimate $\hat{\theta}_i$ of the unknown prior mean θ_* is close with high probability. To prove Lemma 29, we first focus on the case where the event $\mathscr{A}$ defined in Eq. (C.2) holds. We will show that at the end of each epoch, our estimated parameter vector $\dot{\theta}_i$ is close to the true parameter vector θ_i (Lemma 60) with high probability, which implies that the average of our estimated parameters from each epoch $\frac{1}{i}\sum_{j=1}^{i}\dot{\theta}_j$ is close to the average of the true parameters from each epoch $\frac{1}{i}\sum_{j=1}^{i}\theta_j$ (Lemma 62) with high probability. Next, we will show that the latter term $\frac{1}{i}\sum_{j=1}^{i}\theta_j$ is a good approximation of θ_* (Lemma 63). Combining these steps via a triangle inequality and accounting for the probability $\mathscr{A}$ does not hold yields the result in Lemma 29.

We first state two useful lemmas from the literature regarding the concentration of OLS estimates and the matrix Hoeffding bound.

Lemma 60. *When the event $\mathscr{A}$ holds, for any epoch $i \in [N]$ and $\delta \in [0, 2/e]$, conditional on $F_i = \sigma(\dot{\theta}_1, \ldots, \dot{\theta}_{i-1})$, we have*

$$\Pr\left(\left\| \dot{\theta}_i - \theta_i \right\| \geq 2\sigma \sqrt{\frac{2d \log_e(2/\delta)}{\lambda_e}} \;\middle|\; F_i \right) \leq \delta,$$

Proof. (Proof of Lemma 60) When $\mathscr{A}$ holds, the random exploration periods are completed before T time steps, guaranteeing that $\lambda_{\min}(V_{i,T}) \geq \lambda_e$. Thus, this result follows immediately from Theorem 4.1 of [192], where we note that $d + \log_e(2/\delta) \leq 2d \log_e(2/\delta)$ for $\delta < 2/e$. □

Lemma 61 ([111])**.** *Let random vectors $X_1, \ldots, X_n \in \mathscr{R}^d$, satisfy that for all $i \in [n]$ and $u \in \mathscr{R}$,*

$$\mathbb{E}[X_i | \sigma(X_1, \ldots, X_{i-1})] = 0, \quad \Pr(\|X_i\| \geq u | \sigma(X_1, \ldots, X_{i-1})) \leq 2\exp\left(-\frac{u^2}{2\sigma_i^2}\right),$$

then for any $\delta > 0$,

$$\Pr\left(\left\| \sum_{i \in [n]} X_i \right\| \leq 4 \sqrt{\sum_{i \in [n]} \sigma_i^2 \log_e(2d/\delta)} \right) \geq 1 - \delta\,.$$

We now show that the average of our estimated parameters from each epoch is close to the average of the true parameters from each epoch with high probability.

Lemma 62. *When the event $\mathscr{A}$ holds, for any $i \geq 2$, the following holds with probability at least $1 - \delta$:*

$$\left\| \frac{1}{i} \sum_{j=1}^{i} \left(\dot{\theta}_j - \theta_j \right) \right\| \leq 8\sigma \sqrt{\frac{d \log_e(4d/\delta)}{\lambda_e i}}\,.$$

Proof. (Proof of Lemma 62) By Lemma 60, we have for any $u \in \mathscr{R}$,

$$\Pr(\|\dot{\theta}_i - \theta_i\| \geq u \mid F_i) \;\leq\; 2\exp(-\lambda_e u^2 / 8d\sigma^2)\,.$$

Furthermore, since the OLS estimator is unbiased, $\mathbb{E}[\dot{\theta}_i|F_i] = \theta_i$. Thus, we can apply the matrix Hoeffding inequality (Lemma 61) to obtain

$$\Pr\left(\left\|\frac{1}{i-1}\sum_{j=1}^{i-1}(\dot{\theta}_i - \theta_i)\right\| \;\leq\; 8\sqrt{\frac{\sigma^2 d\log_e(4d/\delta)}{\lambda_e(i-1)}}\right) \;\geq\; 1-\delta.$$

Noting that $i \leq 2(i-1)$ for all $i \in \{2,\ldots,N\}$ concludes the proof. □

Lemma 63. *When the event $\mathscr{A}$ holds, for any $i \geq 2$, the following holds with probability at least $1-\delta$:*

$$\left\|\frac{1}{i}\sum_{j=1}^{i}\theta_j - \theta_*\right\| \;\leq\; 8\sqrt{\frac{5\overline{\lambda}d\log_e(4d/\delta)}{i}}.$$

Proof. (Proof of Lemma 63) We first show a concentration inequality for the quantity $\|\theta_j - \theta_*\|$ similar to that of Lemma 60. Note that for any unit vector $s \in \mathscr{R}^{2d}$, $u^\top(\theta_i - \theta_*)$ is a zero-mean normal random variable with variance at most $\overline{\lambda}$. Therefore, for any $u \in \mathscr{R}$,

$$\Pr\left(|s^\top(\theta_j - \theta_*)| \geq u\right) \leq 2\exp\left(-\frac{u^2}{2\overline{\lambda}}\right). \tag{C.8}$$

Consider W, a $(1/2)$-cover of the unit ball in $\mathscr{R}^{2d}$. We know that $|W| \leq 4^{2d}$. Let $s(\theta_j) = \theta_j - \theta_*/\|\theta_j - \theta_*\|$, then there exists $w_{s(\theta_j)} \in W$, such that $\|w_{s(\theta_j)} - s(\theta_j)\| \leq 1/2$ by definition of W. Hence,

$$\begin{aligned}\|\theta_j - \theta_*\| =& \langle s(\theta_j), \theta_j - \theta_*\rangle = \langle s(\theta_j) - w_{s(\theta_j)}, \theta_j - \theta_*\rangle + \langle w_{s(\theta_j)}, \theta_j - \theta_*\rangle \\ \leq& \frac{\|\theta_j - \theta_*\|}{2} + \langle w_{s(\theta_j)}, \theta_j - \theta_*\rangle.\end{aligned}$$

Rearranging the terms yields

$$\|\theta_j - \theta_*\| \leq 2\langle w_{s(\theta_j)}, \theta_j - \theta_*\rangle.$$

Applying an union bound to all possible $w \in W$ with inequality (C.8), we have for any

$u \in \mathscr{R}$,

$$\begin{aligned}\Pr(\|\theta_j - \theta_*\| \geq u) &\leq \Pr(\exists w \in W : \langle w, \theta_j - \theta_* \rangle \geq u/2) \\ &\leq 2 \cdot 4^{2d} \exp\left(-\frac{u^2}{2\overline{\lambda}}\right) \\ &\leq \exp\left(5d - \frac{u^2}{2\overline{\lambda}}\right).\end{aligned}$$

If $u^2 \leq 10\overline{\lambda}d$, we have

$$\Pr(\|\theta_j - \theta_*\| \geq u) \;\leq\; 1 \;\leq\; 2\exp\left(-\frac{u^2}{20\overline{\lambda}d}\right);$$

else if $u^2 = 10\overline{\lambda}d + v$ for some $v \geq 0$, we have

$$\begin{aligned}\Pr(\|\theta_j - \theta_*\| \geq u) &\leq \exp\left(-\frac{v}{2\overline{\lambda}}\right) \\ &\leq 2\exp\left(-\frac{u^2}{20\overline{\lambda}d}\right).\end{aligned}$$

Thus, for any $u \in \mathscr{R}$, we can write

$$\Pr(\|\theta_j - \theta_*\| \geq u) \leq 2\exp\left(-\frac{u^2}{20\overline{\lambda}d}\right). \tag{C.9}$$

Applying Lemma 61, we have

$$\Pr\left(\left\|\frac{\sum_{j=1}^{i-1}\theta_j}{i-1} - \theta_*\right\| \leq 4\sqrt{\frac{10\overline{\lambda}d\log_e(4d/\delta)}{i-1}}\right) \geq 1-\delta.$$

The proof can be concluded by the observation $i \leq 2(i-1)$ for all $i \in\in \{2,\ldots,N\}$. □

We can now combine Lemmas 57, 62 and 63 to prove Lemma 29.

Proof. (Proof of Lemma 29) When the event $\mathscr{A}$ holds, we can use the triangle inequality

and a union bound over Lemmas 62 and 63 to obtain

$$\begin{aligned}\left\|\hat{\theta}_i - \theta_*\right\| &= \left\| \frac{\sum_{j=1}^{i-1} \dot{\theta}_j}{i-1} - \frac{\sum_{j=1}^{i-1} \theta_j}{i-1} + \frac{\sum_{j=1}^{i-1} \theta_j}{i-1} - \theta_* \right\| \\ &\leq \left\| \frac{1}{i-1} \sum_{j=1}^{i-1} \left(\dot{\theta}_j - \theta_j\right) \right\| + \left\| \frac{1}{i-1} \sum_{j=1}^{i-1} \theta_j - \theta_* \right\| \\ &\leq 8\sqrt{\frac{2(\sigma^2/\lambda_e + 5\overline{\lambda}) d \log_e(4dN/\delta)}{i}},\end{aligned}$$

with probability at least $1-2\delta$, where we have use the fact that $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$. By Lemma 57, the event $\mathscr{A}$ does not hold with probability at most $2/(N^2T^2)$. Thus, a second union bound yields the result. □

C.3 `Meta-DP` Regret Analysis

Appendix C.3.1 provides the proof of Lemma 30 and the statement of an intermediate Lemma 64. Appendix C.3.2 provides the proof of Theorem 28, following the proof strategy outlined in Section 4.2.3.

C.3.1 Intermediate Lemmas

Recall that for any $t \in \{\mathscr{T}_i + 1, \cdots, T\}$, the meta oracle maintains and samples from its posterior $\mathscr{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}^{\mathrm{TS}}\right)$ (see Algorithm 5), while our `Meta-DP` algorithm maintains and samples parameters from its posterior $\mathscr{N}\left(\theta_{i,t}^{\mathrm{MD}}, \Sigma_{i,t}^{\mathrm{MD}}\right)$ (see Algorithm 6). Lemma 30 in Section 4.2.3 established the difference in Bayesian posteriors between the meta oracle and our `Meta-DP` algorithm. The proof follows from the standard update rules for Bayesian linear regression and is given below.

Proof. (Proof of Lemma 30)

Using the posterior update rule for Bayesian linear regression [43], the posterior of the

oracle at $t = \mathscr{T}_i + 1$ is

$$\begin{aligned}
\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1} &= \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top\right)^{-1} \left(\Sigma_*^{-1}\theta_* + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} D_{i,t}\right) \\
&= \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top\right)^{-1} \left(\Sigma_*^{-1}\theta_* + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top \theta_i + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} \varepsilon^{\mathrm{TS}}_{i,t}\right), \\
\Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1} &= \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top\right)^{-1}.
\end{aligned}$$

Similarly, the posterior of the `Meta-DP` algorithm at $t = \mathscr{T}_i + 1$ is

$$\begin{aligned}
\theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1} &= \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{t_e} m_{i,t} m_{i,t}^\top\right)^{-1} \left(\Sigma_*^{-1}\hat{\theta}_i + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} D_{i,t}\right) \\
&= \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top\right)^{-1} \left(\Sigma_*^{-1}\hat{\theta}_i + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top \theta_i + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} \varepsilon^{\mathrm{MD}}_{i,t}\right), \\
\Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1} &= \left(\Sigma_*^{-1} + \sigma \sum_{t=1}^{\mathscr{T}_i} m_{i,t} m_{i,t}^\top\right)^{-1}.
\end{aligned}$$

The result follows directly. □

We also note that the prior-independent Thompson sampling algorithm employed in the exploration epochs satisfies a meta regret guarantee:

Lemma 64. *The meta regret of the prior-independent Thompson sampling algorithm in a single epoch is* $\widetilde{O}(d^2 T^{1/2})$.

The proof can be easily adapted from the literature (see, *e.g.*, [10, 5]), and is thus omitted. We note that our normalization implies $\mathbb{E}[\|\theta\|] = \Theta(d^{1/2})$. Lemma 64 ensures that we accrue at most $\widetilde{O}(d^2 N_0 \sqrt{T})$ regret in the N_0 exploration epochs; from Eq. (4.5), we know that N_0 grows merely poly-logarithmically in N and T.

C.3.2 Proof of Theorem 28

Consider any non-exploration epoch $i \geq N_0 + 1$. If upon completion of all exploration steps at time $\mathscr{T}_i + 1$, we have that the posteriors of the meta oracle and our `Meta-DP` algorithm co-

incide — *i.e.*, $(\theta^{\text{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MD}}_{i,\mathscr{T}_i+1}) = (\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1})$ — then both policies would achieve the same expected revenue over the time periods $\mathscr{T}_i+1, \cdots, T$, *i.e.*, we would have

$$\texttt{REV}\left(\theta_i, \theta^{\text{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) = \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right).$$

By Lemma 30, we know that $\Sigma^{\text{TS}}_{i,\mathscr{T}_i+1} = \Sigma^{\text{MD}}_{i,\mathscr{T}_i+1}$ always, so all that remains is establishing when $\theta^{\text{TS}}_{i,\mathscr{T}_i+1} = \theta^{\text{MD}}_{i,\mathscr{T}_i+1}$.

Since the two algorithms begin with different priors but encounter the same covariates $\{x_{i,t}\}_{t=1}^T$ and take the same decisions in $t \in \{1, \cdots, \mathscr{T}_i\}$, their posteriors can only align at time $\mathscr{T}_i+1$ due to the stochasticity in the observations $\varepsilon_{i,t}$. As shown in Eq. (4.10) in Section 4.2.3, alignment occurs with $\theta^{\text{TS}}_{i,\mathscr{T}_i+1} = \theta^{\text{MD}}_{i,\mathscr{T}_i+1}$ if

$$\chi^{\text{MD}}_i - \chi^{\text{TS}}_i = \frac{1}{\sigma}(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right),$$

where we recall $\chi^{\text{TS}}, \chi^{\text{MD}}_i$ were defined in Eqs. (4.8)-(4.9).

Now, we start by defining the clean event

$$\mathscr{E} = \left\{ \left\|\hat{\theta}_i - \theta_*\right\| \le 8\sqrt{\frac{2(\sigma^2/\lambda_e + 5\overline{\lambda})d\log_e(4dN^2T)}{i}}, \quad \mathscr{T}_i \le \mathscr{T}_e \qquad \forall i \ge N_0+1 \right\}, \tag{C.10}$$

which stipulates that for every epoch i after the initial N_0 exploration epochs, (i) our estimated prior mean $\hat{\theta}_i$ is close to the unknown prior mean θ_* (which holds with high probability by Lemma 29), (ii) and the event $\mathscr{A}$ defined in Eq. (C.2) holds, ensuring that the number of exploration periods per epoch is small (which holds with high probability by Lemma 57). Since $\mathscr{E}$ holds with high probability, we first focus on analyzing the meta regret conditioned on $\mathscr{E}$.

Denote the meta regret of epoch i conditioned on the event $\mathscr{E}$ defined in Eq. (C.10) as $\mathscr{R}_{N,T}(i) \mid \mathscr{E}$. The next lemma bounds the meta regret for any epoch $i \ge N_0$ under the event $\mathscr{E}$.

Lemma 65. *The meta regret of an epoch $i \geq N_0 + 1$ satisfies*

$$\mathscr{R}_{N,T}(i) \mid \mathscr{E} \;=\; \tilde{O}\left(d^2\sqrt{\frac{T}{i}} + \frac{\sqrt{d}}{N}\right).$$

Proof. Proof of Lemma 65 As noted earlier, during the exploration periods $1 \leq t \leq \mathscr{T}_i$, the meta oracle and our `Meta-DP` algorithm encounter the same covariates $\{x_{i,t}\}_{t=1}^{T}$ and offer the same prices; thus, by construction, they achieve the same expected revenue and the resulting meta regret is 0. Then, we can write

$$\begin{aligned}
&\mathscr{R}_{N,T}(i) \mid \mathscr{E} \\
&= \mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}},\chi_i^{\mathrm{MD}}}\left[\mathtt{REV}\left(\theta_i, \theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] \\
&= \mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{MD}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] \\
&\quad - \mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right]. \qquad \text{(C.11)}
\end{aligned}$$

We will use our prior alignment technique to express the first term in Eq. (C.11) in terms of the second term in Eq. (C.11); in other words, we will use a change of measure suggested by Eq. (4.10) to express the true regret of our `Meta-DP` algorithm as a function of the true regret of the meta oracle.

We start by expanding the first term of Eq. (C.11) as

$$\begin{aligned}
&\mathbb{E}_{\chi_i^{\mathrm{MD}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] \\
&= \int_{\chi_i^{\mathrm{MD}}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{MD}}\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{t_e/2}} \left(\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right)\right) d\chi_i^{\mathrm{MD}} \mid \mathscr{E}.
\end{aligned}$$

Given a realization of χ_i^{MD}, we denote $\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})$ (with some abuse of notation) as the corresponding realization of χ_i^{TS} that satisfies Eq. (4.10). Note that this is a unique one-to-one mapping. We then perform a change of measure to continue:

$$\int_{\chi_i^{\mathrm{MD}}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{MD}}\right\|^2/2\sigma^2\right)}{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}}$$

$$
\begin{aligned}
&\times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \\
&= \int_{\left\|\chi_i^{\mathrm{MD}}\right\| \le 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{MD}}\right\|^2/2\sigma^2\right)}{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\qquad \times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \\
&\quad + \int_{\left\|\chi_i^{\mathrm{MD}}\right\| \ge 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{MD}}\right\|^2/2\sigma^2\right)}{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\qquad \times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \\
&\le \max_{\left\|\chi_i^{\mathrm{MD}}\right\| \le 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2 - \left\|\chi_i^{\mathrm{MD}}\right\|^2}{2\sigma^2}\right) \\
&\quad \times \int_{\left\|\chi_i^{\mathrm{MD}}\right\| \le 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\qquad \times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \\
&\quad + \int_{\left\|\chi_i^{\mathrm{MD}}\right\| \ge 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{MD}}\right\|^2/2\sigma^2\right)}{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\qquad \times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \\
&\le \max_{\left\|\chi_i^{\mathrm{MD}}\right\| \le 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2 - \left\|\chi_i^{\mathrm{MD}}\right\|^2}{2\sigma^2}\right) \int_{\chi_i^{\mathrm{MD}}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\qquad \times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \\
&\quad + \int_{\left\|\chi_i^{\mathrm{MD}}\right\| \ge 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \frac{\exp\left(-\left\|\chi_i^{\mathrm{MD}}\right\|^2/2\sigma^2\right)}{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)} \frac{\exp\left(-\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\qquad \times \left(\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\mathrm{MD}}\right) \mid \mathscr{E} \qquad \text{(C.12)} \\
&\le \max_{\left\|\chi_i^{\mathrm{MD}}\right\| \le 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2 - \left\|\chi_i^{\mathrm{MD}}\right\|^2}{2\sigma^2}\right) \\
&\quad \times \mathbb{E}_{\chi_i^{\mathrm{TS}}}\left[\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] \\
&\quad + \mathbb{E}_{\chi_i^{\mathrm{MD}}}\left[\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}, \left\|\chi_i^{\mathrm{MD}}\right\| \ge 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right] \\
&\qquad \times \Pr\left(\left\|\chi_i^{\mathrm{MD}}\right\| \ge 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right). \qquad \text{(C.13)}
\end{aligned}
$$

Here, inequality (C.12) follows from the fact that $\texttt{REV}_*(\theta_i, T-T_i) \geq \texttt{REV}(\theta_i, \theta, \Sigma, T-T_i)$ for any choice of θ and Σ. Thus, we have expressed the true regret of our `Meta-DP` algorithm as the sum of a term that is proportional to the true regret of the meta oracle, and an additional term that depends on the tail probability of χ_i^{MD}. To obtain our desired bound, we will argue that (i) the coefficient of the first term decays to 1 as the epoch number i grows large, ensuring that our meta regret goes to 0 for later epochs, and (ii) the second term is negligible with high probability since χ_i^{MD} is a subgaussian random variable.

We start by characterizing the coefficient of the first term in Eq. (C.13):

$$
\begin{aligned}
&\max_{\|\chi_i^{\text{MD}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\text{TS}}(\chi_i^{\text{MD}})\right\|^2 - \left\|\chi_i^{\text{MD}}\right\|^2}{2\sigma^2}\right) \\
&= \max_{\|\chi_i^{\text{MD}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|n_i^{\text{MD}} - \frac{1}{\sigma}(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|^2 - \left\|\chi_i^{\text{MD}}\right\|^2}{2\sigma^2}\right) \\
&= \max_{\|\chi_i^{\text{MD}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left(\chi_i^{\text{MD}}\right)^\top (M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)}{\sigma^3}\right) \\
&\quad \times \exp\left(\frac{\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|^2}{2\sigma^4}\right) \\
&\leq \max_{\|\chi_i^{\text{MD}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\text{MD}}\right\| \left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|}{\sigma^3}\right) \\
&\quad \times \exp\left(\frac{\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|^2}{2\sigma^4}\right) \\
&= \exp\left(\frac{4\sqrt{\mathscr{T}_i \log_e(2NT)}\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|}{\sigma^2} + \frac{\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|^2}{2\sigma^4}\right).
\end{aligned}
\tag{C.14}
$$

Note that

$$
\begin{aligned}
4\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\| &\leq \lambda_{\max}\left((M_i^\top M_i)^{-1}\right)\sqrt{\lambda_{\max}(M_i M_i^\top)}\lambda_{\max}(\Sigma_*^{-1})\left\|\hat{\theta}_i - \theta_*\right\| \\
&\leq 32\sqrt{\frac{\mathscr{T}_i x_{\max}^2 (1+p_{\max}^2)(\sigma^2\lambda_e^{-1} + 5\overline{\lambda})d\log_e(4dN^2T)}{\lambda_e^2 \underline{\lambda}^2 i}} \\
&\leq c_2\sigma^2\sqrt{\frac{d\,\mathscr{T}_i \log_e(4dN^2T)}{i}}.
\end{aligned}
\tag{C.15}
$$

Furthermore, by the definition of N_0 in Eq. (4.5), we have for all $i \geq N_0 + 1$,

$$\frac{4\sqrt{\mathscr{T}_i \log_e(2NT)}\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|}{\sigma^2} + \frac{\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|^2}{2\sigma^4} \leq 1\,. \tag{C.16}$$

Combining Eqs. (C.14) and (C.16), and applying Lemma 76 in Appendix C.7 yields

$$\begin{aligned}
&\max_{\left\|\chi_i^{\mathrm{MD}}\right\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MD}})\right\|^2 - \left\|\chi_i^{\mathrm{MD}}\right\|^2}{2\sigma^2}\right) \\
&\leq 1 + \frac{8\sqrt{\mathscr{T}_i \log_e(2NT)}\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|}{\sigma^2} + \frac{\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|^2}{\sigma^4} \\
&\leq 1 + \frac{16\sqrt{\mathscr{T}_i \log_e(2NT)}\left\|(M_i^\top M_i)^{-1} M_i^\top \Sigma_*^{-1}\left(\theta_* - \hat{\theta}_i\right)\right\|}{\sigma^2} \\
&\leq 1 + 4c_2 \mathscr{T}_i \sqrt{\frac{d \log_e(4dN^2T)\log_e(2NT)}{i}}\,,
\end{aligned} \tag{C.17}$$

where we have used Eq. (C.15) in the last step. Plugging this into Eq. (C.13), we can now bound

$$\begin{aligned}
&\mathbb{E}_{\chi_i^{\mathrm{MD}}}\left[\mathtt{REV}_*\left(\theta_i, T - \mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta_{i,\mathscr{T}_i+1}^{\mathrm{MD}}, \Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MD}}, T - \mathscr{T}_i\right) \mid \mathscr{E}\right] \\
&\leq \left(1 + 4c_2 \mathscr{T}_i \sqrt{\frac{d \log_e(4dN^2T)\log_e(2NT)}{i}}\right) \\
&\quad \times \mathbb{E}_{\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i, T - \mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}}, \Sigma_{i,\mathscr{T}_i+1}^{\mathrm{TS}}, T - \mathscr{T}_i\right) \mid \mathscr{E}\right] \\
&\quad + \mathbb{E}_{\chi_i^{\mathrm{MD}}}\left[\mathtt{REV}_*\left(\theta_i, T - \mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta_{i,\mathscr{T}_i+1}^{\mathrm{MD}}, \Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MD}}, T - \mathscr{T}_i\right) \mid \mathscr{E}, \left\|\chi_i^{\mathrm{MD}}\right\| \geq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right] \\
&\qquad \times \Pr\left(\left\|\chi_i^{\mathrm{MD}}\right\| \geq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right)\,.
\end{aligned} \tag{C.18}$$

As desired, this establishes that the coefficient of our first term decays to 1 as i grows large. Thus, our meta regret from the first term approaches 0 for large i. We now show that the second term in Eq. (C.18) is negligible with high probability. Similar to the proof of Lemma 63, for any $u \in \mathscr{R}$, we can write $\Pr\left(\left\|\chi_i^{\mathrm{MD}}\right\| \geq u\right) \leq 2\exp\left(-u^2/(10\sigma^2\mathscr{T}_i)\right)$, which

implies

$$\Pr\left(\left\|\chi_i^{\mathrm{MD}}\right\| \geq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right) \leq \frac{1}{NT}. \tag{C.19}$$

Moreover, noting that the worst-case regret achievable in a single time period is

$$2p_{\max}x_{\max}\sqrt{1+p_{\max}^2}\|\theta_i\|,$$

and $\mathscr{T}_i \leq \mathscr{T}_e$ on the event $\mathscr{E}$, we can bound

$$\begin{aligned}
&\mathbb{E}_{\chi_i^{\mathrm{MD}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{MD}}_{i,\mathscr{T}_e+1}, \Sigma^{\mathrm{MD}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}, \left\|\chi_i^{\mathrm{MD}}\right\| \geq 4\sigma\sqrt{\mathscr{T}_i\log_e(2NT)}\right] \\
&\leq 2(T-\mathscr{T}_i)p_{\max}x_{\max}\sqrt{1+p_{\max}^2}\mathbb{E}[\|\theta_i\|] \\
&= O(\sqrt{d}T),
\end{aligned} \tag{C.20}$$

where we recall from Eq. (C.7) that $\mathbb{E}[\|\theta_i\|] = O(\sqrt{d})$. Substituting Eqs. (C.19) and (C.20), into Eq. (C.18), we obtain

$$\begin{aligned}
&\left(1 + 4c_2\mathscr{T}_i\sqrt{\frac{d\log_e(4dN^2T)\log_e(2NT)}{i}}\right) \\
&\times \mathbb{E}_{\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] + O\left(\frac{\sqrt{\sqrt{d}}}{N}\right).
\end{aligned}$$

Substituting the above into Eq. (C.11), we can bound the meta regret of epoch i as

$$\begin{aligned}
&\mathscr{R}_{N,T}(i) \mid \mathscr{E} \\
&\leq \left(4c_2\mathscr{T}_i\sqrt{\frac{d\log_e(4dN^2T)\log_e(2NT)}{i}}\right) \\
&\times \mathbb{E}_{\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] + O\left(\frac{\sqrt{d}}{N}\right) \\
&= \tilde{O}\left(d^2\sqrt{\frac{T}{i}} + \frac{\sqrt{d}}{N}\right).
\end{aligned}$$

Here, we have used the fact that the meta oracle's true regret is bounded (Theorem 27), *i.e.*,

$$\mathbb{E}_{\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right)-\mathtt{REV}\left(\theta_i, \theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}}, \Sigma_{i,\mathscr{T}_i+1}^{\mathrm{TS}}, T-\mathscr{T}_i\right) \mid \mathscr{E}\right] \leq \tilde{O}(d^{3/2}\sqrt{T}).$$

□

The remaining proof of Theorem 28 follows straightforwardly.

Proof. (Proof of Theorem 28) The meta regret can then be decomposed as follows:

$$\begin{aligned}\mathscr{R}_{N,T} &= (\mathscr{R}_{N,T} \mid \mathscr{E})\Pr(\mathscr{E}) + (\mathscr{R}_{N,T} \mid \neg\mathscr{E})\Pr(\neg\mathscr{E}) \\ &\leq (\mathscr{R}_{N,T} \mid \mathscr{E}) + (\mathscr{R}_{N,T} \mid \neg\mathscr{E})\Pr(\neg\mathscr{E}).\end{aligned}$$

Recall that the event $\mathscr{E}$ is composed of two events: $\mathscr{A}$ (bounded by Lemma 57) and a bound on $\|\hat{\theta}_i - \theta_*\|$ (bounded by Lemma 29). Applying a union bound over the epochs $i \geq N_0+1$ to Lemma 29 (setting $\delta = 1/(N^2T)$), and applying Lemma 57 yields a bound

$$\Pr(\mathscr{E}) \geq 1 - 1/(NT) - 4/(NT^2) \geq 1 - 5/(NT).$$

Recall that when the event $\mathscr{E}$ is violated, the meta regret is $O(NT)$, so we can bound $(\mathscr{R}_{N,T} \mid \neg\mathscr{E})\Pr(\neg\mathscr{E}) \leq O(NT \times 1/(NT)) = O(1)$. Therefore, the overall meta regret is simply

$$\mathscr{R}_{N,T} \leq (\mathscr{R}_{N,T} \mid \mathscr{E}) + O(1).$$

When $N > N_0$, applying our result in Lemma 65 yields

$$\begin{aligned}&\sum_{i=1}^{N_0}(\mathscr{R}_{N,T}(i) \mid \mathscr{E}) + \sum_{i=N_0+1}^{N}(\mathscr{R}_{N,T}(i) \mid \mathscr{E}) + O(1) \\ &\leq N_0\tilde{O}(d^2\sqrt{T}) + \sum_{i=N_0+1}^{N}\tilde{O}\left(d^2\sqrt{\frac{T}{i}} + \frac{\sqrt{d}}{N}\right) + O(1) \\ &\leq \sum_{i=1}^{N}\tilde{O}\left(d^2\sqrt{\frac{T}{i}} + \frac{\sqrt{d}}{N}\right) + \tilde{O}(d^3\sqrt{T})\end{aligned}$$

$$= \tilde{O}\left(d^2(NT)^{\frac{1}{2}} + d^3\sqrt{T}\right),$$

where we have use the fact that $\sum_{i=1}^{N} 1/\sqrt{i} \leq 2\sqrt{N}$ in the last step. □

C.4 Convergence of Prior Covariance Estimate

Lemma 32 shows that, after observing i epochs of length T, our estimator $\hat{\Sigma}_i$ is close to Σ_* with high probability. To prove Lemma 32, we first focus on the case where the event $\mathscr{A}$ defined in Eq. (C.2) holds. For ease of notation, denote the average of the estimated parameters from each epoch as

$$\bar{\theta}_i = \frac{1}{i-1}\sum_{k=1}^{i-1}\dot{\theta}_k.$$

Then, recall from the definition in Eq. (4.12) that

$$\hat{\Sigma}_i = \frac{1}{i-2}\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \bar{\theta}_i\right)\left(\dot{\theta}_j - \bar{\theta}_i\right)^\top - \frac{\sigma^2}{i-1}\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right].$$

Then, we can expand

$$\begin{aligned}
&\left\|\hat{\Sigma}_i - \Sigma_*\right\|_{op} \\
&= \left\| \frac{1}{i-2}\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \bar{\theta}_i\right)\left(\dot{\theta}_j - \bar{\theta}_i\right)^\top - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1} - \Sigma_* \right\|_{op} \\
&= \left\| \frac{1}{i-2}\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \theta_*\right)\left(\dot{\theta}_j - \theta_*\right)^\top - \frac{i-1}{i-2}\left(\theta_* - \bar{\theta}_i\right)\left(\theta_* - \bar{\theta}_i\right)^\top - \frac{\sigma^2\sum_{j=1}^{i-1}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1} - \Sigma_* \right\|_{op} \\
&= \left\| \frac{1}{i-2}\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \theta_*\right)\left(\dot{\theta}_j - \theta_*\right)^\top - \frac{i-1}{i-2}\Sigma_* - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-2} \right. \\
&\quad \left. - \frac{i-1}{i-2}\left(\theta_* - \bar{\theta}_i\right)\left(\theta_* - \bar{\theta}_i\right)^\top + \frac{1}{i-2}\Sigma_* + \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{(i-1)(i-2)} \right\|_{op} \\
&\leq \frac{i-1}{i-2}\left\| \frac{1}{i-1}\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \theta_*\right)\left(\dot{\theta}_j - \theta_*\right)^\top - \Sigma_* - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1} \right\|_{op}
\end{aligned}$$

$$+\frac{i-1}{i-2}\left\|\left(\theta_*-\bar{\theta}_i\right)\left(\theta_*-\bar{\theta}_i\right)^\top-\frac{1}{i-1}\Sigma_*-\frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{(i-1)^2}\right\|_{op}. \tag{C.21}$$

We proceed by showing that each of the two terms is a subgaussian random variable, and therefore satisfies standard concentration results. The following lemma first establishes that both terms have expectation zero, *i.e.*, $\hat{\Sigma}_i$ is an unbiased estimator of the true prior covariance matrix Σ_*.

Lemma 66. *When the event $\mathscr{A}$ holds, for any epoch $i \geq 3$,*

$$\mathbb{E}\left[\frac{1}{i-1}\sum_{j=1}^{i-1}\left(\dot{\theta}_j-\theta_*\right)\left(\dot{\theta}_j-\theta_*\right)^\top\right]=\Sigma_*+\frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1},$$

$$\mathbb{E}\left[\left(\theta_*-\bar{\theta}_i\right)\left(\theta_*-\bar{\theta}_i\right)^\top\right]=\frac{1}{i-1}\Sigma_*+\frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{(i-1)^2}.$$

Proof. (Proof of Lemma 66) When $\mathscr{A}$ holds, the random exploration time steps are completed before T time steps. Denote

$$\Delta_j=V_{j,\mathscr{T}_j}^{-1}\left(\sum_{t=1}^{\mathscr{T}_j}\varepsilon_{j,t}m_{j,t}\right)=\dot{\theta}_j-\theta_j. \tag{C.22}$$

Then noting that $\mathbb{E}[\theta_j]=\theta_*$, $\mathbb{E}[\Delta_j]=0$, and $\mathbb{E}[\Delta_j\Delta_j^\top]=\sigma^2\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]$, we can write

$$\begin{aligned}\mathbb{E}\left[(\dot{\theta}_j-\theta_*)(\dot{\theta}_j-\theta_*)^\top\right]&=\mathbb{E}\left[(\theta_j+\Delta_j)(\theta_j+\Delta_j)^\top-\theta_*\theta_*^\top\right]\\&=\mathbb{E}\left[\theta_j\theta_j^\top-\theta_*\theta_*^\top\right]+\mathbb{E}\left[\Delta_j\Delta_j^\top\right]\\&=\Sigma_*+\sigma^2\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right].\end{aligned}$$

Summing over j and dividing by $(i-1)$ on both sides yields the first statement. For the second statement, we can write

$$\mathbb{E}\left[(\bar{\theta}_i-\theta_*)(\bar{\theta}_i-\theta_*)^\top\right]=\mathbb{E}\left[\bar{\theta}_i\bar{\theta}_i^\top-\theta_*\theta_*^\top\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\frac{\sum_{k=1}^{i-1}\dot{\theta}_k}{i-1}\right)\left(\frac{\sum_{k=1}^{i-1}\dot{\theta}_k}{i-1}\right)^\top - \theta_*\theta_*^\top\right] \\
&= \mathbb{E}\left[\frac{\sum_{k=1}^{i-1}\theta_k\theta_k^\top + \sum_{k=1}^{i-1}\Delta_k\Delta_k^\top + \sum_{1\le j_1<j_2\le i-1}\theta_{j_1}\theta_{j_2}^\top}{(i-1)^2} - \theta_*\theta_*^\top\right] \\
&= \mathbb{E}\left[\frac{\sum_{k=1}^{i-1}\theta_k\theta_k^\top + \sum_{k=1}^{i-1}\Delta_k\Delta_k^\top}{(i-1)^2} - \frac{1}{i-1}\theta_*\theta_*^\top\right] \\
&= \frac{1}{i-1}\Sigma_* + \frac{\sigma^2\sum_{k=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{(i-1)^2}.
\end{aligned}
$$

□

Having established that both terms in Eq. (C.21) have expectation zero, the following lemma shows that these terms are subgaussian and therefore concentrate with high probability.

Lemma 67. *When the event $\mathscr{A}$ holds, for any $\delta \in [0,1]$, the following holds with probability at least $1-2\delta$:*

$$
\begin{aligned}
&\left\|\frac{\sum_{j=1}^{i-1}\left(\dot{\theta}_j-\theta_*\right)\left(\dot{\theta}_j-\theta_*\right)^\top}{i-1} - \Sigma_* - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1}\right\|_{op} \\
&\le \frac{16(\overline{\lambda}\lambda_e^2+16\sigma^2 d)}{\lambda_e^2}\left(\sqrt{\frac{5d+2\log_e(2/\delta)}{i-1}} \vee \frac{5d+2\log_e(2/\delta)}{i-1}\right), \\
&\left\|\left(\theta_*-\bar{\theta}_i\right)\left(\theta_*-\bar{\theta}_i\right)^\top - \frac{1}{i-1}\Sigma_* - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{(i-1)^2}\right\|_{op} \\
&\le \frac{16(\overline{\lambda}\lambda_e^2+16\sigma^2 d)(5d+2\log_e(2/\delta))}{\lambda_e^2(i-1)}.
\end{aligned}
$$

Proof. Proof of Lemma 67 First, since the OLS estimator is unbiased, we have that

$$
\mathbb{E}\left[\dot{\theta}_j-\theta_*\right] = 0
$$

for all j, and consequently, $\mathbb{E}\left[\bar{\theta}_i-\theta_*\right] = 0$. Recall also our definition of Δ_j from Eq. (C.22).

Then, for any $v \in \mathscr{R}^{2d}$ such that $\|v\| = 1$, we can write for all $u \in \mathscr{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp(u\langle v, \dot{\theta}_j - \theta_*\rangle)\right] =& \mathbb{E}\left[\exp(u\langle v, \theta_j - \theta_*\rangle)\exp(u\langle v, \Delta_j\rangle)\right] \\
&= \mathbb{E}\left[\exp(u\langle v, \theta_j - \theta_*\rangle)\right]\mathbb{E}\left[\exp(u\langle v, \Delta_j\rangle)\right] \\
&= \exp\left(\frac{u^2 v^\top \Sigma_* v}{2}\right)\mathbb{E}\left[\exp(u\langle v, \Delta_j\rangle)\right] \\
&\leq \exp\left(u^2\left(\frac{\overline{\lambda}}{2} + \frac{8\sigma^2 d}{\lambda_e^2}\right)\right),
\end{aligned}$$

where we have re-used Lemmas 60 (from Appendix C.2) and 77 (from Appendix C.7) in the last step. Similarly,

$$\mathbb{E}\left[\exp(u\langle v, \bar{\theta} - \theta_*\rangle)\right] \leq \exp\left(\frac{u^2}{i-1}\left(\frac{\overline{\lambda}}{2} + \frac{8\sigma^2 d}{\lambda_e^2}\right)\right).$$

By definition, along with Lemma 66, this implies that $\dot{\theta}_j - \theta_*$ is a $\left(\sqrt{(\overline{\lambda}\lambda_e^2 + 16\sigma^2 d)/2\lambda_e^2}\right)$-subgaussian vector and, similarly $\bar{\theta} - \theta_*$ is a $\left(\sqrt{(\overline{\lambda}\lambda_e^2 + 16\sigma^2 d)/[\lambda_e^2(i-1)]}\right)$-subgaussian vector. Applying concentration results for subgaussian random variables (see Lemma 78 from Appendix C.7), we have with probability at least $1-\delta$,

$$\begin{aligned}
&\left\| \frac{\sum_{j=1}^{i-1}\left(\dot{\theta}_j - \theta_*\right)\left(\dot{\theta}_j - \theta_*\right)^\top}{i-1} - \Sigma_* - \frac{\sigma^2 \sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{i-1} \right\|_{op} \\
\leq& \frac{16(\overline{\lambda}\lambda_e^2 + 16\sigma^2 d)}{\lambda_e^2}\left(\sqrt{\frac{5d + 2\log_e(2/\delta)}{i-1}} \vee \frac{5d + 2\log_e(2/\delta)}{i-1}\right).
\end{aligned}$$

Similarly, with probability at least $1-\delta$,

$$\begin{aligned}
&\left\| \left(\theta_* - \bar{\theta}_i\right)\left(\theta_* - \bar{\theta}_i\right)^\top - \frac{1}{i-1}\Sigma_* - \frac{\sigma^2 \sum_{j=1}^{i-1}\mathbb{E}\left[V_{j,\mathscr{T}_j}^{-1}\right]}{(i-1)^2} \right\|_{op} \\
\leq& \frac{16(\overline{\lambda}\lambda_e^2 + 16\sigma^2 d)(5d + 2\log_e(2/\delta))}{\lambda_e^2 (i-1)}.
\end{aligned}$$

Combining these with a union bound yields the result. □

The proof of Lemma 32 directly follows as shown below.

Proof. Proof of Lemma 32 When the event $\mathscr{A}$ holds, we can apply Lemma 67 to Eq. (C.21). It is helpful to note that $(i-1)/(i-2) \le 2$ and $1/(i-1) \le 2/i$ for all $i \ge 3$, and $5d + 2\log_e(2/\delta) \le 10d\log_e(2/\delta)$ for all $\delta \in [0, 2/e]$. By Lemma 57, the event $\mathscr{A}$ does not hold with probability at most $2/(N^2T^2)$. Thus, a second union bound yields the result. □

C.5 `Meta-DP++` Regret Analysis

As discussed in Section 4.3.3, we consider two cases; we first focus on the more substantive case where $N > N_1$.

We define a new clean event

$$\mathscr{J} = \Bigg\{ \forall i \ge N_1\,,\ \mathscr{T}_i \le \mathscr{T}_e\,, \quad \left\|\hat{\theta}_i - \theta_*\right\| \le 8\sqrt{\frac{(\sigma^2/\lambda_e + 5\overline{\lambda})d\log_e(4dN^2T)}{i}}\,, \\ \left\|\hat{\Sigma}_i - \Sigma_*\right\|_{op} \le \frac{128(\overline{\lambda}\lambda_e^2 + 16\sigma^2 d)}{\lambda_e^2}\left(\sqrt{\frac{5d\log_e(2N^2T)}{i}} \vee \frac{5d\log_e(2N^2T)}{i}\right), \\ \|\theta_i\| \le S + 5\sigma\sqrt{d\log_e(2N^2T)}\Bigg\}, \tag{C.23}$$

which stipulates that for every epoch after the initial N_1 exploration epochs, (i) the event $\mathscr{A}$ defined in Eq. (C.2) holds, ensuring that the number of exploration periods per epoch is small, (ii) our estimated prior mean $\hat{\theta}_i$ is close to the unknown prior mean θ_*, (iii) our estimated prior covariance $\hat{\Sigma}_i$ is close to the unknown prior covariance Σ_*, and (iv) the true parameter for epoch i $\theta_i \sim \mathcal{N}(\theta_*, \Sigma_*)$ is not too large in the ℓ_2-norm. These events all hold with high probability based on Lemma 57, 29, and 32, and by the properties of multivariate Gaussians respectively; therefore the event $\mathscr{J}$ holds with high probability.

Denote the meta regret of epoch i conditioned on the event $\mathscr{J}$ defined in Eq. (C.23) as $\mathscr{R}_{N,T}(i) \mid \mathscr{J}$. As noted earlier, during the exploration periods $1 \le t \le \mathscr{T}_i$, the meta oracle and our `Meta-DP++` algorithm encounter the same covariates $\{x_{i,t}\}_{t=1}^{T}$ and offer the same prices; thus, by construction, they achieve the same expected revenue and the resulting

meta regret is 0. Then, as in the proof of Theorem 28, we can write

$$
\begin{aligned}
&\mathscr{R}_{N,T}(i) \mid \mathscr{J} \\
=&\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}},\chi_i^{\mathrm{MDP}}}\left[\mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i+1}^{\mathrm{TS}},T-\mathscr{T}_i\right)-\mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},\Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},T-\mathscr{T}_i\right)\mid\mathscr{J}\right] \\
=&\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{MDP}}}\left[\mathtt{REV}_*\left(\theta_i,T-\mathscr{T}_i\right)-\mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},\Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},T-\mathscr{T}_i\right)\mid\mathscr{J}\right] \\
&-\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i,T-\mathscr{T}_i\right)-\mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i+1}^{\mathrm{TS}},T-\mathscr{T}_i\right)\mid\mathscr{J}\right]. \qquad \text{(C.24)}
\end{aligned}
$$

Appendix C.5.1 states two intermediate lemmas and Appendix C.5.2 provides the proof of Theorem 31.

C.5.1 Intermediate Lemmas

First, as we did for the proof of Theorem 28, we characterize the meta regret accrued by aligning the mean of the meta oracle's posterior $\theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}}$ and the mean of our `Meta-DP++` algorithm $\theta_{i,\mathscr{T}_i+1}^{\mathrm{MDP}}$.

Lemma 68. *For an epoch $i \geq N_1$,*

$$
\begin{aligned}
&\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{MDP}}}\left[\mathtt{REV}_*\left(\theta_i,T-\mathscr{T}_i\right)-\mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},\Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},T-\mathscr{T}_i\right)\mid\mathscr{J}\right] \\
\leq&\left(1+\frac{16c_3d^{3/2}\mathscr{T}_i\log_e^{3/2}(4dN^2T)}{\sqrt{i}}\right) \\
&\times\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i,T-\mathscr{T}_i\right)-\mathtt{REV}\left(\theta_i,\theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}},\Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MDP}},T-\mathscr{T}_i\right)\Big|\mathscr{J}\right]+O\left(\frac{1}{N}\right).
\end{aligned}
$$

Proof. (Proof of Lemma 68) By the posterior update rule of Bayesian linear regression [43], we have

$$
\begin{aligned}
\theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}}&=\left(\Sigma_*^{-1}+\sigma\sum_{t=1}^{\mathscr{T}_i}m_{i,t}m_{i,t}^\top\right)^{-1}\left(\Sigma_*^{-1}\theta_*+\sigma\sum_{t=1}^{\mathscr{T}_i}m_{i,t}m_{i,t}^\top\theta_i+\sigma\sum_{t=1}^{\mathscr{T}_i}m_{i,t}\varepsilon_{i,t}^{\mathrm{TS}}\right), \\
\theta_{i,\mathscr{T}_i+1}^{\mathrm{MDP}}&=\left(\left(\hat{\Sigma}_i^w\right)^{-1}+\sigma\sum_{t=1}^{\mathscr{T}_i}m_{i,t}m_{i,t}^\top\right)^{-1}\left(\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i+\sigma\sum_{t=1}^{\mathscr{T}_i}m_{i,t}m_{i,t}^\top\theta_i+\sigma\sum_{t=1}^{\mathscr{T}_i}m_{i,t}\varepsilon_{i,t}^{\mathrm{MDP}}\right).
\end{aligned}
$$

Denoting $M_i = \begin{pmatrix} m_{i,1} & \dots & m_{i,\mathscr{T}_i} \end{pmatrix} \in \mathscr{R}^{2d \times \mathscr{T}_i}$, we observe that prior alignment is achieved with $\theta^{\text{MDP}}_{i,\mathscr{T}_i+1} = \theta^{\text{TS}}_{i,\mathscr{T}_i+1}$ when the following holds:

$$\begin{aligned}
&\chi_i^{\text{TS}} - \chi_i^{\text{MDP}} \\
&= \underbrace{\frac{1}{\sigma}(M_i^\top M_i)^{-1}\left[\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i - \Sigma_*^{-1}\theta_* + \left(\Sigma_*^{-1} - \left(\hat{\Sigma}_i^w\right)^{-1}\right)\left(\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i + \sigma M_i M_i^\top \theta_i + M_i \chi_i^{\text{MDP}}\right)\right]}_{\Delta_n}.
\end{aligned} \tag{C.25}$$

We denote the RHS of the above equation as Δ_n for ease of exposition. While this expression is more complicated than Eq. (4.10), it still induces a mapping between χ_i^{TS} and χ_i^{MDP}. We then proceed similarly to the proof of Lemma 65. We start by expanding

$$\begin{aligned}
&\mathbb{E}_{\chi_i^{\text{MDP}}}\left[\texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \middle| \mathscr{J}\right] \\
&= \int_{\chi_i^{\text{MDP}}} \frac{\exp\left(-\left\|\chi_i^{\text{MDP}}\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \left(\texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right)\right) d\chi_i^{\text{MDP}} | \mathscr{J}.
\end{aligned}$$

Given a realization of χ_i^{MDP}, we denote $\chi_i^{\text{TS}}(\chi_i^{\text{MDP}})$ (with some abuse of notation) as the corresponding realization of χ_i^{TS} that satisfies Eq. (C.25). It is easy to see that this is a unique one-to-one mapping. We then perform a change of measure (similar to Eq. (C.13)) to continue:

$$\begin{aligned}
&\int_{\chi_i^{\text{MDP}}} \frac{\exp\left(-\left\|\chi_i^{\text{MDP}}\right\|^2/2\sigma^2\right)}{\exp\left(-\left\|\chi_i^{\text{TS}}(\chi_i^{\text{MDP}})\right\|^2/2\sigma^2\right)} \frac{\exp\left(-\left\|\chi_i^{\text{TS}}(\chi_i^{\text{MDP}})\right\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\mathscr{T}_i/2}} \\
&\quad \times \left(\texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) d\chi_i^{\text{MDP}}\right) | \mathscr{J} \\
&\leq \max_{\left\|\chi_i^{\text{MDP}}\right\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\text{TS}}(\chi_i^{\text{MDP}})\right\|^2 - \left\|\chi_i^{\text{MDP}}\right\|^2}{2\sigma^2}\right) \\
&\quad \times \mathbb{E}_{\chi_i^{\text{TS}}}\left[\texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \middle| \mathscr{J}\right] \\
&\quad + \mathbb{E}_{\chi_i^{\text{MDP}}}\left[\texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right) \middle| \mathscr{J}, \left\|\chi_i^{\text{MDP}}\right\| \geq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right] \\
&\qquad \times \Pr\left(\left\|\chi_i^{\text{MDP}}\right\| \geq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MDP}})\right\|^2 - \left\|\chi_i^{\mathrm{MDP}}\right\|^2}{2\sigma^2}\right) \\
&\quad \times \mathbb{E}_{\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \mathtt{REV}\left(\theta_i, \theta_{i,\mathscr{T}_i+1}^{\mathrm{TS}}, \Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MDP}}, T-\mathscr{T}_i\right) \middle| \mathscr{J}\right] \\
&\quad + \frac{\sqrt{\kappa+S^2}\, p_{\max} x_{\max}\sqrt{1+p_{\max}^2}}{N},
\end{aligned} \tag{C.26}
$$

where the last step follows from Eqs. (C.19) and (C.20). Thus, we have expressed the true regret of our `Meta-DP++` algorithm as the sum of a term that is proportional to the true regret of a policy that is aligned with the meta oracle (*i.e.*, it employs the prior $\mathscr{N}(\theta_{i,\mathscr{T}_i+1}^{\mathrm{MDP}}, \Sigma_{i,\mathscr{T}_i+1}^{\mathrm{MDP}})$), and an additional term that is small (*i.e.*, scales as $1/N$).

We now characterize the coefficient of the first term in Eq. (C.26):

$$
\begin{aligned}
&\max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MDP}})\right\|^2 - \left\|\chi_i^{\mathrm{MDP}}\right\|^2}{2\sigma^2}\right) \\
&= \max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{MDP}} + \Delta_n\right\|^2 - \left\|\chi_i^{\mathrm{MDP}}\right\|^2}{2\sigma^2}\right) \\
&= \max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left(\chi_i^{\mathrm{MDP}}\right)^\top \Delta_n}{\sigma^2} + \frac{\|\Delta_n\|^2}{2\sigma^2}\right) \\
&\leq \max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{MDP}}\right\| \|\Delta_n\|}{\sigma^2} + \frac{\|\Delta_n\|^2}{2\sigma^2}\right) \\
&= \max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{4\sqrt{t_e \log_e(2NT)}\|\Delta_n\|}{\sigma} + \frac{\|\Delta_n\|^2}{2\sigma^2}\right).
\end{aligned} \tag{C.27}
$$

To continue, we must characterize $\|\Delta_n\|$. Applying the triangle inequality, we have that

$$
\begin{aligned}
&\|\Delta_n\| \\
&\leq \frac{1}{\sigma\lambda_e}\left\|\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i - \Sigma_*^{-1}\theta_*\right\| + \frac{1}{\sigma\lambda_e}\left\|\left(\Sigma_*^{-1} - \left(\hat{\Sigma}_i^w\right)^{-1}\right)\left(\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i + \sigma M_i M_i^\top \theta_i + M_i \chi_i^{\mathrm{MDP}}\right)\right\|.
\end{aligned} \tag{C.28}
$$

The first term of Eq. (C.28) satisfies

$$
\begin{aligned}
&\frac{1}{\sigma\lambda_e}\left\|\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i - \Sigma_*^{-1}\theta_*\right\| \\
&= \frac{1}{\sigma\lambda_e}\left\|\Sigma_*^{-1}\left(\hat{\theta}_i - \theta_*\right) + \left(\left(\hat{\Sigma}_i^w\right)^{-1} - \Sigma_*^{-1}\right)\left(\hat{\theta}_i - \theta_*\right) + \left(\left(\hat{\Sigma}_i^w\right)^{-1} - \Sigma_*^{-1}\right)\theta_*\right\|
\end{aligned}
$$

$$\leq \frac{1}{\sigma\lambda_e}\left\|\Sigma_*^{-1}\left(\hat{\theta}_i-\theta_*\right)\right\|+\frac{1}{\sigma\lambda_e}\left\|\left(\left(\hat{\Sigma}_i^w\right)^{-1}-\Sigma_*^{-1}\right)\left(\hat{\theta}_i-\theta_*\right)\right\|+\frac{1}{\sigma\lambda_e}\left\|\left(\left(\hat{\Sigma}_i^w\right)^{-1}-\Sigma_*^{-1}\right)\theta_*\right\|$$
$$\leq 8\sqrt{\frac{(\sigma^2/\lambda_e+5\overline{\lambda})d\log_e(4dN^2T)}{\sigma^2\lambda_e^2 i}}\left(\frac{1}{\underline{\lambda}}+\left\|\left(\hat{\Sigma}_i^w\right)^{-1}-\Sigma_*^{-1}\right\|_{op}\right)+\frac{S}{\sigma\lambda_e}\left\|\left(\hat{\Sigma}_i^w\right)^{-1}-\Sigma_*^{-1}\right\|_{op}. \tag{C.29}$$

Next, the second term of Eq. (C.28) satisfies

$$\begin{aligned}
&\frac{1}{\sigma\lambda_e}\left\|\left(\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right)\left(\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i+\sigma M_i M_i^\top\theta_i+M_i\chi_i^{\text{MDP}}\right)\right\| \\
&\leq \frac{\left\|\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}}{\sigma\lambda_e}\left(\left\|\left(\hat{\Sigma}_i^w\right)^{-1}\hat{\theta}_i\right\|+\left\|\sigma M_i M_i^\top\theta_i\right\|+\left\|M_i\chi_i^{\text{MDP}}\right\|\right) \\
&\leq \frac{\left\|\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}}{\sigma\lambda_e}\left(\left\|\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}(S+1)+\sigma\mathscr{T}_i x_{\max}^2(p_{\max}^2+p_{\max}^4)\right. \\
&\quad \left.+4\sigma p_{\max}x_{\max}\sqrt{\mathscr{T}_i(1+p_{\max}^2)\log_e(2NT)}\right) \\
&\leq \frac{\left\|\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}}{\sigma\lambda_e}\left(\left\|\Sigma_*^{-1}\right\|_{op}(S+1)+\sigma\mathscr{T}_i x_{\max}^2(p_{\max}^2+p_{\max}^4)\right. \qquad \text{(C.30)} \\
&\quad \left.+4\sigma p_{\max}x_{\max}\sqrt{\mathscr{T}_i(1+p_{\max}^2)\log_e(2NT)}\right) \\
&\leq \frac{8p_{\max}x_{\max}\sqrt{\mathscr{T}_i(1+p_{\max}^2)\log_e(2NT)}\left\|\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}}{\lambda_e}, \qquad \text{(C.31)}
\end{aligned}$$

where Eq. (C.30) follows from the fact that $\|\hat{\Sigma}_i^w\|_{op}\geq\|\Sigma_*\|_{op}$ (on the event $\mathscr{J}$) and because both matrices are positive semi-definite (since they are covariance matrices). Applying Lemma 79, we can simplify the term

$$\begin{aligned}
\left\|\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op} &= \left\|\left(\hat{\Sigma}_i^w\right)^{-1}(\hat{\Sigma}_i^w-\Sigma_*)\Sigma_*^{-1}\right\|_{op} \\
&\leq \left\|\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}\left\|\hat{\Sigma}_i^w-\Sigma_*\right\|_{op}\left\|\Sigma_*^{-1}\right\|_{op} \\
&\leq \frac{256(\overline{\lambda}\lambda_e^2+16\sigma^2 d)}{\lambda_e^2\underline{\lambda}^2}\sqrt{\frac{5d\log_e(2N^2T)}{i}}. \qquad \text{(C.32)}
\end{aligned}$$

Combining Eqs. (C.28)–(C.32), we have

$$\|\Delta_n\| \leq c_3 \sigma d \sqrt{\frac{d \mathscr{T}_i \log_e(4dN^2T) \log_e(2N^2T)}{i}}.$$

Substituting this expression into Eq. (C.27), we can bound the coefficient

$$\begin{aligned}
&\max_{\|\chi_i^{\mathrm{MDP}}\| \leq 4\sigma\sqrt{\mathscr{T}_i \log_e(2NT)}} \exp\left(\frac{\left\|\chi_i^{\mathrm{TS}}(\chi_i^{\mathrm{MDP}})\right\|^2 - \left\|\chi_i^{\mathrm{MDP}}\right\|^2}{2\sigma^2}\right) \\
&\leq \exp\left(8c_3 d \mathscr{T}_i \log_e(2N^2T) \sqrt{\frac{d \log_e(4dN^2T)}{i}}\right) \\
&\leq 1 + 16 c_3 d \mathscr{T}_i \log_e^4(4dN^2T) \sqrt{\frac{d}{i}},
\end{aligned}$$

where we used Lemma 76 in the last step. Substituting into Eq. (C.26) yields the result. □

We will use Lemma 68 in the proof of Theorem 31 to characterize the meta regret from prior alignment. The next lemma will help us characterize the remaining meta regret due to the difference in the covariance matrices post-alignment.

Lemma 69. *When the event $\mathscr{J}$ holds, we can write*

$$\prod_{t=\mathscr{T}_i+1}^{T} \max_{\theta : \|\theta - \theta_{i,t}^{\mathrm{TS}}\| \leq C} \frac{d\mathcal{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}'^{\mathrm{MDP}}\right)}{d\mathcal{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}^{\mathrm{TS}}\right)} \leq 1 + \frac{2c_4 d^{5/2} T \log_e^{3/2}(2N^2T)}{\sqrt{i}} \leq 3.$$

Proof. (Proof of Lemma 69) By the definition of the multivariate normal distribution, we have

$$\begin{aligned}
&\max_{\theta : \|\theta - \theta_{i,t}^{\mathrm{TS}}\| \leq C} \frac{d\mathcal{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}'^{\mathrm{MDP}}\right)}{d\mathcal{N}\left(\theta_{i,t}^{\mathrm{TS}}, \Sigma_{i,t}^{\mathrm{TS}}\right)} \\
&= \sqrt{\frac{\det\left(\Sigma_{i,t}^{\mathrm{TS}}\right)}{\det\left(\Sigma_{i,t}'^{\mathrm{MDP}}\right)}} \times \\
&\max_{\theta : \|\theta - \theta_{i,t}^{\mathrm{TS}}\| \leq C} \exp\left(\frac{\left(\theta - \theta_{i,t}^{\mathrm{TS}}\right)^\top \left(\Sigma_{i,t}^{\mathrm{TS}}\right)^{-1} \left(\theta - \theta_{i,t}^{\mathrm{TS}}\right)}{2} - \frac{\left(\theta - \theta_{i,t}^{\mathrm{TS}}\right)^\top \left(\Sigma_{i,t}'^{\mathrm{MDP}}\right)^{-1} \left(\theta - \theta_{i,t}^{\mathrm{TS}}\right)}{2}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\frac{\det(\Sigma_{i,t}^{\mathrm{TS}})}{\det(\Sigma_{i,t}^{\prime\mathrm{MDP}})}} \max_{\theta:\left\|\theta-\theta_{i,t}^{\prime\mathrm{MDP}}\right\|\leq C} \exp\left(\frac{\left(\theta-\theta_{i,t}^{\mathrm{TS}}\right)^{\top}\left(\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right)\left(\theta-\theta_{i,t}^{\mathrm{TS}}\right)}{2}\right) \\
&\leq \sqrt{\frac{\det\left(\left(\hat{\Sigma}_i^w\right)^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)}{\det\left(\Sigma_*^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)}} \exp\left(\frac{C^2\left\|\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}\right\|_{op}}{2}\right) \\
&\leq \sqrt{\frac{\det\left(\left(\hat{\Sigma}_i^w\right)^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)}{\det\left(\Sigma_*^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)}} \exp\left(\frac{128C^2(\overline{\lambda}\lambda_e^2+16\sigma^2 d)}{\lambda_e^2\underline{\lambda}^2}\sqrt{\frac{5d\log_e(2N^2T)}{i}}\right),
\end{aligned}
$$

where we have used Eq. (C.32) in the last step. Since our estimated covariance matrix is widened, we know that on the event $\mathscr{J}$, $\Sigma_*^{-1}-\left(\hat{\Sigma}_i^w\right)^{-1}=\Sigma_*^{-1}\left(\hat{\Sigma}_i^w-\Sigma_*\right)\left(\hat{\Sigma}_i^w\right)^{-1}$ is positive semi-definite, and thus it is evident that $\left(\Sigma_*^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)-\left(\left(\hat{\Sigma}_i^w\right)^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)$ is also positive semi-definite. Therefore, conditioned on the clean event $\mathscr{J}$,

$$
\sqrt{\frac{\det\left(\left(\hat{\Sigma}_i^w\right)^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)}{\det\left(\Sigma_*^{-1}+\sum_{\tau=1}^{t-1} w_{i,\tau}w_{i,\tau}^{\top}\right)}} \leq 1\,.
$$

The result follows directly. □

C.5.2 Proof of Theorem 31

Proof. (Proof of Theorem 31) First, we consider the "small N" regime, where $N \leq N_1$. In this case, our `Meta-DP++` algorithm simply executes N instances prior-independent Thompson sampling. Then, an immediate consequence of Lemma 64 is that the meta regret is bounded by $N \times \widetilde{O}\left(d^2T^{1/2}\right) = \widetilde{O}\left(d^3(NT)^{5/6}\right)$ because $N \leq N_1 = O(d^4T^2)$. Thus, the result already holds in this case.

We now turn our attention to the "large N" regime, *i.e.*, $N > N_1$. The meta regret can be decomposed as

$$
\begin{aligned}
\mathscr{R}_{N,T} &= (\mathscr{R}_{N,T}|\,\mathscr{J}\,)\Pr(\,\mathscr{J}\,) + (\mathscr{R}_{N,T}|\neg\mathscr{J}\,)\Pr(\neg\mathscr{J}\,) \\
&\leq (\mathscr{R}_{N,T}|\,\mathscr{J}\,) + (\mathscr{R}_{N,T}|\neg\mathscr{J}\,)\Pr(\neg\mathscr{J}\,).
\end{aligned}
$$

Recall that the event $\mathscr{J}$ is composed of four events, each of which hold with high probability. Applying a union bound over the epochs $i \geq N_1 + 1$ to Lemma 57, Lemma 29 (setting $\delta = 1/(N^2T)$), Lemma 32 (with $\delta = 1/(N^2T)$), and Eq. (C.9) (with $u = 5\sigma\sqrt{d\log_e(2N^2T)}$), we obtain that

$$\Pr(\mathscr{J}) \geq 1 - 4/(NT) - 6/(NT^2) \geq 1 - 10/(NT).$$

Recall that when the event $\mathscr{J}$ is violated, the meta regret is $O(NT)$, so we can bound $(\mathscr{R}_{N,T}|\neg\mathscr{J})\Pr(\neg\mathscr{J}) = O(NT \times 1/(NT)) = O(1)$. Therefore, the overall meta regret is simply

$$\mathscr{R}_{N,T} \leq (\mathscr{R}_{N,T} \mid \mathscr{J}) + O(1). \tag{C.33}$$

Thus, it suffices to bound $\mathscr{R}_{N,T} \mid \mathscr{J}$. As described in Section 4.3.3, we consider bounding the meta regret post-alignment ($t = \mathscr{T}_i + 1, \cdots, T$), where our `Meta-DP++` algorithm follows the aligned posterior $\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})$. Let $\mathscr{N}(\theta^{\text{TS}}_{i,t}, \Sigma'^{\text{MDP}}_{i,t})$ denote the posterior of our `Meta-DP++` algorithm at time step t, if it begins with the prior $\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})$ in time step $\mathscr{T}_i + 1$, but follows the randomness of the oracle. Then, we can write

$$\begin{aligned}
&\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right)\middle|\mathscr{J}\right] \\
=&\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\int_\theta \texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}(\theta_i, \theta, 0, 1)\right. \\
&\left.-\texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+2}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+2}, T-\mathscr{T}_i-1\right) d\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})\middle|\mathscr{J}\right] \\
=&\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\int_{\theta:\|\theta\|\leq C} \texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}(\theta_i, \theta, 0, 1)\right. \\
&\left.-\texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+2}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+2}, T-\mathscr{T}_i-1\right) d\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})\middle|\mathscr{J}\right] \\
&+\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\int_{\theta:\|\theta\|> C} \texttt{REV}_*(\theta_i, T-\mathscr{T}_i) - \texttt{REV}(\theta_i, \theta, 0, 1)\right. \\
&\left.-\texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+2}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+2}, T-\mathscr{T}_i-1\right) d\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1})\middle|E\right] \\
\leq&\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\max_{\theta:\left\|\theta-\theta'^{\text{MDP}}_{i,t}\right\|\leq C} \frac{d\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma'^{\text{MDP}}_{i,\mathscr{T}_i+1})}{d\mathscr{N}(\theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1})}\left(\texttt{REV}_*(\theta_i, 1) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1}, 1\right)\right)\middle|\mathscr{J}\right]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\max_{\theta:\left\|\theta-\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}\right\|\le C}\frac{d\mathscr{N}(\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma'^{\mathrm{MDP}}_{i,\mathscr{T}_i+1})}{d\mathscr{N}(\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1})}\right.\\
&\quad\left.\times\left(\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i-1)-\mathtt{REV}\left(\theta_i,\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+2},\Sigma'^{\mathrm{MDP}}_{i,\mathscr{T}_i+2},T-\mathscr{T}_i-1\right)\right)\middle|\mathscr{J}\right]\\
&+\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\int_{\theta:\left\|\theta-\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1}\right\|>C}\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i)\,d\mathscr{N}(\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma^{\mathrm{MDP}}_{i,\mathscr{T}_i+1})\middle|\mathscr{J}\right],
\end{aligned}
$$

where $C=5\sigma\sqrt{d\log_e(NT)}$. Inductively, we have

$$
\begin{aligned}
&\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i)-\mathtt{REV}\left(\theta_i,\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma^{\mathrm{MDP}}_{i,\mathscr{T}_i+1},T-\mathscr{T}_i\right)\middle|\mathscr{J}\right]\\
\le&\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\prod_{t=\mathscr{T}_i+1}^{T}\max_{\theta:\|\theta-\theta^{\mathrm{TS}}_{i,t}\|\le C}\frac{d\mathscr{N}(\theta^{\mathrm{TS}}_{i,t},\Sigma'^{\mathrm{MDP}}_{i,t})}{d\mathscr{N}(\theta^{\mathrm{TS}}_{i,t},\Sigma^{\mathrm{TS}}_{i,t})}\right.\\
&\quad\left.\times\left(\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i)-\mathtt{REV}\left(\theta_i,\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1},T-\mathscr{T}_i\right)\right)\middle|\mathscr{J}\right]\\
&\quad+\sum_{t=\mathscr{T}_i+1}^{T}\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\prod_{t=\mathscr{T}_i+2}^{T}\max_{\theta:\|\theta-\theta^{\mathrm{TS}}_{i,t}\|\le C}\frac{d\mathscr{N}(\theta^{\mathrm{TS}}_{i,t},\Sigma^{\mathrm{MDP}}_{i,t})}{d\mathscr{N}(\theta^{\mathrm{TS}}_{i,t},\Sigma^{\mathrm{TS}}_{i,t})}\right.\\
&\qquad\left.\times\int_{\theta:\|\theta\|>C}\mathtt{REV}_*(\theta_i,T-t)\,d\mathscr{N}(\theta^{\mathrm{TS}}_{i,t},\Sigma'^{\mathrm{MDP}}_{i,t})\middle|\mathscr{J}\right].
\end{aligned}
\tag{C.34}
$$

Applying Lemma 69, we can bound Eq. (C.34) as

$$
\begin{aligned}
&\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\mathrm{TS}}}\left[\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i)-\mathtt{REV}\left(\theta_i,\theta^{\mathrm{TS}}_{i,t_e+1},\Sigma^{\mathrm{MDP}}_{i,t_e+1},T-\mathscr{T}_i\right)\middle|\mathscr{J}\right]\\
\le&\left(1+\frac{2c_4d^{5/2}T\log_e^{3/2}(2N^2T)}{\sqrt{i}}\right)\\
&\quad\times\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i)-\mathtt{REV}\left(\theta_i,\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1},T-\mathscr{T}_i\right)\middle|\mathscr{J}\right]\\
&\quad+\sum_{t=\mathscr{T}_i+1}^{T}\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[3\int_{\theta:\|\theta\|>C}\mathtt{REV}_*(\theta_i,T-t)\,d\mathscr{N}(\theta^{\mathrm{TS}}_{i,t},\Sigma'^{\mathrm{MDP}}_{i,t})\middle|\mathscr{J}\right]\\
=&\left(1+\frac{2c_4d^{5/2}T\log_e^{3/2}(2N^2T)}{\sqrt{i}}\right)\\
&\quad\times\mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\mathtt{REV}_*(\theta_i,T-\mathscr{T}_i)-\mathtt{REV}\left(\theta_i,\theta^{\mathrm{TS}}_{i,\mathscr{T}_i+1},\Sigma^{\mathrm{TS}}_{i,\mathscr{T}_i+1},T-\mathscr{T}_i\right)\middle|\mathscr{J}\right]+O\left(\frac{1}{N}\right),
\end{aligned}
$$

where we used Eq. (C.9) in the last step. Thus, we have expressed the post-alignment meta regret as the sum of a term that is proportional to the true regret of the meta oracle and

a negligibly small term. We can now apply Lemma 68 to further include the meta regret accrued from our prior alignment step to obtain

$$\begin{aligned}
&\mathbb{E}_{\theta_i,\hat{\theta}_i,\chi_i^{\text{MDP}}}\left[\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\text{MDP}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{MDP}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right)\middle|\mathscr{J}\right] \\
&\leq \left(1+\frac{16c_3 d^{3/2}\mathscr{T}_i \log_e^{3/2}(4dN^2T)}{\sqrt{i}}\right)\left(1+\frac{2c_4 d^{5/2}T\log_e^{3/2}(2N^2T)}{\sqrt{i}}\right) \\
&\quad \times \mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right)\middle|E\right] + O\left(\frac{1}{N}\right).
\end{aligned}$$

As desired, this establishes that the coefficient of our first term decays to 1 as i grows large. Thus, our meta regret from the first term approaches 0 for large i, and all other terms are clearly negligible.

Noting that $N > N_1 = \tilde{O}(d^4T^2)$ in the "large N" regime, we can upper bound the meta regret as

$$\begin{aligned}
&\sum_{i=N_1+1}^{N}\left[\left(1+\frac{16c_3 d^{3/2}\mathscr{T}_i \log_e^{3/2}(4dN^2T)}{\sqrt{i}}\right)\left(1+\frac{2c_4 d^{5/2}T\log_e^{3/2}(2N^2T)}{\sqrt{i}}\right)-1\right] \\
&\quad \times \mathbb{E}_{\theta_i,\hat{\theta}_i,}\left[\texttt{REV}_*\left(\theta_i, T-\mathscr{T}_i\right) - \texttt{REV}\left(\theta_i, \theta^{\text{TS}}_{i,\mathscr{T}_i+1}, \Sigma^{\text{TS}}_{i,\mathscr{T}_i+1}, T-\mathscr{T}_i\right)\middle|\mathscr{J}\right] + O\left(\frac{1}{N}\right) \\
&= \tilde{O}\left(\sum_{i=N_1+1}^{N}\frac{d^4T^{\frac{3}{2}}}{\sqrt{i}}\right) \;=\; \tilde{O}\left(d^4N^{\frac{1}{2}}T^{\frac{3}{2}}\right) \;=\; \tilde{O}\left(d^2(NT)^{\frac{5}{6}}\right).
\end{aligned}$$

□

C.6 Extension to Multiple Products with Substitution Effects

Thus far, we have considered the setting where the seller offers a single product in each epoch. In practice, there may be many products offered simultaneously in an epoch, and there may be substitution effects across these products (within a single epoch) that must be additionally modeled. We now show that our transfer learning approach extends straightforwardly to this setting.

C.6.1 Formulation

We extend our single-product epoch formulation from Section 4.1 to a multi-product epoch formulation, where K products are offered in each epoch. To capture substitution effects *within* an epoch, we will employ an epoch-level joint demand model across all K products. Our demand model is an extension of the multi-product demand model proposed by [117], with the addition of (exogenous, product-specific and customer-specific) features. The seller will now choose a price *vector* (one for each product), observe a demand vector, and estimate the demand function jointly across all products given the price/demand data.

As before, in epoch $i \in [N]$ at time $t \in [T]$, the seller observes a random feature vector $x_{i,t} \in \mathscr{R}^d$, which is sampled i.i.d. from a known distribution $\mathscr{P}_i^{mp}$. She then chooses a price vector $p_{i,t}^{mp} = \begin{pmatrix} p_{i,t,1}^{mp} & \dots p_{i,t,K}^{mp} \end{pmatrix}^\top \in \mathscr{R}^K$, where $p_{i,t,k}^{mp}$ is the chosen price for product $k \in [K]$ in time t and epoch i. Recall that, owing to practical constraints, we assume that the allowable price range is bounded across periods and products, *i.e.*, $p_{i,t}^{mp} \in [p_{\min}, 1]^K$ and that $0 < p_{\min} < 1$.[1] The seller then observes the resulting induced demand for product $k \in [K]$,

$$D_{i,t,k}^{mp}(p_{i,t}^{mp}, x_{i,t}) = \langle \alpha_{i,k}^{mp}, x_{i,t} \rangle + \sum_{j=1}^{K} p_{i,t,j}^{mp} \langle \beta_{i,k,j}^{mp}, x_{i,t} \rangle + \varepsilon_{i,t,k}^{mp},$$

where $\alpha_{i,k}^{mp} \in \mathscr{R}^d$ and $\beta_{i,k,j}^{mp} \in \mathscr{R}^d$ are unknown fixed constants throughout epoch i, and $\varepsilon_{i,t,k}^{mp} \sim \mathscr{N}(0, \sigma^2)$ is i.i.d. Gaussian noise with variance σ^2.

Observe that the demand for product k now depends not only on the price of product k but also on the prices of all other products in this epoch — in particular, $\beta_{i,k,j}$ for $j \neq k$, captures the substitution effects between products k and j under feature vector $x_{i,t}$. For ease of notation, we collectively denote the demand vector

$$D_{i,t}^{mp}(p_{i,t}^{mp}, x_{i,t}) = \begin{pmatrix} D_{i,t,1}^{mp}(p_{i,t}^{mp}, x_{i,t}) & \dots & D_{i,t,K}^{mp}(p_{i,t}^{mp}, x_{i,t}) \end{pmatrix}. \tag{C.35}$$

[1] Note that we have set $p_{\max} = 1$; this is done WLOG since we can always normalize our parameters appropriately.

Shared Structure: For ease of notation, we additionally define the matrix

$$\theta_i^{mp} = \begin{pmatrix} \alpha_{i,1}^{mp} & \dots & \alpha_{i,K}^{mp} \\ \beta_{i,1,1}^{mp} & \dots & \beta_{i,K,1}^{mp} \\ \vdots & \dots & \vdots \\ \beta_{i,1,K}^{mp} & \dots & \beta_{i,K,K}^{mp} \end{pmatrix} \in \mathscr{R}^{(K+1)d \times K},$$

where θ_i^{mp} is the unknown parameter matrix 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 $\{\theta_i^{mp}\}_{i=1}^N$, our problem reduces to N independent dynamic pricing problems.

However, as discussed in the main chapter, we may have some shared structure that can be related across products. We model the shared structure by positing that product demand parameters $\{\theta_i^{mp}\}_{i=1}^N$ are independent and identically distributed draws from a common unknown matrix normal distribution,[2] *i.e.*, $\theta_i^{mp} \sim \mathscr{MN}(\theta_*^{mp}, \Sigma_*^{mp}, I_K)$ for each $i \in [N]$. (The third argument is I_K because the noise terms are uncorrelated by assumption.)

Assumptions: We impose the same assumptions made in Section 4.1.2. However, since we are now learning $(K^2+K)d$ instead of just $2d$ parameters (in the single-product case), we may naturally expect that the constants to differ. Specifically, we take the constants in Assumption 8 to be $x_{\max}$ and S^{mp}; similarly, we take the constant in Assumption 10 to be $\overline{\lambda}^{mp}$ and $\underline{\lambda}^{mp}$ for the multi-product setting.

Meta Oracle: As before, we define our meta oracle to be Thompson Sampling with a known prior. Here, our meta oracle is $\mathrm{TS}\left(\mathscr{MN}\left(\theta_*^{mp}, \Sigma_*^{mp}, I_K\right), \lambda_e^{mp}\right)$, the Thompson sampling algorithm with prior $\mathscr{MN}\left(\theta_*^{mp}, \Sigma_*^{mp}, I_K\right)$ and an input parameter λ_e^{mp}. The description is formally given in Algorithm 10 below. As before, we perform random price

[2] See, *e.g.*, [95] for the definition and properties of a matrix normal distribution.

exploration for $\tilde{O}(1)$ time periods by offering initial prices

$$p^{(1)} = \begin{pmatrix} p_{\min} \\ p_{\min} \\ p_{\min} \\ \vdots \\ p_{\min} \end{pmatrix}, \quad p^{(2)} = \begin{pmatrix} 1 \\ p_{\min} \\ p_{\min} \\ \vdots \\ p_{\min} \end{pmatrix}, \quad p^{(3)} = \begin{pmatrix} p_{\min} \\ 1 \\ p_{\min} \\ \vdots \\ p_{\min} \end{pmatrix}, \quad \cdots \quad p^{(K+1)} = \begin{pmatrix} p_{\min} \\ p_{\min} \\ \vdots \\ p_{\min} \\ 1 \end{pmatrix}. \tag{C.36}$$

The random exploration period ends once the minimum eigenvalue of the matrix

$$\sum_{s=1}^{t} \begin{pmatrix} x_{i,s}^\top & p_{i,s,1}^{mp} x_{i,s}^\top & \dots p_{i,s,K}^{mp} x_{i,s}^\top \end{pmatrix}^\top \begin{pmatrix} x_{i,s}^\top & p_{i,s,1}^{mp} x_{i,s}^\top & \dots p_{i,s,K}^{mp} x_{i,s}^\top \end{pmatrix},$$

exceeds λ_e^{mp}. For each subsequent time step, the meta oracle (1) samples the unknown product demand parameters

$$\mathring{\theta}_{i,t}^{mp} = \begin{pmatrix} \mathring{\alpha}_{i,t,1}^{mp} & \dots & \mathring{\alpha}_{i,t,K}^{mp} \\ \mathring{\beta}_{i,t,1,1}^{mp} & \dots & \mathring{\beta}_{i,t,K,1}^{mp} \\ \vdots & \dots & \vdots \\ \mathring{\beta}_{i,t,1,K}^{mp} & \dots & \mathring{\beta}_{i,t,K,K}^{mp} \end{pmatrix},$$

from the posterior $\mathscr{N}\left(\theta_{i,t}^{\mathrm{TS}}, I_K \otimes \Sigma_{i,t}^{\mathrm{TS}}\right)$, and (2) solves and offers the resulting optimal price based on the demand function given by the sampled parameters

$$p_{i,t}^{\mathrm{TS}} = \underset{p \in [p_{\min}, 1]^K}{\arg\max} \sum_{k=1}^{K} \left[p_k \left(\left\langle \mathring{\alpha}_{i,t,k}, x_{i,t} \right\rangle + \sum_{j=1}^{K} p_j \cdot \left\langle \mathring{\beta}_{i,t,k,j}, x_{i,t} \right\rangle \right) \right]. \tag{C.37}$$

Upon observing the actual realized demand $D_{i,t}\left(p_{i,t}^{\mathrm{TS}}, x_{i,t}\right)$, the algorithm computes the posterior $\mathscr{MN}\left(\theta_{i,t+1}^{\mathrm{TS}}, \Sigma_{i,t+1}^{\mathrm{TS}}, I_K\right)$ for round $t+1$ [158]. The same algorithm is applied independently to each epoch $i \in [N]$.

The following theorem bounds the Bayes regret of our meta oracle:

Algorithm 10 TS$\left(\mathscr{MN}\left(\theta_*^{mp},\Sigma_*^{mp},I_K\right),\lambda_e^{mp}\right)$: Thompson Sampling Algorithm

Input: The prior mean matrix θ_*^{mp} and covariance matrix Σ_*^{mp}, the index i of epoch, the length of each epoch T, the noise parameter σ, exploration parameter λ_e.
Initialization: $t \leftarrow 1$, $\left(\theta_{i,t}^{\text{TS}},\Sigma_{i,t}^{\text{TS}}\right) \leftarrow \left(\theta_*^{mp},\Sigma_*^{mp}\right)$,
while $\lambda_{\min}\left(\sum_{s=1}^{t-1}\left(x_{i,s}^\top \quad p_{i,s,1}x_{i,s}^\top \quad \dots p_{i,s,K}x_{i,s}^\top\right)^\top\left(x_{i,s}^\top \quad p_{i,s,1}x_{i,s}^\top \quad \dots p_{i,s,K}x_{i,s}^\top\right)\right) \leq \lambda_e$ **do**
    Observe feature vector $x_{i,t}$, and offer price $p_{i,t}^{\text{TS}} \leftarrow p^{(t \mod K)}$
    Observe demand $D_{i,t}\left(p_{i,t}^{\text{TS}},x_{i,t}\right)$, and compute the posterior $\mathscr{MN}\left(\theta_{i,t+1}^{\text{TS}},\Sigma_{i,t+1}^{\text{TS}},I_K\right)$.
    $t \leftarrow t+1$.
end while
while $t \leq T$ **do**
    Observe feature vector $x_{i,t}$.
    Sample parameter $\mathring{\theta}_{i,t} \sim \mathscr{MN}\left(\theta_{i,t}^{\text{TS}},\Sigma_{i,t}^{\text{TS}},I_K\right)$.
    Offer $p_{i,t}^{\text{TS}}$ according to eq. (C.37).
    Observe demand $D_{i,t}\left(p_{i,t}^{\text{TS}},x_i\right)$, and compute the posterior $\mathscr{MN}\left(\theta_{i,t+1}^{\text{TS}},\Sigma_{i,t+1}^{\text{TS}},I_K\right)$.
    $t \leftarrow t+1$.
end while

Corollary 70 (Multi-Product meta oracle)**.** *The Bayes regret of Algorithm 10 satisfies*

$$\text{Bayes Regret}_{N,T}(\pi) = \tilde{O}\left(K^3 d^{\frac{3}{2}} N\sqrt{T}\right),$$

when the prior over the product demand parameters is known.

Corollary 70 follows directly from Theorem 27 in the single-product case. This is because, if a matrix X follows the matrix Gaussian distribution $\mathscr{MN}(A,B,C)$, then vec(X), (*i.e.*, the vectorized version of X that concatenates each column of X to form a vector), follows the multivariate Gaussian distribution $\mathscr{N}(A,C\otimes B)$ [95]. Thus, since we still maintain a linear demand model, the only mathematical change is that the unknown parameter has dimension $(K^2+K)d$ instead of $2d$. Thus, the same result applies by replacing the d in Theorem 27 with $(K^2+K)d$.

C.6.2 Multi-Product `Meta-DP` Algorithm

The multi-product `Meta-DP` algorithm is presented in Algorithm 11. We first define some additional notation, and then describe the algorithm in detail.

Additional Notation: Analogous to our previous notation, we use

$$m_{i,t}^{mp} = \begin{pmatrix} x_{i,t} \\ p_{i,t,1}^{mp} x_{i,t} \\ \vdots \\ p_{i,t,K}^{mp} x_{i,t} \end{pmatrix},$$

to denote the price and feature information and $V_{i,t}^{mp} = \sum_{\tau=1}^{t} m_{i,t}^{mp} \left(m_{i,t}^{mp}\right)^{\top}$ to denote the Fisher information matrix of round t in epoch i for all $i \in [N]$ and $t \in [T]$.

Algorithm Description: The first N_0^{mp} epochs are treated as exploration epochs, where we define

$$N_0^{mp} = (c_2^{mp})^2 d^2 (K^2+K)^2 \log_e(4d(K^2+K)N^2T) \log_e(2NT) \lambda_e^{mp}, \tag{C.38}$$

and the constant c_2^{mp} is defined as

$$c_2^{mp} = \frac{32\sqrt{2x_{\max}^2(\sigma^2 \left(\lambda_e^{mp}\right)^{-1} + 5\overline{\lambda}^{mp})}}{\lambda_e^{mp} \underline{\lambda}^{mp} \sigma^2}.$$

As before, the `Meta-DP` algorithm proceeds differently for earlier exploration epochs and later epochs:

1. **Epoch i $\leq \mathbf{N_0^{mp}}$** : The `Meta-DP` algorithm runs the prior-independent Thompson sampling algorithm [10, 5] TS$(\mathscr{MN}(0, \Psi^{mp} \cdot I_{(K+1)d}, I_K), \lambda_e)$, where

$$\Psi^{mp} = \sigma\sqrt{2d\log_e(T(1+2x_{\max}^2 T))} + \sqrt{20\overline{\lambda}^{mp} d \log_e(2T)}.$$

2. **Epoch i $>$ $\mathbf{N_0^{mp}}$** : the `Meta-DP` algorithm computes the ordinary least square (OLS) estimate of the parameter vector θ_i for each of the past epochs; then, it averages these OLS estimates to arrive at an estimate $\hat{\theta}_i^{mp}$ of the prior mean θ_*, *i.e.*,

$$\hat{\theta}_i^{mp} = \frac{\sum_{j=1}^{i-1} \left(V_{j,T}^{mp}\right)^{-1} \left(\sum_{t=1}^{T} m_{j,t}^{mp} D_{j,t}^{mp}(p_{j,t}^{mp}, x_{j,t})\right)}{i-1}. \tag{C.39}$$

Then, the `Meta-DP` algorithm runs the Thompson sampling algorithm (see Algorithm 10) with the estimated prior $\mathscr{MN}(\hat{\theta}_i^{mp}, \Sigma_*^{mp}, I_K)$.

Algorithm 11 Meta-Personalized Dynamic Pricing Algorithm

Input: The prior covariance matrix Σ_*^{mp}, the total number of epochs N, the length of each epoch T, the subgaussian parameter σ, and the set of feasible prices $[p_{\min}, 1]$.
Initialization: N_0 as defined in eq. (C.38).
for each epoch $i = 1, \ldots, N$ **do**
    if $i \leq N_0$ **then**
        Run TS$(\mathscr{MN}(0, \Psi^{mp} \cdot I_{(K+1)d}, I_K), \lambda_e^{mp})$.
    else
        Update $\hat{\theta}_i^{mp}$ according to eq. (C.39), and run TS$\left(\mathscr{MN}\left(\hat{\theta}_i^{mp}, \Sigma_*^{mp}, I_K\right), \lambda_e^{mp}\right)$.
    end if
end for

We now translate our previous upper bound on the meta regret of the single-product `Meta-DP` algorithm to the multi-product setting.

Corollary 71 (Multi-Product `Meta-DP`)**.** *The meta regret of multi-product `Meta-DP` satisfies*

$$\mathscr{R}_{N,T}(\texttt{Meta-DP algorithm}) = \tilde{O}\left(K^4 d^2 (NT)^{\frac{1}{2}}\right).$$

Corollary 71 is again an immediate consequence of Theorem 28. Again, this is because, if a matrix X follows the matrix Gaussian distribution $\mathscr{MN}(A, B, C)$, then vec(X), (*i.e.*, the vectorized version of X that concatenates each column of X to form a vector), follows the multivariate Gaussian distribution $\mathscr{N}(A, C \otimes B)$ [95]. In other words, we can map the multi-product prior $\mathscr{MN}\left(\theta_*^{mp}, \Sigma_*^{mp}, I_K\right)$ to the same form as a single-product prior $\mathscr{N}\left(\theta_*^{mp}, I_k \otimes \Sigma_*^{mp}\right)$, by taking the unknown prior mean to be the vectorized vec(θ_*^{mp}) and

the prior covariance to be $\begin{pmatrix} 1 & p_{i,t,1}^{mp} & \dots & p_{i,t,K}^{mp} \end{pmatrix}^\top \otimes x_{i,t} \otimes \mathbf{1}_K$ ($\mathbf{1}_K$ is the $K \times 1$ column vector with all entries equal to 1). Thus, since we still maintain a linear demand model, the only mathematical change is that the unknown parameter has dimension $(K^2+K)d$ instead of $2d$. Thus, the same result applies by replacing the d in Theorem 28 with $(K^2+K)d$.

C.6.3 Multi-Product `Meta-DP++` algorithm

The multi-product `Meta-DP++` algorithm is presented in Algorithm 12. We first define some additional notation, and then describe the algorithm in detail.

Algorithm Description: The first N_1^{mp} epochs are treated as exploration epochs, where we define

$$N_1^{mp} = \max\{N_0,\ 32(c_3^{mp})^2 d^3 (K^2+K)^3 \mathscr{T}_e^2 \log_e^3(2d(K^2+K)N^2T), \\ (c_4^{mp})^2 d^4 (K^2+K)^4 T^2 \log_e^3(2N^2T)\}, \tag{C.40}$$

and the constants are defined as

$$c_3^{mp} = \frac{16\sqrt{\sigma^2(\lambda_e^{mp})^{-1} + 5\overline{\lambda}^{mp}}}{\sigma\lambda_e^{mp}\underline{\lambda}^{mp}} + \frac{256(\overline{\lambda}^{mp}(\lambda_e^{mp})^2 + 16\sigma^2)}{(\lambda_e^{mp}\underline{\lambda}^{mp})^2}\left(\frac{8\sqrt{2}x_{\max}}{\lambda_e^{mp}} + \frac{S^{mp}}{\sigma\lambda_e^{mp}}\right),$$
$$c_4 = \frac{10^4\sigma(\overline{\lambda}^{mp}(\lambda_e^{mp})^2 + 16\sigma^2)}{(\lambda_e^{mp}\underline{\lambda}^{mp})^2}.$$

As before, the `Meta-DP++` algorithm proceeds differently for earlier exploration epochs and later epochs:

1. **Epoch $\mathbf{i \le N_1^{mp}}$:** the `Meta-DP++` algorithm runs the prior-independent Thompson sampling algorithm $\mathrm{TS}(\mathscr{MN}(0, \Psi^{mp} \cdot I_{(K+1)d}, I_K), \lambda_e)$, where

$$\Psi^{mp} = \sigma\sqrt{2d\log_e(T(1+2x_{\max}^2 T))} + \sqrt{20\overline{\lambda}^{mp} d\log_e(2T)}.$$

2. **Epoch $\mathbf{i > N_1^{mp}}$:** the `Meta-DP++` algorithm computes an estimator $\hat{\theta}_i^{mp}$ of the prior mean θ_*^{mp} using Eq. (C.39) (same as the multi-product `Meta-DP` algorithm), and an

estimator $\hat{\Sigma}_i^{mp}$ of the prior covariance Σ_*^{mp} as

$$\hat{\Sigma}_i^{mp} = \frac{1}{i-2}\sum_{j=1}^{i-1}\left(\dot{\theta}_j^{mp} - \frac{\sum_{k=1}^{i-1}\dot{\theta}_k^{mp}}{i-1}\right)\left(\dot{\theta}_j^{mp} - \frac{\sum_{k=1}^{i-1}\dot{\theta}_k^{mp}}{i-1}\right)^\top - \frac{\sigma^2\sum_{j=1}^{i-1}\mathbb{E}\left[\left(V_{j,\mathscr{T}_j}^{mp}\right)^{-1}\right]}{i-1}, \tag{C.41}$$

where, following the single-product `Meta-DP++` algorithm, we define

$$\dot{\theta}_i^{mp} = \left(V_{i,\mathscr{T}_i}^{mp}\right)^{-1}\left(\sum_{t=1}^{\mathscr{T}_i} D_{i,t}^{mp}(p_{i,t}^{mp}, x_{i,t})m_{i,t}^{mp}\right).$$

The widened posterior covariance is thus

$$\hat{\Sigma}_i^{mp,w} = \hat{\Sigma}_i + \frac{128(\overline{\lambda}^{mp}(\lambda_e^{mp})^2 + 8\sigma^2 dK(K+1))}{(\lambda_e^{mp})^2}\sqrt{\frac{5dK(K+1)\log_e(2N^2T)}{i}}\cdot I_{K(K+1)d}, \tag{C.42}$$

where $I_{K(K+1)d}$ is the $(K(K+1)d)$-dimensional identity matrix.

Then, the `Meta-DP++` algorithm runs the Thompson Sampling algorithm (see Algorithm 10) with the estimated prior $\mathscr{MN}\left(\hat{\theta}_i^{mp}, \hat{\Sigma}_i^{mp,w}, I_K\right)$.

Algorithm 12 Meta-Dynamic Pricing++ Algorithm

Input: The total number of products N, the length of each epoch T, the noise parameter σ, and the set of feasible prices $[p_{\min}, 1]$.
for epoch $i = 1, \ldots, N$ **do**
  if $i \le N_1^{mp}$ **then**
    Run TS$(\mathscr{MN}(0, \Psi^{mp}\cdot I_{(K+1)d}, I_K), \lambda_e^{mp})$.
  else
    Update $\hat{\theta}_i^{mp}$ and $\hat{\Sigma}_i^{mp}$ according to Eqs. (C.39) and (C.41) respectively.
    Compute widened prior mean estimate $\hat{\Sigma}_i^{mp,w}$ according to Eq. (C.42).
    Run TS$\left(\mathscr{MN}\left(\hat{\theta}_i^{mp}, \hat{\Sigma}_i^{mp,w}, I_K\right), \lambda_e^{mp}\right)$.
  end if
end for

We now translate our previous upper bound on the meta regret of the single-product `Meta-DP++` algorithm to the multi-product setting.

Corollary 72 (Multi-Product `Meta-DP++`)**.** *The meta regret of multi-product* `Meta-DP++` *satisfies*

$$\mathscr{R}_{N,T}(\texttt{Meta-DP++ algorithm}) = \tilde{O}\left(\min\left\{K^4 d^2 N T^{\frac{1}{2}},\ K^8 d^4 N^{\frac{1}{2}} T^{\frac{3}{2}}\right\}\right) = \tilde{O}\left(K^6 d^3 (NT)^{\frac{5}{6}}\right).$$

Corollary 72 is again an immediate consequence of Theorem 31. The reasoning is exactly the same as for Corollary 71, so we omit it. Essentially, we can map the multi-product prior to the same form as a single-product prior, so that the only mathematical change is that the unknown parameter has dimension $(K^2+K)d$ instead of $2d$. Thus, the same result applies by replacing the d in Theorem 31 with $(K^2+K)d$.

C.7 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 73 (Proposition 3 of [162])**.** *Fix positive constants σ, c, and c'. Denote the set of all possible parameters as $\Theta \in \mathscr{R}^d$, the mean reward function as $f_\theta(a) = \langle \phi(a), \theta \rangle$ for some $\phi : \mathscr{A} \to \mathscr{R}$, $\sup_{\rho\in\Theta} \|\rho\| \le c$, and $\sup_{a\in\mathscr{A}} \|\phi(a)\| \le c'$, 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 characterizes the eigenvalues of a matrix Kronecker product.

Lemma 74 (Corollary 13.11 of [126])**.** *Let A be a real-valued matrix with singular values $\lambda_1 \ge \ldots \ge \lambda_r > 0$, and let B be a real-valued matrix with singular values $\lambda'_1 \ge \ldots \ge \lambda'_s > 0$, then $A \otimes B$ has $r \cdot s$ singular values $\lambda_i \lambda'_j$ $(i \in [r]\ j \in [s])$.*

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

Lemma 75 ([171])**.** *For the d-dimensional unit ball, its δ covering number is upper bounded by $d \log_e(1+2/\delta)$.*

The following lemma provides an upper bound for the quantity $\exp(1/a)$ when $a > 1$.

Lemma 76. *For any number $a \in [0,1]$, $\exp(a) \leq 1 + 2a$.*

Proof. Proof of Lemma 76. We note that the function $f(a) = \exp(a) - 1 - 2a$ is a convex function as

$$f''(a) = e^a > 0, \tag{C.43}$$

as well as that $f(0) = 1 - 1 = 0$ and $f(1) = e - 3 < 0$, so $f(a) \leq 0$ for all $a \in [0,1]$. □

The following lemma makes a connection between the tail probability of a random variable and its moment generating function.

Lemma 77 (Lemma 1.5 of [154])**.** *For a random variable $X \in \mathscr{R}$ such that $\mathbb{E}[X] = 0$ and for any $u > 0$,*

$$\Pr(|X| > u) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right),$$

we have for any $v \in \mathscr{R}$,

$$\mathbb{E}[\exp(vX)] \leq \exp(4v^2\sigma^2).$$

The following lemma provides a concentration inequality for estimating the empirical covariance matrix.

Lemma 78 (Theorem 7.1 of [155] and Theorem 6.5 of [171])**.** *Let $X_1, \ldots, X_n$ be n i.i.d. copies of the random vector X such that $\mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma$, and X is σ-subgaussian vector. Then, the operator norm of the difference between the empirical covariance $\sum_{i=1}^n X_i X_i^\top / n$ and Σ satisfies*

$$\Pr\left(\left\|\frac{\sum_{i=1}^n X_i X_i^\top}{n} - \Sigma\right\|_{op} \leq 32\sigma^2\left(\sqrt{\frac{5d + 2\log_e(2/\delta)}{n}} \vee \frac{5d + 2\log_e(2/\delta)}{n}\right)\right) \geq 1 - \delta$$

for any $\delta \in [0,1]$.

The following lemma shows that the operator norm of the product of two matrices is upper bounded by the product of the operator norms of those matrices.

Lemma 79. *For two positive semi-definite matrices A and B, we have*

$$\|AB\|_{op} \leq \|A\|_{op}\|B\|_{op}.$$

Proof. The statement can be easily concluded as follows.

$$\begin{aligned}
\|AB\|_{op} = \max_{x:\|x\|=1} \|ABx\| &= \max_{x:\|x\|=1} \frac{\|ABx\|}{\|Bx\|}\|Bx\| \\
&\leq \max_{x:\|x\|=1} \frac{\|ABx\|}{\|Bx\|} \max_{y:\|y\|=1} \|By\| \\
&= \max_{Bx:\|x\|=1} \frac{\|ABx/\|Bx\|\|}{\|Bx/\|Bx\|\|} \max_{y:\|y\|=1} \|By\| \\
&= \|A\|_{op}\|B\|_{op}.
\end{aligned}$$

□

The following lemma compares the determinants of two positive semi-definite matrices.

Lemma 80. *For two symmetric positive semi-definite matrices A and B, if $A-B$ is positive semi-definite, then* $\det(A) \geq \det(B)$.

Proof. Note that

$$\begin{aligned}
\det(A) = \det(B+(A-B)) &= \det\left(B^{\frac{1}{2}}\left(I+B^{-\frac{1}{2}}(A-B)B^{-\frac{1}{2}}\right)B^{\frac{1}{2}}\right) \\
&= \det(B)\det\left(\left(I+B^{-\frac{1}{2}}(A-B)B^{-\frac{1}{2}}\right)\right) \\
&\geq \det(B)\left(1+\det\left(B^{-\frac{1}{2}}(A-B)B^{-\frac{1}{2}}\right)\right) \qquad \text{(C.44)} \\
&= \det(B)+\det(A-B) \\
&\geq \det(B). \qquad \text{(C.45)}
\end{aligned}$$

Here, inequality (C.44) holds because $\prod_{k=1}^{2d}(1+\mu_k) \geq 1+\prod_{k=1}^{2d}\mu_k$ where μ_k is the k^{th} eigenvalue of $B^{-\frac{1}{2}}(A-B)B^{-\frac{1}{2}}$, and inequality (C.45) holds because $A-B$ is positive semi-definite. □

Appendix D

Proofs for Chapter 5

D.1 Proof of Proposition 33

We let $x_1 = 1, x_2 = T, x_t = t-1 \ \forall t \in [3,T]$, and

$$y_t = \begin{cases} 1 & \text{if } x_t \text{ is odd and } x_t < T; \\ 0 & \text{if } x_t \text{ is even and } x_t < T; \\ \frac{1}{2} & \text{if } x_t = T. \end{cases}$$

Then, we can see that the oracle can employ a constant function $f(x_t) = 1/2$ and achieve a cumulative loss at most

$$\inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq \sum_{t=1}^{T} \ell\left(\frac{1}{2}, y_t\right) = \sum_{t=1}^{T-1} \left(\frac{1}{2}\right)^2 = \frac{T-1}{4}.$$

According to [121], the solution to the least square problem 5.6 can be computed by the Pool Adjacent Violators Algorithm (PAVA) [24]. The algorithm is based on the observation that if the labels of any two consecutive labels y_i, y_{i+1} violate isotonicity (*i.e.*, $x_i \leq x_{i+1}$ but $y_i \geq y_{i+1}$), then we must have $\hat{f}(x_i) = \hat{f}(x_{i+1})$ in the optimal solution of the least square and we may merge both points to their average. This process repeats and terminates until every historical data is passed. For every time step t, there are two important properties of the fitted function $\hat{f}_t$ [156, 121]:

1. The function $\hat{f}_t(\cdot)$ is piecewise constant and thus its level sets partition $\{1,\ldots,T\}$.

2. The value of $\hat{f}_t(\cdot)$ on any level set is equal to the weighted average of labels within that set.

Now, we show that even if the ILS policy knows $f(x_1) = f(T) = 1/2 + 1/T$ and $f(x_2) = f(1) = 1$ (*i.e.*, the ILS policy does not need to perform extrapolation) would have to incur a cumulative loss of $13(T-2)/32$. To see this, we distinguish two cases for every $t \geq 2$:

- **Case 1.** x_t is odd: In this case, one can easily verify $\hat{f}_t(x) = 1/2$ and hence,

$$\ell(\hat{f}_t(x_t), y_t) = \left(\frac{1}{2} - 1\right)^2 = \frac{1}{4}.$$

- **Case 2.** x_t is even: In this case, one can easily verify

$$\hat{f}_t(x) = \begin{cases} \frac{1}{2} & \text{if } x \leq x_{t-1}; \\ \frac{3}{4} & \text{oterhwise.} \end{cases}$$

  Hence,

$$\ell(\hat{f}_t(x_t), y_t) = \left(\frac{3}{4} - 0\right)^2 = \frac{9}{16}.$$

Summing up the two cases

$$\sum_{t=1}^{T} \ell(\hat{f}_t(x_t), y_t) \geq \left(\frac{1}{4} + \frac{9}{16}\right) \frac{T-2}{2} = \frac{13(T-2)}{32} \geq \frac{12(T-1)}{32},$$

where we use the assumption $T \geq 14$. The statement thus follows.

D.2 Relaxation and Admissibility

In [152], a rate-optimal algorithmic recipe for the general problem of online non-parametric regression is proposed based on the relaxation framework introduced in [151]. The relaxation framework follows a backward induction approach to characterized the minimax-optimal regret bounds for general online learning problems. Although the induced algo-

rithm is computationally inefficient in general, we will adopt the framework to characterize the minimax-optimal regret upper bound, and design a computationally-efficient algorithm in Section 5.2 for our setting by exploiting the special properties of random-design online isotonic regression.

Following the zero-sum sequential game formulation for online learning (see, *e.g.*, Section 7.3 of [53]), the minimax-optimal regret of our setting can be written as (we recall that $\{x_t\}_{t=1}^T \sim \mathscr{D}$, $\hat{y}_t \sim q_t \in \Delta([0,1])$, and $y_t \sim p_t \in \Delta[0,1]$ for all time step $t \in [T]$, but for brevity, we often do not write these out explicitly)

$$\mathscr{R}_T := \inf \mathscr{R}_T(\pi|\mathscr{F}) = \mathop{\mathbb{E}}_{x_1} \inf_{q_1} \sup_{p_1} \mathop{\mathbb{E}\mathbb{E}}_{\hat{y}_1 y_1} \cdots \mathop{\mathbb{E}}_{x_T} \inf_{q_T} \sup_{p_T} \mathop{\mathbb{E}\mathbb{E}}_{\hat{y}_T y_T} \left[\sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^T \ell(f(x_t), y_t) \right], \tag{D.1}$$

where the DM picks the q_t's to minimize the terminating loss, *i.e.*,

$$\sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^T \ell(f(x_t), y_t)$$

, while an adversary picks the p_t's to maximize the terminating loss, which is also her gain.

Remark 32. *The relaxation framework is not specifically tied to any particular online non-parametric regression setting, and $\mathscr{F}$ can be replaced by any other non-parametric function class.*

Formally, a relaxation $\mathbf{Rel} = \{\mathbf{Rel}_t\}_{t=0}^T$ is a sequence of mappings from the history information and the covariate distributions to real numbers. Specifically, a relaxation $\mathbf{Rel}$ is called *admissible* if the following conditions (in eq. (D.2) and eq. (D.3)) are satisfied, *i.e.*,

$$\mathbf{Rel}_T(\mathscr{H}_T) \geq - \inf_{f \in \mathscr{F}} \sum_{t=1}^T \ell(f(x_t), y_t), \tag{D.2}$$

and for every $t \in [T]$, there exists a distribution $\tilde{q}_t \in \Delta([0,1])$ (recall that for a set $\mathscr{A} \in \mathbb{R}$,

$\Delta(\mathscr{A})$ is the simplex over $\mathscr{A}$), such that

$$\mathop{\mathbb{E}}_{x_t} \sup_{\tilde{p}_t \in \Delta([0,1])} \mathop{\mathbb{E}}_{\hat{y}_t \sim \tilde{q}_t} \mathop{\mathbb{E}}_{y_t \sim \tilde{p}_t} \left[\ell\left(\hat{y}_t, y_t\right) + \mathbf{Rel}_t\left(\mathscr{H}_t\right)\right] \leq \mathbf{Rel}_{t-1}\left(\mathscr{H}_{t-1}\right). \tag{D.3}$$

Here, we do not explicitly include the dependence on $\mathscr{D}$ for brevity. It can be readily shown that an admissible relaxation-policy pair can lead to the following regret bound guarantee.

Lemma 81. *If **Rel** is admissible, then $\mathscr{R}_T \leq \mathbf{Rel}_0(\mathscr{H}_0)$.*

The proof of Lemma 81 is similar to the proof of Proposition 1 in [151]. For completeness, we include it in Section D.2.1.

Suppose there exists a relaxation **Rel** that is admissible for a certain online non-parametric regression problem, [152] proposed a natural way to derive π to attain the regret upper bound specified in Proposition 81: For each time step t, after the covariate x_t is revealed, the DM predicts by sampling $\hat{y}_t$ according to the distribution q_t defined as follows:

$$q_t := \mathop{\arg\min}_{q_t' \in \Delta[0,1]} \sup_{p_t \in \Delta([0,1])} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t'} \mathop{\mathbb{E}}_{y_t \sim p_t} \left[\ell\left(\hat{y}_t, y_t\right) + \mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]. \tag{D.4}$$

Remark 33. *Although it can be easily seen that this specific choice of policy is can lead to the regret upper bound shown in Lemma 81, it is often computationally hard to solve the optimization problem (D.4) as pointed out by [152]. In particular, it is unclear how to adopt this framework to the online isotonic regression problem even under the fixed-design setting (see Section 1.1 of [121]).*

One useful corollary for the relaxation result is that if the induced probability of a policy π satisfies eq. (D.2) and (D.3), then the regret of π is upper bounded by $\mathbf{Rel}_0(\mathscr{H}_0)$.

Corollary 82. *Following the notations in Section 3.1, let q_t be the induced distribution of π_t, and **Rel** be any admissible relaxation, if it satisfies*

$$\mathbf{Rel}_T\left(\mathscr{H}_T\right) \geq -\inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t),$$

and for every $t \in [T]$,

$$\mathop{\mathbb{E}}_{x_t} \sup_{\tilde{p}_t \in \Delta([0,1])} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \mathop{\mathbb{E}}_{y_t \sim \tilde{p}_t} \left[\ell\left(\hat{y}_t, y_t\right) + \boldsymbol{Rel}_t\left(\mathscr{H}_t\right) \right] \leq \boldsymbol{Rel}_{t-1}\left(\mathscr{H}_{t-1}\right).$$

then the regret of $\pi = \{\pi_t\}_{t=1}^T$ *is upper bounded by* $\boldsymbol{Rel}_0(\mathscr{H}_0)$.

The proof of 82 is very similar to that of Lemma 81, and it is thus omitted.

D.2.1 Proof of Lemma 81

By definition, the regret of π can be written as (we recall that $x_t \sim \mathscr{D}_t$, $\hat{y}_t \sim \tilde{q}_t \in \Delta([0,1])$, and $y_t \sim p_t \in \Delta[0,1]$ for all time step $t \in [T]$)

$$\begin{aligned}
\mathscr{R}_T =& \mathop{\mathbb{E}}_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{\hat{y}_1 y_1} \mathbb{E} \ldots \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \mathbb{E} \left[\sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^T \ell(f(x_t), y_t) \right] \\
\leq& \mathop{\mathbb{E}}_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{\hat{y}_1 y_1} \mathbb{E} \ldots \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \mathbb{E} \left[\sum_{t=1}^T \ell(\hat{y}_t, y_t) + \mathbf{Rel}_T(\mathscr{H}_T) \right] \qquad \text{(D.5)} \\
=& \mathop{\mathbb{E}}_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{\hat{y}_1 y_1} \mathbb{E} \ldots \mathop{\mathbb{E}}_{x_{T-1}} \sup_{p_{T-1}} \mathop{\mathbb{E}}_{\hat{y}_{T-1} y_{T-1}} \left\{ \sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \mathbb{E} \left[\ell(\hat{y}_t, y_t) + \mathbf{Rel}_T(\mathscr{H}_T) \right] \right\},
\end{aligned}$$

where we have used the admissibility condition (D.2) in inequality (D.5). Continue to apply the admissibility condition (D.3), we can further upper bound the above as

$$\mathscr{R}_T \leq \mathop{\mathbb{E}}_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{\hat{y}_1 y_1} \mathbb{E} \ldots \mathop{\mathbb{E}}_{x_{T-1}} \sup_{p_{T-1}} \mathop{\mathbb{E}}_{\hat{y}_{T-1} y_{T-1}} \left[\sum_{t=1}^{T-1} \ell(\hat{y}_t, y_t) + \mathbf{Rel}_{T-1}(\mathscr{H}_{T-1}) \right].$$

Recursively applying admissibility condition (D.3) for $t = T-1, \ldots, 1$, we have

$$\mathscr{R}_T \leq \mathop{\mathbb{E}}_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{\hat{y}_1 y_1} \mathbb{E} \left[\ell(\hat{y}_1, y_1) + \mathbf{Rel}_1(\mathscr{H}_1) \right] \leq \mathbf{Rel}_0(\mathscr{H}_0).$$

D.3 Proof of Theorem 34

Following the discussion in Section 5.2.2, we know that the difficulty in analyzing the regret bound of the `SEW` policy lies in its design, *i.e.*, it utilizes simulated future covariates to make

predictions. It turns out that the relaxation framework also leverages a backward induction principle to establish regret bounds for online non-parametric regression problems [152]. We thus exploit this approach throughout the proof.

D.3.1 Arriving at the Relaxation

In this section, we first derive the relaxation (*i.e.*, the potential functions V_t's will be the $\mathbf{Rel}_t(\mathscr{H}_t)$'s then) for our setting. Following the discussion in Section 5.1.3, it is sufficient to show a regret upper bound against the data-dependent discrete offline oracle $\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\sum_{t=1}^T \ell(f(x_t),y_t)$. We thus follow eq. (D.2) and (D.3) to work in a backward manner, and begin by upper bounding the term $-\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\sum_{t=1}^T \ell(f(x_t),y_t)$. Note that for any positive real number λ (>0),

$$\begin{aligned}-\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\sum_{t=1}^T \ell(f(x_t),y_t) =& \sup_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)} -\sum_{t=1}^T (f(x_t)-y_t)^2 \\ =&\lambda^{-1}\sup_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)} -\lambda\sum_{t=1}^T (f(x_t)-y_t)^2.\end{aligned}$$

From the fact that $\forall x\in\mathbb{R}$, $x=\log_e(\exp(x))$, we have

$$\begin{aligned}-\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\sum_{t=1}^T \ell(f(x_t),y_t) =&\lambda^{-1}\log_e\left[\sup_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\exp\left(-\lambda\sum_{t=1}^T (f(x_t)-y_t)^2\right)\right] \\ \leq&\lambda^{-1}\log_e\left[\sum_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\exp\left(-\lambda\sum_{t=1}^T (f(x_t)-y_t)^2\right)\right],\end{aligned}$$

where we have used the non-negativity of the exponential function to replace the supremum by the summation. We can thus define

$$\mathbf{Rel}_T(\mathscr{H}_T)=\lambda^{-1}\log_e\left[\sum_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\exp\left(-\lambda\sum_{t=1}^T (f(x_t)-y_t)^2\right)\right] \tag{D.6}$$

with λ (>0) to be specified in the forthcoming Lemma 83. It is evident that this choice satisfies inequality (D.2) by definition. We then compute an upper bound for the quantity

$\mathop{\mathbb{E}}_{x_T}\mathrm{infsup}\mathop{\mathbb{E}}_{q_T}\mathop{\mathbb{E}}_{p_T}\mathop{\mathbb{E}}_{\hat{y}_T y_T}[\ell(\hat{y}_T, y_T) + \mathbf{Rel}_T(\mathscr{H}_T)]$ to guide our design of the relaxation.

Lemma 83. *Setting $\lambda = 1/2$ in eq. (D.6), we have*

$$\mathop{\mathbb{E}}_{x_T}\mathrm{infsup}\mathop{\mathbb{E}}_{q_T}\mathop{\mathbb{E}}_{p_T}\mathop{\mathbb{E}}_{\hat{y}_T y_T}[\ell(\hat{y}_T, y_T) + \boldsymbol{Rel}_T(\mathscr{H}_T)] \le 2\mathop{\mathbb{E}}_{x_T}\log_e\left\{\sum_{f\in\mathscr{F}(\{x_t\}_{t=1}^T)}\exp\left[-\frac{\sum_{t=1}^{T-1}(f(x_t)-y_t)^2}{2}\right]\right\}.$$

The proof of Lemma 83 is provided in Section D.3.3 of the appendix. Lemma 83 motivates us to define the following relaxation for, *i.e.*,

$$\forall t\in[T] \qquad \mathbf{Rel}_t(\mathscr{H}_t) = 2\mathop{\mathbb{E}}_{x_{t+1}}\ldots\mathop{\mathbb{E}}_{x_T}\log_e\left\{\sum_{f\in\mathscr{F}(\{x_t\}_{t=1}^T)}\exp\left[-\frac{\sum_{s=1}^{t}(f(x_s)-y_s)^2}{2}\right]\right\}. \tag{D.7}$$

We now show that this choice of relaxation is admissible.

Lemma 84. *The relaxation (D.7) is admissible.*

The proof of Lemma 84 is provided in Section D.3.4 of the appendix. To this end, we have come up with one possible relaxation for our problem. In what follows, we shall see how to utilize the relaxation to show that the regret of the `SEW` policy is indeed $O(T^{1/3})$.

D.3.2 Completing the Proof

In this section, we verify that the `SEW` policy and the relaxation defined in eq. (D.7) satisfies the precondition of Corollary 82, and the regret of the `SEW` policy is thus upper bounded by $\mathbf{Rel}_0(H_0) = O(T^{1/3})$.

We first write out the output $\hat{y}_t$ by the `SEW` policy at each time step $t\in[T]$ explicitly. For each time step t, conditioned on the sampled future covariates $x'_{t+1},\ldots,x'_T$, we define a distribution $q_t \in \Delta\left(\mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x_s\}_{s=t+1}^T\right)\right) : \forall f\in\mathscr{F}\left(\{x_s\}_{s=1}^t\cup\{x'_s\}_{s=t+1}^T\right)$

$$\Pr(f(x_t)) = \frac{\exp\left(-\sum_{s=1}^{t-1}(f(x_s)-y_s)^2/2\right)}{\sum_{f'\in\mathscr{F}\left(\{x_s\}_{s=1}^t\cup\{x'_j\}_{j=t+1}^T\right)}\exp\left(-\sum_{s=1}^{t-1}(f'(x_s)-y_s)^2/2\right)}. \tag{D.8}$$

We can thus express $\hat{y}_t$ output by the SEW policy as $\hat{y}_t = \mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)]$. We now verify that the SEW policy and the relaxation defined in (D.7) satisfy the admissibility condition in (D.3). Note that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{\hat{y}_t\sim\pi_t}\mathop{\mathbb{E}}_{y_t\sim p_t}\left[\ell\left(\hat{y}_t,y_t\right)+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]\\
=&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{x'_{t+1}}\ldots\mathop{\mathbb{E}}_{x'_T y_t\sim p_t}\left[\ell\left(\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)],y_t\right)+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]\\
=&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{x'_{t+1}}\ldots\mathop{\mathbb{E}}_{x'_T y_t\sim p_t}\left[\left(\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)]-y_t\right)^2+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]\\
=&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{x'_{t+1}}\ldots\mathop{\mathbb{E}}_{x'_T y_t\sim p_t}\left[2\log_e\exp\left(\frac{\left(\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)]-y_t\right)^2}{2}\right)+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right].
\end{aligned}\tag{D.9}$$

To proceed, we show the following lemma.

Lemma 85. *Conditioned on $x'_{t+1},\ldots,x'_T$, we have*

$$\exp\left(\frac{\left(\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)]-y_t\right)^2}{2}\right)\le\frac{\sum_{f'\in\mathscr{F}\left(\{x_s\}_{s=1}^t\cup\{x'_s\}_{s=t+1}^T\right)}\exp\left(-\sum_{s=1}^{t-1}(f'(x_s)-y_s)^2/2\right)}{\sum_{f\in\mathscr{F}\left(\{x_s\}_{t=1}^t\cup\{x'_s\}_{s=t+1}^T\right)}\exp\left(-\sum_{s=1}^{t}(f(x_s)-y_s)^2/2\right)}\tag{D.10}$$

The proof of Lemma 85 is provided in Section D.3.5. Applying Lemma 85 to the RHS of (D.9), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{\hat{y}_t\sim\pi_t}\mathop{\mathbb{E}}_{y_t\sim p_t}\left[\ell\left(\hat{y}_t,y_t\right)+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]\\
\le&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{x'_{t+1}}\ldots\mathop{\mathbb{E}}_{x'_T y_t\sim p_t}\left[2\log_e\frac{\sum_{f'\in\mathscr{F}\left(\{x_s\}_{s=1}^t\cup\{x'_s\}_{s=t+1}^T\right)}\exp\left(-\sum_{s=1}^{t-1}(f'(x_s)-y_s)^2/2\right)}{\sum_{f\in\mathscr{F}\left(\{x_s\}_{t=1}^t\cup\{x'_s\}_{s=t+1}^T\right)}\exp\left(-\sum_{s=1}^{t}(f(x_s)-y_s)^2/2\right)}+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right].
\end{aligned}\tag{D.11}$$

Further note that $x'_{t+1},\ldots,x_T$ are i.i.d. copies of $x_{t+1},\ldots,x_T$, the RHS of (D.11) can be

rewritten as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{\hat{y}_t\sim\pi_t}\mathop{\mathbb{E}}_{y_t\sim p_t}\left[\ell\left(\hat{y}_t,y_t\right)+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]\\
\leq&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{x_{t+1}}\ldots\mathop{\mathbb{E}}_{x_T y_t\sim p_t}\left[2\log_e\frac{\sum_{f'\in\mathscr{F}\left(\{x_s\}_{s=1}^T\right)}\exp\left(-\sum_{s=1}^{t-1}(f'(x_s)-y_s)^2/2\right)}{\sum_{f\in\mathscr{F}\left(\{x_s\}_{s=1}^T\right)}\exp\left(-\sum_{s=1}^{t}(f(x_s)-y_s)^2/2\right)}+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right] \qquad \text{(D.12)}\\
=&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{x_{t+1}}\ldots\mathop{\mathbb{E}}_{x_T y_t\sim p_t}\left\{2\log_e\left[\sum_{f'\in\mathscr{F}\left(\{x_s\}_{s=1}^T\right)}\exp\left(-\frac{\sum_{s=1}^{t-1}(f'(x_s)-y_s)^2}{2}\right)\right]\right\}, \qquad \text{(D.13)}
\end{aligned}$$

where we have recalled the definition of $\mathbf{Rel}_t(\mathscr{H}_t)$ as

$$\mathbf{Rel}_t(\mathscr{H}_t)=2\mathop{\mathbb{E}}_{x_{t+1}}\ldots\mathop{\mathbb{E}}_{x_T}\log_e\left\{\sum_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\exp\left[-\frac{\sum_{s=1}^{t}(f(x_s)-y_s)^2}{2}\right]\right\}. \qquad \text{(D.14)}$$

in inequality (D.12). Continue with (D.13), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t}\sup_{p_t}\mathop{\mathbb{E}}_{\hat{y}_t\sim\pi_t}\mathop{\mathbb{E}}_{y_t\sim p_t}\left[\ell\left(\hat{y}_t,y_t\right)+\mathbf{Rel}_t\left(\mathscr{H}_t\right)\right]\\
\leq&\mathop{\mathbb{E}}_{x_t}\ldots\mathop{\mathbb{E}}_{x_T}\left\{2\log_e\left[\sum_{f\in\mathscr{F}\left(\{x_s\}_{s=1}^T\right)}\exp\left(-\frac{\sum_{s=1}^{t-1}(f(x_s)-y_s)^2}{2}\right)\right]\right\}\\
=&\mathbf{Rel}_{t-1}(\mathscr{H}_{t-1}).
\end{aligned}$$

Therefore, we have established that the relaxation defined in eq. (D.7) and the policy induced by the `SEW` policy algorithm satisfy eq. (D.3). Together with Corollary 82, we can derive the following regret upper bound for the `SEW` policy against the oracle

$$\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^T\right)}\sum_{t=1}^{T}\ell(f(x_t),y_t).$$

Lemma 86. *The regret of the `SEW` policy against the discrete isotonic function class* $\mathscr{F}\left(\{x_t\}_{t=1}^T\right)$

is upper bounded as

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\hat{y}_t(\pi_t),y_t)-\inf_{f\in\mathscr{F}\left(\{x_t\}_{t=1}^{T}\right)}\sum_{t=1}^{T}\ell(f(x_t),y_t)\right]\leq \textbf{\textit{Rel}}_0(\mathscr{H}_0)\leq 2K\log_e(T+1).$$

The proof of Lemma 86 is provided in Section D.3.6 of the appendix. By Lemma 86 and inequality (5.5), we know that the regret of the SEW policy against the isotonic function class $\mathscr{F}$ is upper bounded as

$$\mathscr{R}_T=\mathbb{E}\left[\sum_{t=1}^{T}\ell(\hat{y}_t(\pi_t),y_t)-\inf_{f\in\mathscr{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t)\right]\leq 2K\log_e(T+1)+T/4K^2.$$

By setting $K=\left\lceil T^{1/3}/[4(\log_e(T+1))^{1/3}]\right\rceil$, this is of order $\mathscr{R}_T=\tilde{O}(T^{1/3})$.

D.3.3 Proof of Lemma 83

Recall that for a set $\mathscr{A}\subseteq\mathbb{R}$, $\Delta(\mathscr{A})$ is the simplex over $\mathscr{A}$, by the minimax theorem [141], we have

$$\mathop{\mathbb{E}}_{x_T}\inf_{q_T}\sup_{p_T}\mathop{\mathbb{E}}_{\hat{y}_T y_T}\left[\ell\left(\hat{y}_T,y_T\right)+\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right]=\mathop{\mathbb{E}}_{x_T}\sup_{p_T}\inf_{q_T}\mathop{\mathbb{E}}_{\hat{y}_T y_T}\left[\ell\left(\hat{y}_T,y_T\right)+\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right].$$

Since the RHS is convex in $\hat{y}_T$, the infimum over $q_T\in\Delta([0,1])$ is attained at a point mass, and we can further rewrite

$$\begin{aligned}\mathop{\mathbb{E}}_{x_T}\inf_{q_T}\sup_{p_T}\mathop{\mathbb{E}}_{\hat{y}_T y_T}\left[\ell\left(\hat{y}_T,y_T\right)+\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right]&=\mathop{\mathbb{E}}_{x_T}\sup_{p_T}\inf_{\hat{y}_T}\mathop{\mathbb{E}}_{y_T}\left[\ell\left(\hat{y}_T,y_T\right)+\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right]\\&=\mathop{\mathbb{E}}_{x_T}\sup_{p_T}\inf_{\hat{y}_T}\mathop{\mathbb{E}}_{y_T}\left[\left(\hat{y}_T-y_T\right)^2+\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right]\\&=\mathop{\mathbb{E}}_{x_T}\sup_{p_T}\left[\inf_{\hat{y}_T}\mathop{\mathbb{E}}_{y_T}\left(\hat{y}_T-y_T\right)^2+\mathop{\mathbb{E}}_{y_T}\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right].\end{aligned}$$

Note that the minimum of $\mathop{\mathbb{E}}_{y_T}\left(\hat{y}_T-y_T\right)^2$ is attained at $\hat{y}_T=\mathbb{E}[y_T]$, and by definition of $\mathbf{Rel}_T(\mathscr{H}_T)$ in eq. (D.6),

$$\mathop{\mathbb{E}}_{x_T}\inf_{q_T}\sup_{p_T}\mathop{\mathbb{E}}_{\hat{y}_T y_T}\left[\ell\left(\hat{y}_T,y_T\right)+\mathbf{Rel}_T\left(\mathscr{H}_T\right)\right]$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{y_T} \left[(\mathbb{E}[y_T] - y_T)^2 + \mathbf{Rel}_T(\mathscr{H}_T) \right] \qquad \text{(D.15)} \\
&= \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{y_T} \left[(\mathbb{E}[y_T] - y_T)^2 + \lambda^{-1} \log_e \left(\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(-\lambda \sum_{s=1}^T (f(x_s) - y_s)^2 \right) \right) \right]
\end{aligned}
$$

Observing that $x = \log_e(\exp(x))$, we can proceed as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x_T} \inf_{q_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \mathbb{E}\left[\ell(\hat{y}_T, y_T) + \mathbf{Rel}_T(\mathscr{H}_T) \right] \\
&= \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{y_T} \left[\lambda^{-1} \log_e \left(\exp\left(\lambda \left(\mathbb{E}[y_t] - y_t \right)^2 \right) \right) \right. \\
&\qquad \left. + \lambda^{-1} \log_e \left(\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(-\lambda \sum_{s=1}^T (f(x_s) - y_s)^2 \right) \right) \right] \\
&= \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{y_T} \lambda^{-1} \log_e \left[\exp\left(\lambda \left(\mathbb{E}[y_t] - y_t \right)^2 \right) \left(\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(-\lambda \sum_{s=1}^T (f(x_s) - y_s)^2 \right) \right) \right] \\
&= \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{y_T} \lambda^{-1} \log_e \left[\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(\lambda \left(\mathbb{E}[y_T] - y_T \right)^2 - \lambda \sum_{s=1}^T (f(x_s) - y_s)^2 \right) \right].
\end{aligned}
$$

Observing that $(\mathbb{E}[y_T] - y_T)^2 - (f(x_T) - y_T)^2 = 2(y_T - \mathbb{E}[y_T])(f(x_T) - \mathbb{E}[y_T]) - (f(x_T) - \mathbb{E}[y_T])^2$, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x_T} \inf_{q_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \mathbb{E}\left[\ell(\hat{y}_T, y_T) + \mathbf{Rel}_T(\mathscr{H}_T) \right] \\
&= \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{y_T} \lambda^{-1} \log_e \left[\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(2\lambda (y_T - \mathbb{E}[y_T])(f(x_T) - \mathbb{E}[y_T]) - \lambda (f(x_T) - \mathbb{E}[y_T])^2 \right. \right. \\
&\qquad \left. \left. - \lambda \sum_{s=1}^{T-1} (f(x_s) - y_s)^2 \right) \right].
\end{aligned}
$$

By concavity of the logarithm function and the Jensen inequality, we can further upper bound

$$
\mathop{\mathbb{E}}_{x_T} \inf_{q_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \mathbb{E}\left[\ell(\hat{y}_T, y_T) + \mathbf{Rel}_T(\mathscr{H}_T) \right]
$$

$$\leq \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \lambda^{-1} \log_e \left\{ \sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \mathop{\mathbb{E}}_{y_T} \exp\left[2\lambda (y_T - \mathbb{E}[y_T])(f(x_T) - \mathbb{E}[y_T]) - \lambda (f(x_T) - \mathbb{E}[y_T])^2 \right.\right. \tag{D.16}$$

$$\left.\left. -\lambda \sum_{t=1}^{T-1} (f(x_t) - y_t)^2 \right] \right\}.$$

Note that $y_t - \mathbb{E}[y_t] \in [-1, 1]$, which implies it is 1-subGaussian (Lemma 1.8 of [153]), and hence

$$\mathop{\mathbb{E}}_{y_t} \exp\left[2\lambda (y_t - \mathbb{E}[y_t])(f(x_t) - \mathbb{E}[y_t])\right] \leq \exp\left[2\lambda^2 (f(x_T) - \mathbb{E}[y_T])^2\right]. \tag{D.17}$$

Applying inequality (D.17) to the RHS of (D.16), we arrive at an upper bound

$$\mathop{\mathbb{E}}_{x_T} \inf_{q_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \left[\ell(\hat{y}_T, y_T) + \mathbf{Rel}_T(\mathscr{H}_T)\right]$$

$$\leq \mathop{\mathbb{E}}_{x_T} \sup_{p_T} \lambda^{-1} \log_e \left\{ \sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left[\left(2\lambda^2 - \lambda\right) (f(x_T) - \mathbb{E}[y_T])^2 - \lambda \sum_{t=1}^{T-1} (f(x_t) - y_t)^2 \right] \right\}. \tag{D.18}$$

Taking $\lambda = 1/2$ in the RHS of (D.18), we further have

$$\mathop{\mathbb{E}}_{x_T} \inf_{q_T} \sup_{p_T} \mathop{\mathbb{E}}_{\hat{y}_T y_T} \left[\ell(\hat{y}_T, y_T) + \mathbf{Rel}_T(\mathscr{H}_T)\right] \leq 2 \mathop{\mathbb{E}}_{x_T} \log_e \left\{ \sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left[-\frac{\sum_{t=1}^{T-1} (f(x_t) - y_t)^2}{2} \right] \right\}.$$

D.3.4 Proof of Lemma 84

Following exactly the same steps until eq. (D.15) in the proof of Lemma 83, we have

$$\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \sup_{p_t} \mathop{\mathbb{E}}_{\hat{y}_t y_t} \left[\ell(\hat{y}_t, y_t) + \mathbf{Rel}_t(\mathscr{H}_t)\right] = \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{y_t} \left[(\mathbb{E}[y_t] - y_t)^2 + \mathbf{Rel}_t(\mathscr{H}_t) \right].$$

By definition of the relaxation in (D.7) and minimax theorem, we have

$$\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \sup_{p_t} \mathop{\mathbb{E}}_{\hat{y}_t y_t} \left[\ell(\hat{y}_t, y_t) + \mathbf{Rel}_t(\mathscr{H}_t)\right]$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \inf_{q_t} \mathop{\mathbb{E}\mathbb{E}}_{\hat{y}_t y_t} \left[\ell\left(\hat{y}_t, y_t\right) + \mathbf{Rel}_t\left(\mathscr{H}_t\right) \right] \\
&= \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{y_t} \mathop{\mathbb{E}}_{x_{t+1}} \ldots \mathop{\mathbb{E}}_{x_T} \left[\left(\mathbb{E}[y_t] - y_t\right)^2 + 2\log_e \left(\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(- \frac{\sum_{s=1}^t \left(f(x_s) - y_s\right)^2}{2} \right) \right) \right],
\end{aligned}
$$

Observing that $x = \log_e(\exp(x))$, we can proceed as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \sup_{p_t} \mathop{\mathbb{E}\mathbb{E}}_{\hat{y}_t y_t} \left[\ell\left(\hat{y}_t, y_t\right) + \mathbf{Rel}_t\left(\mathscr{H}_t\right) \right] \\
&= \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{y_t} \mathop{\mathbb{E}}_{x_{t+1}} \ldots \mathop{\mathbb{E}}_{x_T} \left[2\log_e \left(\exp\left(\frac{\left(\mathbb{E}[y_t] - y_t\right)^2}{2} \right) \right) \right. \\
&\quad \left. + 2\log_e \left(\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(- \frac{\sum_{s=1}^t \left(f(x_s) - y_s\right)^2}{2} \right) \right) \right] \\
&= \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{y_t} \mathop{\mathbb{E}}_{x_{t+1}} \ldots \mathop{\mathbb{E}}_{x_T} 2\log_e \left[\exp\left(\frac{\left(\mathbb{E}[y_t] - y_t\right)^2}{2} \right) \left(\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(- \frac{\sum_{s=1}^t \left(f(x_s) - y_s\right)^2}{2} \right) \right) \right] \\
&= \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{y_t} \mathop{\mathbb{E}}_{x_{t+1}} \ldots \mathop{\mathbb{E}}_{x_T} 2\log_e \left[\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(\frac{\left(\mathbb{E}[y_t] - y_t\right)^2 - \sum_{s=1}^t \left(f(x_s) - y_s\right)^2}{2} \right) \right].
\end{aligned}
$$

Note that $\left(\mathbb{E}[y_t] - y_t\right)^2 - \left(f(x_t) - y_t\right)^2 = 2(y_t - \mathbb{E}[y_t])(f(x_t) - \mathbb{E}[y_t]) - (f(x_t) - \mathbb{E}[y_t])^2$, we can continue as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \sup_{p_t} \mathop{\mathbb{E}\mathbb{E}}_{\hat{y}_t y_t} \left[\ell\left(\hat{y}_t, y_t\right) + \mathbf{Rel}_t\left(\mathscr{H}_t\right) \right] \\
&= \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{y_t} \mathop{\mathbb{E}}_{x_{t+1}} \ldots \mathop{\mathbb{E}}_{x_T} 2\log_e \left[\right. \\
&\qquad \left. \sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(\frac{2(y_t - \mathbb{E}[y_t])(f(x_t) - \mathbb{E}[y_t]) - (f(x_t) - \mathbb{E}[y_t])^2 - \sum_{s=1}^{t-1} \left(f(x_s) - y_s\right)^2}{2} \right) \right].
\end{aligned}
$$

By concavity of the logarithm function and the Jensen inequality, we can further upper bound

$$
\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \sup_{p_t} \mathop{\mathbb{E}\mathbb{E}}_{\hat{y}_t y_t} \left[\ell\left(\hat{y}_t, y_t\right) + \mathbf{Rel}_t\left(\mathscr{H}_t\right) \right]
$$

$$\leq \mathop{\mathbb{E}}_{x_t} \sup_{p_t} \mathop{\mathbb{E}}_{x_{t+1}} \dots \mathop{\mathbb{E}}_{x_T} 2\log_e \Bigg[\tag{D.19}$$

$$\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \mathop{\mathbb{E}}_{y_t} \exp\left(\frac{2(y_t - \mathbb{E}[y_t])(f(x_t) - \mathbb{E}[y_t]) - (f(x_t) - \mathbb{E}[y_t])^2 - \sum_{s=1}^{t-1}(f(x_s) - y_s)^2}{2}\right)\Bigg].$$

Note that $y_t - \mathbb{E}[y_t] \in [-1,1]$, which implies it is 1-subGaussian (Lemma 1.8 of [153]), and hence

$$\mathop{\mathbb{E}}_{y_t} \exp\left[\frac{2(y_t - \mathbb{E}[y_t])(f(x_t) - \mathbb{E}[y_t])}{2}\right] \leq \exp\left[\frac{(f(x_t) - \mathbb{E}[y_t])^2}{2}\right]. \tag{D.20}$$

Applying inequality (D.20) to the RHS of (D.19), we arrive at an upper bound

$$\begin{aligned}\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \sup_{p_t} \mathbb{E}\mathbb{E}[\ell(\hat{y}_t, y_t) + \mathbf{Rel}_t(\mathscr{H}_t)] \leq & \mathop{\mathbb{E}}_{x_t} \mathop{\mathbb{E}}_{x_{t+1}} \dots \mathop{\mathbb{E}}_{x_T} 2\log_e \left[\sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^T\right)} \exp\left(\frac{-\sum_{s=1}^{t-1}(f(x_s) - y_s)^2}{2}\right)\right] \\ = & \mathbf{Rel}_{t-1}(\mathscr{H}_{t-1}).\end{aligned} \tag{D.21}$$

D.3.5 Proof of Lemma 85

From pg. 46 of [53], we know that $\exp(-(a-b)^2/2)$ is concave in a if $(a-b)^2 \leq 1$. Note that $\left(\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)] - y_t\right)^2 \leq 1$ since both $\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)]$ and y_t belong to $[0,1]$. We can thus apply the Jensen's inequality as follows

$$\begin{aligned}& \exp\left(-\frac{\left(\mathop{\mathbb{E}}_{f(x_t)\sim q_t}[f(x_t)] - y_t\right)^2}{2}\right) \\ \geq & \mathop{\mathbb{E}}_{f(x_t)\sim q_t}\left[\exp\left(-\frac{(f(x_t) - y_t)^2}{2}\right)\right] \\ = & \sum_{f \in \mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)} \exp\left(-\frac{(f(x_t) - y_t)^2}{2}\right) \\ & \times \frac{\exp\left(-\sum_{s=1}^{t-1}(f(x_s) - y_s)^2/2\right)}{\sum_{f' \in \mathscr{F}\left(\{x_s\}_{s=1}^t \cup \{x'_j\}_{j=t+1}^T\right)} \exp\left(-\sum_{s=1}^{t-1}(f'(x_s) - y_s)^2/2\right)}\end{aligned} \tag{D.22}$$

$$= \frac{\sum_{f \in \mathscr{F}\left(\{x_s\}_{s=1}^{t} \cup \{x'_j\}_{j=t+1}^{T}\right)} \exp\left(-\sum_{s=1}^{t}(f(x_s)-y_s)^2/2\right)}{\sum_{f' \in \mathscr{F}\left(\{x_s\}_{s=1}^{t} \cup \{x'_j\}_{j=t+1}^{T}\right)} \exp\left(-\sum_{s=1}^{t-1}(f'(x_s)-y_s)^2/2\right)},$$

where we have used the definition of q_t (defined in eq. (D.8)) in eq. (D.22). Rearranging the terms, we can conclude the proof.

D.3.6 Proof of Lemma 86

The first inequality is an immediate consequence of Corollary 82. To see the second part, we note that by definition

$$\begin{aligned}
\mathbf{Rel}_0(\mathscr{H}_0) =& 2\mathop{\mathbb{E}}_{x_1} \ldots \mathop{\mathbb{E}}_{x_T} \log_e \left\{ \sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^{T}\right)} \exp\left[-\frac{\sum_{s=1}^{0}(f(x_s)-y_s)^2}{2} \right] \right\} \\
=& 2\mathop{\mathbb{E}}_{x_1} \ldots \mathop{\mathbb{E}}_{x_T} \log_e \left\{ \sum_{f \in \mathscr{F}\left(\{x_t\}_{t=1}^{T}\right)} 1 \right\} \\
=& 2\log_e\left(\left|\mathscr{F}\left(\{x_t\}_{t=1}^{T}\right)\right|\right). \qquad \text{(D.23)}
\end{aligned}$$

From the proof of Theorem 4 of [121], we know that $\left|\mathscr{F}\left(\{x_t\}_{t=1}^{T}\right)\right| \leq (T+1)^K$. Therefore, (D.23) is at most $K\log_e(T+1)$.

D.4 Dynamic Programming Acceleration

Following [121], we can define for each $k \in \{0,\ldots,K\}$ and $s \in [T]$

$$\begin{aligned}
w_s^k &= \sum_{0 \leq f(z_1) \leq \ldots \leq f(z_s) = \frac{k}{K}} \exp\left(-\frac{\sum_{q<t:x_q<z_s} \left(f\left(x_q\right)-y_q\right)^2}{2} \right), \\
v_s^k &= \sum_{\frac{k}{K} = f(z_s) \leq \ldots \leq f(z_T) \leq 1} \exp\left(-\frac{\sum_{q<t:x_q>z_s} \left(f\left(x_q\right)-y_q\right)^2}{2} \right).
\end{aligned}$$

Suppose x_t is the i^{th} smallest in all the covariates, *i.e.*, $x_t = z_i$, then it can be readily verified that

$$\begin{aligned}
&\sum_{f\in\mathscr{F}_t} f(x_t)\exp\left(-\frac{\sum_{j=1}^{t-1}(f(x_j)-y_j)^2}{2}\right)\\
&=\sum_{k=0}^{K}\sum_{f\in\mathscr{F}_t:f(x_t)=\frac{k}{K}} f(x_t)\exp\left(-\frac{\sum_{j=1}^{t-1}(f(x_j)-y_j)^2}{2}\right)\\
&=\sum_{k=0}^{K}\frac{k}{K}\sum_{f\in\mathscr{F}_t:f(x_t)=\frac{k}{K}} \exp\left(-\frac{\sum_{q<t:x_q<x_t=z_i}(f(x_j)-y_j)^2}{2}\right)\exp\left(-\frac{\sum_{q<t:x_q>x_t=z_i}(f(x_j)-y_j)^2}{2}\right)\\
&=\sum_{k=0}^{K}\frac{k}{K}\left[\sum_{0\le f(z_1)\le\ldots\le f(z_i)=f(x_t)=\frac{k}{K}} \exp\left(-\frac{\sum_{q<t:x_q<x_t=z_i}(f(x_j)-y_j)^2}{2}\right)\right]\\
&\quad\times\left[\sum_{\frac{k}{K}=f(x_t)=f(z_i)\le\ldots\le f(z_T)\le 1} \exp\left(-\frac{\sum_{q<t:x_q>x_t=z_i}(f(x_j)-y_j)^2}{2}\right)\right]\\
&=\sum_{k=0}^{K}\frac{k}{K}w_i^k v_i^k
\end{aligned}\tag{D.24}$$

and similarly, $\sum_{f\in\mathscr{F}_t}\exp\left(-\frac{\sum_{j=1}^{t-1}(f(x_j)-y_j)^2}{2}\right)=\sum_{k=0}^{K}w_i^k v_i^k$. By definition in (5.7), $\hat{y}_t=\frac{\sum_{k=0}^{K}\frac{k}{K}w_i^k v_i^k}{\sum_{k=0}^{K}w_i^k v_i^k}$. To this end, we can compute w_s^k from $s=1$ to i for all $k\in\{0,\ldots,K\}$ as follows. We recall we have defined for every $s\in[T]$ and every $k\in\{0,\ldots,K\}$,

$$u_s^k=\exp\left(-\mathbf{1}\left[\left(z_s\in\{x_j\}_{j=1}^{t}\right)\right]\cdot\frac{\left(k/K-y_{\sigma(s)}\right)^2}{2}\right),$$

where $\sigma(s)$ is the corresponding subscript of the x_q that is equal to z_s if $z_s\in\{x_j\}_{j=1}^{t}$, *i.e.*, $\sigma(s)=q$.

Hence, starting from $w_0^k=1$ for every k, we have the recursive equations for all $k\in\{0,\ldots,K\}$,

$$\begin{aligned}
&w_{s+1}^k\\
&=\sum_{0\le f(z_1)\le\ldots\le f(z_{s+1})=\frac{k}{K}}\exp\left(-\frac{\sum_{q<t:x_q<z_{s+1}}\left(f\left(x_q\right)-y_q\right)^2}{2}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=0}^{k} \exp\left(-\mathbf{1}\left[\left(z_s \in \{x_j\}_{j=1}^{t}\right)\right] \cdot \frac{\left(j/K - y_{\sigma(s)}\right)}{2}\right) \\
&\quad \times \left[\sum_{0 \le f(z_1) \le \ldots < f(z_s) = \frac{j}{K}} \exp\left(-\frac{\sum_{q<t: x_q < z_s} \left(f\left(x_q\right) - y_q\right)^2}{2}\right)\right] \\
&= \sum_{j=0}^{k} u_s^j w_s^j. \qquad \text{(D.25)}
\end{aligned}
$$

For v_s^k's, starting from $u_T^k = 1$ for every k, we have the recursive equations for all $k \in \{0, \ldots, K\}$,

$$
\begin{aligned}
&v_{s-1}^k \\
&= \sum_{\frac{k}{K} = f(z_{s-1}) \le \ldots \le f(z_T) \le 1} \exp\left(-\frac{\sum_{q<t: x_q > z_{s-1}} \left(f\left(x_q\right) - y_q\right)^2}{2}\right) \\
&= \sum_{j=k}^{K} \exp\left(-\mathbf{1}\left[\left(z_s \in \{x_j\}_{j=1}^{t}\right)\right] \cdot \frac{\left(j/K - y_{\sigma(s)}\right)}{2}\right) \\
&\quad \times \left[\sum_{\frac{j}{K} = f(z_s) \le \ldots f(z_T) \le 1} \exp\left(-\frac{\sum_{q<t: x_q > z_s} \left(f\left(x_q\right) - y_q\right)^2}{2}\right)\right] \\
&= \sum_{j=k}^{K} u_s^j v_s^j. \qquad \text{(D.26)}
\end{aligned}
$$