Improved Online Confidence Bounds for Multinomial Logistic Bandits

Joongkyu Lee [1] **Min-hwan Oh** [1]

Abstract

In this paper, we propose an improved online confidence bound for multinomial logistic (MNL) models and apply this result to MNL bandits, achieving variance-dependent optimal regret. Recently, Lee & Oh (2024) established an online confidence bound for MNL models and achieved nearly minimax-optimal regret in MNL bandits. However, their results still depend on the norm-boundedness of the unknown parameter B and the maximum size of possible outcomes K. To address this, we first derive an online confidence bound of $\mathcal{O}\left(\sqrt{d\log t} + B\sqrt{d}\right)$, which is a significant improvement over the previous bound of $\mathcal{O}(B\sqrt{d}\log t\log K)$ (Lee & Oh, 2024). This is mainly achieved by establishing tighter self-concordant properties of the MNL loss and applying Ville's inequality to bound the estimation error. Using this new online confidence bound, we propose a constant-time algorithm, `OFU-MNL++`, which achieves a variance-dependent regret bound of $\mathcal{O}\left(d\log T\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$ for sufficiently large T, where σ_t^2 denotes the variance of the rewards at round t, d is the dimension of the contexts, and T is the total number of rounds. Furthermore, we introduce a Maximum Likelihood Estimation (MLE)-based algorithm, `OFU-M`^{2}`NL`, which achieves an anytime $\mathrm{poly}(B)$-free regret of $\mathcal{O}\left(d\log(BT)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$.

1. Introduction

The multinomial logistic (MNL) bandit framework (Rusmevichientong et al., 2010; Sauré & Zeevi, 2013; Agrawal et al., 2017; 2019; Oh & Iyengar, 2019; 2021; Perivier & Goyal, 2022; Agrawal et al., 2023; Lee & Oh, 2024) provides a principled approach to tackling sequential assortment selection problems. At every round t, an agent offers an assortment of at most K items among total N items and receives feedback *only* for the chosen decisions. The user choice probability follows an MNL model (McFadden, 1977). This framework is widely deployed in industry, with applications ranging from news recommendation systems to online retail, where assortment selections are optimized based on user-choice feedback from the offered options. In such applications, the agent often has access to item features and, potentially, contextual information about the user. This setup is referred to as the *contextual* MNL bandit problem (Agrawal et al., 2019; 2017; Ou et al., 2018; Chen et al., 2020; Oh & Iyengar, 2019; 2021; Perivier & Goyal, 2022; Agrawal et al., 2023; Lee & Oh, 2024).

Recently, in contextual MNL bandits, Lee & Oh (2024) proposed a constant-time algorithm and obtained a regret of $\mathcal{O}(B^{3/2}d\log K(\log T)^{3/2}\sqrt{T})$. Although this result is nearly minimax optimal when ignoring B and logarithmic terms, it still depends on the maximum assortment size K and the norm-boundedness of the parameter B. Intuitively, a larger K may provide more information (Lee & Oh, 2024), suggesting that the regret should not scale with any factor involving K. Moreover, while $\mathrm{poly}(B)$-free regret bound has been established for generalized linear model (GLM) bandits (Lee et al., 2024b) using the MLE, it remains unclear whether such a bound can be obtained while maintaining a constant per-round computational cost.

Our main goal is to design a constant-time algorithm that achieves improved regret with respect to $\mathrm{poly}(B)$ and K. The main challenge in achieving such regret lies in deriving a tight confidence bound. Currently, the best-known online confidence bound is $\mathcal{O}(B\sqrt{d}\log t\log K)$ (Lee & Oh, 2024), which explicitly depends on both B and $\log K$. This dependency poses a significant bottleneck for obtaining improved regret. Furthermore, to the best of our knowledge, there is no variance-dependent regret in contextual MNL bandits. Hence, the following research questions arise:

- *Can we derive a B, K-improved confidence bound for online parameter estimation in MNL models?*
- *Can we design a constant-time algorithm that achieves B, K-improved (or free) and variance-dependent regret in contextual MNL bandits?*

[1]Seoul National University, Seoul, Korea. Correspondence to: Min-hwan Oh <minoh@snu.ac.kr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

In the first part of our main results (Section 4.1), we construct a K-free online confidence bound with improved dependence on B, which depends on a *update condition parameter* α. This significantly improves upon previous results in online parameter estimation in MNL models (Zhang & Sugiyama, 2024; Lee & Oh, 2024). To achieve this, we first establish self-concordant-like properties with respect to the ℓ_∞-norm (instead of the traditional ℓ_2-norm) (Propositions B.3 to B.6). This improvement enhances the existing self-concordant properties of MNL models (Tran-Dinh et al., 2015), which is of independent interest. Then, unlike Zhang & Sugiyama (2024); Lee & Oh (2024), we apply Ville's inequality (Ville, 1939) to bound estimation errors, thereby avoiding the need for a smoothing technique (Foster et al., 2018), which would otherwise lead to a $\mathcal{O}(\sqrt{\log t \log K})$ looser confidence bound.

In the second part (Section 4.2), we propose a constant-time algorithm, called `OFU-MNL++`, that achieves B-improved, K-free, and variance-dependent regret. This algorithm updates the parameter only within a specific space constructed during the *adaptive warm-up* rounds. With high probability, this space contains the true parameter $\mathbf{w}^\star$, while also shrinking sufficiently relative to the current feature set $\mathcal{X}_t$. This is the key to keeping the update condition parameter α of the online confidence bound small (or constant). Note that since the parameter is updated in a fully online manner, the computational cost per round of `OFU-MNL++` remains constant throughout all rounds. Furthermore, we introduce a novel regret decomposition, which ultimately allows us to achieve a variance-dependent regret bound of $\mathcal{O}\Big(\big(d\log T + Bd\sqrt{\log T}\big)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\Big)$, where σ_t^2 denotes the variance of the (chosen) rewards at round t.

In the final part (Section 4.3), as an independent contribution and inspired by Lee et al. (2024b), we propose an MLE-based algorithm, `OFU-M`^{2}`NL`, that leverages an MLE confidence bound and achieves completely $\mathrm{poly}(B), K$-free regret with only $\log B$ dependence. However, note that the per-round computational cost of `OFU-M`^{2}`NL` increases linearly with t due to the use of the MLE, whereas the per-round computational cost of `OFU-MNL++` remains constant.

Our main contributions are summarized as follows:

- **Sharper online confidence bound for the MNL models**: We first establish a confidence bound for online parameter updates in MNL models, which depends on the update condition parameter α (defined later). In Theorem 4.2, when the parameter is updated over the entire space $\mathbb{B}^d(B)$, as is common in prior works (Faury et al., 2022; Zhang & Sugiyama, 2024; Lee & Oh, 2024), we achieve a confidence bound of $\mathcal{O}(B\sqrt{d\log t} + B^{3/2}\sqrt{d} + B^2)$, significantly improving upon the previous bound of $\mathcal{O}(B\sqrt{d}\log t \log K + B^{3/2}\sqrt{d\log K})$ (Lee & Oh, 2024). More importantly, when the parameter is updated within a specific space where the update condition parameter α is bounded by a constant, we achieve a confidence bound of $\mathcal{O}(\sqrt{d\log t} + B\sqrt{d})$, which is completely independent of $\mathrm{poly}(B)$ and K.

- **New B-improved, K-free, variance-dependent regret bound**: To apply our new online confidence bound to MNL bandits and achieve a tighter regret in terms of $\mathrm{poly}(B)$ and K, we propose an algorithm called `OFU-MNL++`. In addition, through a novel regret decomposition, we derive a variance-dependent optimal regret of $\mathcal{O}\Big(\big(d\log T + Bd\sqrt{\log T}\big)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\Big)$, where $\sigma_t^2 \leqslant 1$ represents the variance of the rewards at round t. For sufficiently large T, we obtain a $\tilde{\mathcal{O}}\Big(d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\Big)$ regret. To the best of our knowledge, this is the first B, K-free and variance-dependent optimal regret bound in contextual MNL bandits.

- **Completely poly$(B), K$-free confidence and regret bound using MLE**: We propose an MLE-based algorithm, `OFU-M`^{2}`NL`, which achieves $\mathrm{poly}(B), K$-free variance-dependent optimal regret of $\mathcal{O}(d\log(BT)\sqrt{\sum_{t=1}^{T}\sigma_t^2})$ by leveraging a B-free MLE confidence bound.

2. Related Work

Logistic bandits. The logistic bandit problem (Dong et al., 2019; Faury et al., 2020; Abeille et al., 2021; Faury et al., 2022; Lee et al., 2024a;b) is a special case of the MNL bandit problem. In this setting, the agent offers only a single item (i.e., $K = 1$) and receives 0-1 binary feedback, restricting the problem to the uniform rewards setting. As summarized in Table 1, recent works have successfully eliminated the harmful dependency on $1/\kappa$ (which can be exponentially large) in the leading term, achieving instance-dependent regret (i.e., $\kappa_t^\star$-dependent regret). However, most of these approaches still suffer from an unnecessary dependency on the norm-boundedness of the unknown parameter, $\mathrm{poly}(B)$. While a recent work by Lee et al. (2024b) successfully eliminated the $\mathrm{poly}(B)$ factors, their approach incurs a per-round computational cost that grows linearly with t. Thus, the question of whether it is possible to design a B-free, computationally efficient algorithm remains open.

MNL bandits. The MNL bandits (Agrawal et al., 2019; 2017; Ou et al., 2018; Chen et al., 2020; Oh & Iyengar, 2019; 2021; Perivier & Goyal, 2022; Agrawal et al., 2023; Lee & Oh, 2024) address more sophisticated problems compared to logistic bandits, as they involve selecting a set of items (thus highlighting their combinatorial nature) and consider

Table 1: Comparisons of regret bounds in recent works on contextual logistic and MNL bandits with T rounds, the maximum size of assortments K, d-dimensional feature vectors, the norm-boundedness of the unknown parameter B, problem-dependent constants $1/\kappa = \mathcal{O}(K^2 e^{3B})$ and $\kappa_t^\star := \sum_{i \in S_t^\star} p_t(i|S_t^\star, \mathbf{w}^\star) p_t(0|S_t^\star, \mathbf{w}^\star) \leqslant 1$, and the variance of the rewards $\sigma_t^2 \leqslant 1$ at round t (formally defined in (9)). For the computational cost (abbreviated as "Comput."), we consider only the dependence on t. The term "Intractable" refers to computational runtimes that are non-polynomial.

	Algorithm	Regret	Rewards	Comput. per Round
Logistic Bandits	Abeille et al. (2021) (OFULog)	$\mathcal{O}\Big(B^{3/2} d \log T \sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	Uniform	Intractable
	Abeille et al. (2021) (OFULog-r)	$\mathcal{O}\Big(B^{5/2} d \log T \sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	Uniform	$\mathcal{O}(t)$
	Faury et al. (2022) (ada-OFU-ECOLog)	$\mathcal{O}\Big(B d \log T \sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	Uniform	$\mathcal{O}(\log t)$
	Lee et al. (2024b) (OFUGLB)	$\mathcal{O}\Big(d \log(BT) \sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	Uniform	$\mathcal{O}(t)$
MNL Bandits	Chen et al. (2020) (MLE-UCB)	$\mathcal{O}\Big(B d \log(KT) \sqrt{T}\Big)$	Uniform/Non-Uniform	Intractable
	Oh & Iyengar (2021) (UCB-MNL)	$\mathcal{O}\Big(\frac{1}{\kappa} d \log T \sqrt{T}\Big) = \mathcal{O}\Big(K^2 e^B d \log T \sqrt{T}\Big)$	Uniform/Non-Uniform	$\mathcal{O}(t)$
	Perivier & Goyal (2022) (OFU-MNL)	$\mathcal{O}\Big(BK d \log(KT) \sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	Uniform	Intractable
	Lee & Oh (2024) (OFU-MNL+)	$\mathcal{O}\Big(B^{3/2} d \log K (\log T)^{3/2} \sqrt{T}\Big)$	Uniform/Non-Uniform	$\mathcal{O}(1)$
	This work (OFU-MNL++, Theorem 4.5)	$\mathcal{O}\Big(\big(d \log T + B d \sqrt{\log T}\big) \sqrt{\sum_{t=1}^T \sigma_t^2} \Big)$	Uniform/Non-Uniform	$\mathcal{O}(1)$
	This work (OFU-M^{2}NL, Theorem 4.12)	$\mathcal{O}\Big(d \log(BT) \sqrt{\sum_{t=1}^T \sigma_t^2}\Big)$	Uniform/Non-Uniform	$\mathcal{O}(t)$

non-uniform rewards rather than binary feedback. Recently, Lee & Oh (2024) made significant progress by resolving the long-standing open problem of establishing the minimax optimal regret (ignoring factors of B and logarithmic terms) with computational efficiency. However, as shown in Table 1, all existing regret bounds increase with B and K. Furthermore, the tightest regret bound by Lee & Oh (2024) includes an additional $(\log T)^{1/2}$ term, arising from a loose confidence bound. To address these limitations, in this paper, we construct the sharper online confidence bound to date and, leveraging this, achieve (asymptotically) B, K-free regret while maintaining computational efficiency.

RL with MNL models. There has been growing interest in incorporating MNL models into reinforcement learning (RL). One line of work extends MNL bandits to the RL setting. Recently, Lee & Oh (2025) proposed a new framework, called *combinatorial RL with preference feedback*, in which the agent selects a subset of items in each round to maximize long-term cumulative reward based on MNL-modeled preferences, and established the minimax-optimal regret bound in linear MDPs (Jin et al., 2020).

Another direction focuses on RL with MNL-based transition models. Hwang & Oh (2022) introduced *MNL-MDPs*, a class of MDPs where the transition probabilities are parameterized by an MNL model. Building on this, Cho et al. (2024) and Li et al. (2024) concurrently improved the dependency on $1/\kappa = \mathcal{O}(K^2 e^{3B})$ in their regret bounds. Park et al. (2024) further extended this direction to the infinite-horizon setting.

3. Preliminaries

Notations. For a positive integer n, we define $[n]$ as the set $\{1, 2, \ldots, n\}$. The ℓ_2- and ℓ_∞-norm of a vector x is denoted by $\|x\|_2$ and $\|x\|_\infty$, respectively. For a positive semi-definite matrix A and a vector x, we use $\|x\|_A$ to represent $\sqrt{x^\top A x}$. For any two symmetric matrices A and B of the same dimensions, $A \succeq B$ indicates that $A - B$ is a positive semi-definite matrix. Finally, we define $\mathcal{S}$ as the set of candidate assortments with a size constraint of at most K, i.e., $\mathcal{S} = \{S \subseteq [N] : |S| \leqslant K\}$.

3.1. Problem Setting

We consider the contextual MNL bandit problem, where an agent selects assortments (sets of items) and receives feedback based on user choices. Specifically, at each round t, the agent receives a feature vector $x_{ti} \in \mathbb{R}^d$ and a reward r_{ti} for every item $i \in [N]$. Note that the feature set $\mathcal{X}_t := \{x_{ti}\}_{i=1}^N$ and rewards $\{r_{ti}\}_{i=1}^N$ can be *arbitrarily* chosen by an adversary. The agent then offers an assortment $S_t = \{i_1, \ldots, i_l\} \in \mathcal{S}$, where $l \leqslant K$. After presenting the assortment, the agent observes the user's purchase decision

$c_t \in S_t \cup \{0\}$, where $\{0\}$ represents the "outside option", indicating that the user did not choose any item from S_t. The user choices are modeled using the Multinomial Logistic (MNL) framework (McFadden, 1977), where the probability of selecting an item $i \in S_t \cup \{0\}$ is defined as:

$$p_t(i|S_t, \mathbf{w}^\star) := \frac{\exp(x_{ti}^\top \mathbf{w}^\star)}{1+\sum_{j\in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)},$$

where $\mathbf{w}^\star \in \mathbb{R}^d$ is an *unknown* parameter and $x_{t0} = \mathbf{0}$.

The choice response for each item $i \in S_t \cup \{0\}$ is defined as $y_{ti} := \mathbb{1}(c_t = i) \in \{0, 1\}$. Hence, the choice feedback vector $\mathbf{y}_t := (y_{t0}, y_{ti_1}, \dots y_{ti_l})$ is sampled from the multinomial (MNL) distribution $\mathbf{y}_t \sim \mathrm{MNL}\{1, (p_t(0|S_t, \mathbf{w}^\star), \dots, p_t(i_l|S_t, \mathbf{w}^\star))\}$, where the parameter 1 indicates that $\mathbf{y}_t$ is a single-trial sample, meaning $y_{t0} + \sum_{k=1}^{l} y_{ti_k} = 1$. Then, the expected revenue of an assortment S is defined as:

$$R_t(S, \mathbf{w}^\star) := \sum_{i\in S} p_t(i|S, \mathbf{w}^\star) r_{ti} = \frac{\sum_{i\in S} \exp(x_{ti}^\top \mathbf{w}^\star) r_{ti}}{1+\sum_{j\in S} \exp(x_{tj}^\top \mathbf{w}^\star)}.$$

We denote $S_t^\star$ as the optimal assortment at time t, i.e., $S_t^\star := \mathrm{argmax}_{S\in\mathcal{S}} R_t(S, \mathbf{w}^\star)$. The goal of the agent is to minimize the cumulative regret over the T rounds:

$$\mathbf{Reg}_T(\mathbf{w}^\star) := \sum_{t=1}^{T} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star).$$

When $K = 1$ and $r_{t1} = 1$, the MNL bandit reduces to the binary logistic bandit with $R_t(S = \{x\}, \mathbf{w}^\star) = \sigma\left(x^\top \mathbf{w}^\star\right) = 1/(1 + \exp(-x^\top \mathbf{w}^\star))$, where $\sigma(\cdot)$ is the sigmoid function.

We will work under the standard boundedness assumption.

Assumption 3.1 (Bounded assumption)**.** We assume that, for all $t \geqslant 1$, $i \in [N]$, $\|x_{ti}\|_2 \leqslant 1$ and $r_{ti} \in [0, 1]$. There exists a *known* constant such that $\|\mathbf{w}^\star\|_2 \leqslant B$,

Following the previous contextual MNL bandit literature (Oh & Iyengar, 2021; Perivier & Goyal, 2022; Zhang & Sugiyama, 2024; Lee & Oh, 2024), we introduce the problem-dependent constant:

Definition 3.2. Let $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\|_2 \leqslant B\}$. There exists $\kappa > 0$ such that, for any $i \in S$, $S \in \mathcal{S}$, and $t \in [T]$, we have $\min_{\mathbf{w}\in\mathcal{W}} p_t(i|S, \mathbf{w}) p_t(0|S, \mathbf{w}) \geqslant \kappa$.

A small κ signifies a greater deviation from the linear model. Notably, $1/\kappa$ can be exponentially large, growing on the order of $\mathcal{O}(K^2 e^{3B})$. Therefore, it is crucial to ensure that our regret bound does not depend on $1/\kappa$.

4. Main Results

4.1. Sharper Online Confidence Bound for MNL Model

Instead of performing Maximum Likelihood Estimation (MLE) as done in previous studies (Chen et al., 2020; Oh & Iyengar, 2021; Perivier & Goyal, 2022), we follow the approach of Zhang & Sugiyama (2024); Lee & Oh (2024) and adopt the online mirror descent (OMD) algorithm for parameter estimation. To begin, we define the multinomial logistic loss function for round t as:

$$\ell_t(\mathbf{w}) := -\sum_{i\in S_t} y_{ti} \log p_t(i|S_t, \mathbf{w}). \tag{1}$$

In this paper, we present a *general* description of online parameter estimation. We consider a (possibly) *time-varying* compact convex search space $\mathcal{W}_t \subseteq \mathbb{R}^d$ and allow for *occasional updates* to the parameter rather than requiring updates at every round. We denote $\mathcal{T} \subseteq [T]$ as all the update rounds. At the update round $t \in \mathcal{T}$, the true parameter $\mathbf{w}^\star$ is estimated as follows:

$$\begin{aligned}
\mathbf{w}'_t &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_t} \|\mathbf{w} - \mathbf{w}_t\|_{H_t}, && \text{(projection onto } \mathcal{W}_t\text{)} \\
\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_t} \langle \nabla\ell_t(\mathbf{w}'_t), \mathbf{w}\rangle + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}'_t\|^2_{\tilde{H}_t}, && (2)
\end{aligned}$$

where $\eta > 0$ is the step-size parameter, and $\mathcal{W}_t \subseteq \mathbb{R}^d$ is the compact convex set, which will be specified later. The matrix $\tilde{H}_t$ is defined as $\tilde{H}_t := H_t + \eta\nabla^2\ell_t(\mathbf{w}'_t)$, where $H_t := \lambda\mathbf{I}_d + \sum_{s\in\mathcal{T}\setminus\{t,\dots,T\}} \nabla^2\ell_s(\mathbf{w}_{s+1})$ with $\lambda > 0$.

If no update is performed, $\mathbf{w}_t$, $\tilde{H}_t$ and H_t remain unchanged. Formally, let $t' \in \mathcal{T}$ denote the last update round prior to t (i.e., $t' < t$). Then, we have $\mathbf{w}_{t'+1} = \cdots = \mathbf{w}_t$, $H_{t'+1} = \cdots = H_t$, and $\tilde{H}_{t'+1} = \cdots = \tilde{H}_t$.

In the optimization problem (2), we first solve the unconstrained optimization problem in closed form, obtaining $\mathbf{w}'_{t+1}$. Then, we project $\mathbf{w}'_{t+1}$ back into the feasible set.

$$\begin{aligned}
\mathbf{w}'_{t+1} &= \mathbf{w}'_t - \eta\tilde{H}_t^{-1}\nabla\ell_t(\mathbf{w}'_t), \\
\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_t} \|\mathbf{w} - \mathbf{w}'_{t+1}\|_{\tilde{H}_t}, . && (3)
\end{aligned}$$

This estimator is efficient in both computation and storage.

Remark 4.1 (Computational cost). For a general convex set $\mathcal{W}_t$, the projection optimization problem (e.g., Equation (3)) can be solved up to $\epsilon > 0$ accuracy using the Projected Gradient Descent algorithm (e.g., Algorithm 2 in (Hazan et al., 2016)), requiring computational cost of $\mathcal{O}(Kd^3 \log(1/\epsilon))$. As a special case, if $\mathcal{W}_t$ is an ellipsoid, the optimization problem can be solved in a single projection step (via a closed-form projection), which needs only $\mathcal{O}(Kd^3)$ cost.

In terms of storage, the estimator avoids retaining all historical data, as $\tilde{H}_t$, and H_t can be updated incrementally, requiring only $\mathcal{O}(d^2)$ storage.

Our first main contribution is the development of an improved online confidence bound for MNL bandits, which depends on the update condition parameter α. The proof is deferred to Appendix C.

Theorem 4.2 (Improved online confidence bound)**.** *Let $\delta \in (0, 1]$ and $\mathcal{T} \subseteq [T]$ denote the set of update rounds. For all $t \in \mathcal{T}$, we assume the following update conditions hold:*

$$\sup_{\mathbf{w} \in \mathcal{W}_t} |x_{ti}^\top (\mathbf{w} - \mathbf{w}^\star)| \leqslant \alpha, \quad \forall i \in S_t,$$

where $\mathcal{W}_t$ is a compact convex set, and $\alpha > 0$. We set $\eta = (1 + 3\sqrt{2}\alpha)/2$ and $\lambda = \max\{12\sqrt{2}\eta\alpha, 144\eta d, 2\}$. Then, under Assumption 3.1, with probability at least $1 - \delta$, we have:

$$\|\mathbf{w}_t - \mathbf{w}^\star\|_{H_t} = \mathcal{O}\left(\alpha\sqrt{d \log(t/\delta)} + B\sqrt{\lambda}\right).$$

Remark 4.3 (Condition of Theorem 4.2). Note that the condition in Theorem 4.2 is easy to satisfy and has already been addressed in prior works (Faury et al., 2022; Zhang & Sugiyama, 2024; Lee & Oh, 2024). Specifically, if the parameter is updated at every round (i.e., $\mathcal{T} = [T]$) over the entire parameter space (i.e., $\mathcal{W}_t = \mathcal{W}$), as is common in previous works (Faury et al., 2022; Zhang & Sugiyama, 2024; Lee & Oh, 2024), it follows directly that $\alpha = B$.

Discussion of Theorem 4.2. When the parameter is updated at every round (so $\alpha = 2B$) and λ is set to $\lambda = \Theta(Bd + B^2)$, we obtain a completely K-free confidence bound of $\mathcal{O}(B\sqrt{d \log t} + B^{3/2}\sqrt{d} + B^2)$. Compared to the recently established confidence bound $\mathcal{O}(B\sqrt{d} \log t \log K + B^{3/2}\sqrt{d \log K})$ (Lee & Oh, 2024), our bound is tighter by a factor of $\sqrt{\log t} \log K$ in the leading term.

More importantly, and perhaps more interestingly, if we can construct $\mathcal{W}_t$ such that α remains small (or constant) and updates occur *only* when this condition is met, we achieve a confidence bound of $\mathcal{O}(\sqrt{d \log t} + B\sqrt{d})$. For sufficiently large t, i.e., $t \geqslant \mathcal{O}(e^{B^2})$, this further simplifies to $\mathcal{O}(\sqrt{d \log t})$, representing a significant improvement over the previous bound $\mathcal{O}(B\sqrt{d} \log t \log K)$ (Lee & Oh, 2024), with no dependence on B or K.

Proof sketch of Theorem 4.2. We provide a proof sketch and highlight the technical novelties of Theorem 4.2.

Following the previous works (Zhang & Sugiyama, 2024; Lee & Oh, 2024), we first bound the estimation error between $\mathbf{w}_{t+1}$ and $\mathbf{w}^\star$ as follows:

$$\|\mathbf{w}_{t+1} - \mathbf{w}^\star\|_{H_{t+1}}^2 \lesssim \eta \sum_{s \in \mathcal{T}_{t+1}} (\ell_s(\mathbf{w}^\star) - \ell_s(\mathbf{w}_{s+1})) + B^2\lambda, \tag{4}$$

where $\mathcal{T}_{t+1} \subseteq \mathcal{T}$ is the set of update rounds prior to $t + 1$.

1) B, K-independent step size η. In Zhang & Sugiyama (2024); Lee & Oh (2024), η is set as $\eta \simeq \log K + B$, based on Lemma 4 from Jézéquel et al. (2021). To eliminate the dependency on B and $\log K$, we establish Proposition B.3, which shows that the MNL loss is $3\sqrt{2}$-self-concordant with respect to the ℓ_∞-norm (rather than the ℓ_2-norm, as shown in Tran-Dinh et al. (2015)), which may be of independent interest. This result enables us to set $\eta \simeq \alpha$ (Proposition C.5), making it independent of B and K.

2) Intermediary term. Inspired by Zhang & Sugiyama (2024); Lee & Oh (2024), we introduce an intermediary parameter:

$$\tilde{\mathbf{z}}_s := \boldsymbol{\sigma}_s^+ \left(\mathbb{E}_{\mathbf{w} \sim P_s}\left[\boldsymbol{\sigma}_s\left((x_{sj}^\top \mathbf{w})_{j \in S_s}\right)\right]\right),$$

where $\boldsymbol{\sigma}_s$ is the softmax function, $\boldsymbol{\sigma}_s^+$ is its pseudo-inverse, and P_s is a multivariate normal distribution with mean $\mathbf{w}_s'$ and covariance cH_s^{-1} for some $c > 0$. Then, we decompose the sum of loss gaps appearing in the first term of Equation (4) as follows:

$$\begin{aligned}
&\sum_{s \in \mathcal{T}_{t+1}} (\ell_s(\mathbf{w}^\star) - \ell_s(\mathbf{w}_{s+1})) \\
&= \underbrace{\sum_{s \in \mathcal{T}_{t+1}} \left(\ell_s(\mathbf{w}^\star) - \bar{\ell}_s(\tilde{\mathbf{z}}_s)\right)}_{(a)} + \underbrace{\sum_{s \in \mathcal{T}_{t+1}} \left(\bar{\ell}_s(\tilde{\mathbf{z}}_s) - \ell_s(\mathbf{w}_{s+1})\right)}_{(b)}.
\end{aligned}$$

3) Tighter bound for term (a) via Ville's inequality. In Zhang & Sugiyama (2024); Lee & Oh (2024), term (a) is bounded using a Bernstein-type inequality. However, the intermediary parameter cannot be used directly, as it is generally unbounded (Foster et al., 2018). To address this, they employ a *smoothed* version of the parameter, but this leads to a bound of $\mathcal{O}(\log K (\log t)^2)$ for term (a), resulting in a significantly looser confidence bound.

In contrast, we apply *Ville's inequality* (Ville, 1939) without resorting to smoothing. To do so, we first show that the following quantity forms a supermartingale:

$$A_t := \exp\left(\sum_{s \in \mathcal{T}_{t+1}} \left(\ell_s(\mathbf{w}^\star) - \bar{\ell}_s(\tilde{\mathbf{z}}_s)\right)\right).$$

Then, by Ville's inequality, with probability at least $1 - \delta$ (set $\delta \approx \frac{1}{t}$), we can bound term (a) as follows:

$$\sum_{s \in \mathcal{T}_{t+1}} \left(\ell_s(\mathbf{w}^\star) - \bar{\ell}_s(\tilde{\mathbf{z}}_s)\right) \leqslant \log \frac{1}{\delta} \approx \log t, \tag{5}$$

which is an improvement by a factor of $\mathcal{O}(\log K \log t)$ compared to $\mathcal{O}(\log K (\log t)^2)$ (Zhang & Sugiyama, 2024; Lee & Oh, 2024).

On the other hand, we can bound term (b) by applying LemmaF.3 of Lee et al. (2024a) (or Lemma14 of Zhang & Sugiyama (2024)):

$$\sum_{s \in \mathcal{T}_{t+1}} \left(\bar{\ell}_s(\tilde{\mathbf{z}}_s) - \ell_s(\mathbf{w}_{s+1})\right) \lesssim \alpha d \log t. \tag{6}$$

Combining (4), (5), and (6), we complete the proof.

Algorithm 1 OFU-MNL++

Input: failure level δ, confidence radii $\beta_t(\delta)$ and $\zeta_t(\delta)$.
Initialize: $\mathcal{W}_1^w(\delta) = \mathcal{W}$, $H_1 = \lambda \mathbf{I}_d$, $H_1^w = \lambda_1^w \mathbf{I}_d$, $\mathbf{w}_1, \mathbf{w}_1^w \in \mathcal{W}$, $\eta := 1$, $\eta^w := \frac{1}{2} + 3\sqrt{2}B$, $\lambda := 144d$, $\lambda^w := \max\{12\sqrt{2}\eta^w B, 144\eta^w d, 2\}$, $\tau_t := 6\sqrt{2}\zeta_t(\delta)$.
for round $t = 1, \ldots, T$ **do**
  Observe feature set $\mathcal{X}_t = \{x_{ti}\}_{i=1}^N$ and rewards $\{r_{ti}\}_{i=1}^N$.
  if $\max_{x \in \mathcal{X}_t} \|x\|^2_{(H_t^w)^{-1}} \geqslant 1/\tau_t^2$ **then** ▷ *Adaptive warm-up*
    Offer $S_t = \{i_t\}$, where $x_{ti_t} = \operatorname{argmax}_{x \in \mathcal{X}_t} \|x\|^2_{(H_t^w)^{-1}}$, and observe $\mathbf{y}_t$.
    Update $(\mathbf{w}_{t+1}^w, H_{t+1}^w) \leftarrow$ RS-OMD$(\mathcal{W}, \ell_t, H_t^w, \mathbf{w}_t^w, \eta^w)$ by Algorithm 2.
    Calculate $\mathcal{W}_{t+1}^w(\delta) \leftarrow \left\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w} - \mathbf{w}_{t+1}^w\|_{H_{t+1}^w} \leqslant \zeta_{t+1}(\delta)\right\}$.
    Update $H_{t+1} \leftarrow H_t$ and $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$.
  else ▷ *Planning & Learning*
    Offer $S_t = \operatorname{argmax}_{S \in \mathcal{S}} \tilde{R}_t(S)$ and observe $\mathbf{y}_t$.
    Update $(\mathbf{w}_{t+1}, H_{t+1}) \leftarrow$ RS-OMD$(\mathcal{W}_t^w(\delta), \ell_t, H_t, \mathbf{w}_t, \eta)$ by Algorithm 2.
    Update $H_{t+1}^w \leftarrow H_t^w$, $\mathbf{w}_{t+1}^w \leftarrow \mathbf{w}_t^w$, and $\mathcal{W}_{t+1}^w(\delta) \leftarrow \mathcal{W}_t^w(\delta)$.
  end if
end for

Algorithm 2 RS-OMD, **R**estricted **S**pace **O**nline **M**irror **D**escent

Input: convex set $\mathcal{W}_t$, ℓ_t, H_t, $\mathbf{w}_t$, η.
  Update $\tilde{H}_t \leftarrow H_t + \eta \nabla^2 \ell_t(\mathbf{w}_t)$.
  Calculate $\mathbf{w}_{t+1}$ by Equation (2).
  Update $H_{t+1} \leftarrow H_t + \nabla^2 \ell_t(\mathbf{w}_{t+1})$.
Return: $\mathbf{w}_{t+1}$, H_{t+1}.

4.2. Online Update with Adaptive Warm-Up

In this subsection, we introduce OFU-MNL++ (Algorithm 1), which employs a novel two-phase online update approach leveraging the improved confidence bound from Theorem 4.2 to achieve the tightest regret bound in MNL bandits. Note that the feature set $\mathcal{X}_t$ can be arbitrarily given at each round t, without imposing any distributional assumptions on the exogenous contexts.

Intuition. Theorem 4.2 indicates that if α is constant, a confidence bound of $\mathcal{O}(\sqrt{d \log t})$ can be obtained for sufficiently large t. Our primary objective is to design the search space $\mathcal{W}_t$ to ensure that α remains constant in most rounds. To achieve this, we enforce the condition by rejecting, *on-the-fly*, any $\mathcal{X}_t$ that might violate the constancy of α. Specifically, it is sufficient to verify the following condition:

$$\max_{x \in \mathcal{X}_t} \|x\|^2_{(H_t^w)^{-1}} \geqslant 1/\tau_t^2, \tag{7}$$

where $H_t^w := \lambda^w \mathbf{I}_d + \sum_{s \in \mathcal{T}^w \setminus \{t, \ldots, T\}} \nabla^2 \ell_s(\mathbf{w}_{s+1}^w)$ is the warm-up version of H_t, i.e., the regularized sum of Hessians corresponding to all assortments played during the adaptive warm-up rounds $\mathcal{T}^w$. Here, τ_t is a carefully chosen threshold and $\lambda^w > 0$ is a regularization parameter.

Online adaptive warm-up. At round t, given $\mathcal{X}_t$, if for any feature $x \in \mathcal{X}_t$, the quantity $\|x\|^2_{(H_t^w)^{-1}}$ is greater than or equal to the threshold $1/\tau_t^2$ (as specified in Equation (7)), we do not update our current estimate $\mathbf{w}_t$. Instead, we update a separate *warm-up parameter* $\mathbf{w}_t^w$ to ensure that the condition in (7) is more likely to hold in the future.

In such cases, we offer only the single item that maximizes $\|x\|^2_{(H_t^w)^{-1}}$ (Line 6). Subsequently, we update the warm-up parameter $\mathbf{w}_t^w$ by invoking **R**estricted **S**pace **O**nline **M**irror **D**escent (RS-OMD, Algorithm 2) as a subroutine (Line 7). Then, we construct the following parameter set (Line 8):

$$\mathcal{W}_{t+1}^w(\delta) = \left\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w} - \mathbf{w}_{t+1}^w\|_{H_{t+1}^w} \leqslant \zeta_{t+1}(\delta)\right\},$$

where $\zeta_{t+1}(\delta) = \mathcal{O}(B\sqrt{d \log(t/\delta)} + B^{3/2}\sqrt{d} + B^2)$. This ellipsoid is then used in the RS-OMD procedure during the *Planning & Learning* rounds (Line 11- 13). Note that, since the search space is the entire parameter space $\mathcal{W}$, we can set $\alpha = B$ for the condition of Theorem 4.2 to obtain the warm-up confidence bound $\zeta_{t+1}(\delta)$. The quantities H_t and $\mathbf{w}_t$ remain unchanged during the warm-up rounds (Line 9).

Parameter update within restricted space $\mathcal{W}_t^w(\delta)$. When the condition in Equation (7) does not hold, the parameter $\mathbf{w}_t$ is updated by searching only within $\mathcal{W}_{t+1}^w(\delta)$ using RS-OMD as a subroutine (Line 12). In this scenario, α can be set as a constant (with high probability), leading to a confidence bound of $\mathcal{O}(\sqrt{d \log t} + B\sqrt{d})$ (by Theorem 4.2).

Corollary 4.4 (Informal, B-improved & K-free online confidence bound)**.** *Let $\delta \in (0, 1]$ and $\beta_t(\delta) = \mathcal{O}\left(\sqrt{d \log(t/\delta)} + B\sqrt{d}\right)$. Suppose $\mathbf{w}^\star \in \mathcal{W}_t^w(\delta)$ for all $t \geqslant 1$. Define the following confidence set as follows:*

$$\mathcal{C}_t(\delta) := \left\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w} - \mathbf{w}_t\|_{H_t} \leqslant \beta_t(\delta)\right\}.$$

Then, we have $\Pr\left[\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{C}_t(\delta)\right] \geqslant 1-\delta$.

Efficient assortment selection. Given the confidence set in Corollary 4.4, we calculate the optimistic utility UCB_{ti} as:

$$\mathrm{UCB}_{ti} := x_{ti}^\top \mathbf{w}_t + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}, \quad \forall i \in [N].$$

If the true parameter $\mathbf{w}^\star$ lies within the confidence set $\mathcal{C}_t(\delta)$, the value UCB_{ti} serves as an upper bound for $x_{ti}^\top \mathbf{w}^\star$. Using UCB_{ti}, we define the optimistic expected revenue for an assortment S as:

$$\tilde{R}_t(S) := \frac{\sum_{i\in S} \exp(\mathrm{UCB}_{ti}) r_{ti}}{1+\sum_{j\in S}\exp(\mathrm{UCB}_{tj})}, \tag{8}$$

where $r_{ti} \in [0,1]$. We then offer the assortment S_t that maximizes $\tilde{R}_t(S)$, i.e., $S_t = \mathrm{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$ (Line 11). The quantities H_t^w, $\mathbf{w}_t^w$, and $\mathcal{W}_t^w(\delta)$ remain unchanged during the *planning & learning rounds* (Line 13). Note that the optimization problem in (8) can be efficiently solved in polynomial time, $\mathcal{O}(\mathrm{poly}(N))$, independent of t (Rusmevichientong et al., 2010; Davis et al., 2014).

Variance-dependent optimal regret. We establish a variance-dependent optimal regret bound through a novel regret decomposition. Specifically, we show that the regret is bounded by the sum of covariances between r_{ti} and $\|x_{ti}\|_{H_t^{-1}}$, given S_t. Thus, with some slight notational abuse (as the expressions do not strictly denote random variables), the regret can be bounded as follows:

$$\begin{aligned}\mathbf{Reg}_T(\mathbf{w}^\star) &\lesssim \beta_T(\delta)\sum_{t\notin\mathcal{T}^w} \mathrm{Cov}_t\left(r_{ti}, \|x_{ti}\|_{H_t^{-1}}\right)\\ &\lesssim \beta_T(\delta)\sqrt{\sum_{t\notin\mathcal{T}^w}\mathbb{V}_t(r_{ti})}\sqrt{\sum_{t\notin\mathcal{T}^w}\mathbb{V}_t(\|x_{ti}\|_{H_t^{-1}})},\end{aligned}$$

where $\mathrm{Cov}_t(\cdot,\cdot)$ and $\mathbb{V}_t(\cdot)$ is the covariance and variance, respectively, given S_t. For simplicity, rewrite $\mathbb{V}_t(r_{ti})$ as

$$\sigma_t^2 := \mathbb{E}_{i\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[\left(r_{ti} - \mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}[r_{tj}]\right)^2\right], \tag{9}$$

where $r_{t0}=0$. By applying the elliptical potential lemma (Lemma D.7) to the sum of the variances of $\|x_{ti}\|_{H_t^{-1}}$, we derive a variance-dependent regret bound. The complete proof is provided in Appendix D.

Theorem 4.5. *Let $\delta \in (0,1]$, and assume that Assumption 3.1 holds. Then, with probability at least $1-\delta$, the regret of* `OFU-MNL++` *(Algorithm 1) satisfies*

$$\begin{aligned}\mathbf{Reg}_T(\mathbf{w}^\star) \lesssim \left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^T \sigma_t^2}\\ + \frac{1}{\kappa}B^3d^2(\log T)^2 + \frac{1}{\kappa}B^4 d\log T.\end{aligned}$$

Discussion of Theorem 4.5. For sufficiently large, i.e., $T \geqslant \tilde{\mathcal{O}}(e^{B^2} + \frac{1}{\kappa^2}B^8d^2)$, `OFU-MNL++` achieves a regret of $\mathcal{O}\left(d\log T\sqrt{\sum_{t=1}^T\sigma_t^2}\right)$. To the best of our knowledge, this is the first variance-dependent and $\mathrm{poly}(B), K$-free regret bound in contextual MNL bandits. Compared to the recent minimax optimal result of $\mathcal{O}\left(B^{3/2}d\log K(\log T)^{3/2}\sqrt{T}\right)$ by Lee & Oh (2024), our method improves the regret by a factor of $\mathcal{O}\left(B^{3/2}\log K\sqrt{\log T}\right)$. Moreover, the $\tilde{\mathcal{O}}(\sqrt{T})$ term in Lee & Oh (2024) is replaced in our result by $\tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^T\sigma_t^2}\right)$. Since $\sigma_t^2 \leqslant 1$ always holds, this represents a strict improvement over $\sqrt{T}$.

Remark 4.6 (Computational cost of `OFU-MNL++`). The proposed algorithm, `OFU-MNL++`, maintains a constant computational cost per round of $\mathcal{O}(Kd^3 + \mathrm{poly}(N))$, which is entirely independent of t. For parameter updates, we utilize the linearized loss, inspired by Zhang & Sugiyama (2024), and work within ellipsoidal search spaces ($\mathcal{W}$ and $\mathcal{W}_t(\delta)$ in both phases. As a result, the update process incurs only a cost of $\mathcal{O}(Kd^3)$. Moreover, the assortment optimization problem can be solved in $\mathcal{O}(\mathrm{poly}(N))$ (Davis et al., 2014).

Remark 4.7 (Lower bound and optimality). For the worst-case regret, we achieve $\tilde{\mathcal{O}}(d\sqrt{T})$ (since $\sigma_t = \mathcal{O}(1)$), which matches the minimax lower bound of $\Omega(d\sqrt{T})$ established by Lee & Oh (2024). When the rewards are uniform, i.e., $r_{ti}=1$, we obtain $\tilde{\mathcal{O}}(d\sqrt{T/K})$, as $\sigma_t^2 \simeq p_t(0|S_t,\mathbf{w}^\star) \simeq 1/K$. This result also matches the uniform reward minimax lower bound of $\Omega(d\sqrt{T/K})$ (Lee & Oh, 2024).

Comparison to related works. While our approach shares some similarities with previous works (Faury et al., 2022; Sawarni et al., 2024) that also use a similar warm-up phase, there are significant differences.

Remark 4.8 (Comparison to Faury et al. (2022)). Faury et al. (2022) incurs a $\mathrm{poly}(B)$ dependence in the leading term, whereas our method avoids this entirely by exploiting the self-concordant structure of the MNL loss (see Appendix B). Additionally, their use of MLE in the adaptive warm-up phase results in a per-round computation cost that grows linearly with the number of warm-up rounds. In contrast, our method uses an online update rule, resulting in significantly better computational efficiency. Finally, their approach requires prior knowledge of κ, which is often unknown or hard to estimate in practice.

Remark 4.9 (Comparison to Sawarni et al. (2024)). Unlike Sawarni et al. (2024), which requires prior knowledge of κ—an impractical assumption in real-world scenarios—our approach does not rely on knowing κ in advance. Additionally, their method fully updates parameters using MLE rather than an online update. As a result, the per-round computation cost of their algorithm scales linearly with t, while ours remains constant.

Discussion on instance-dependent regret. As a special

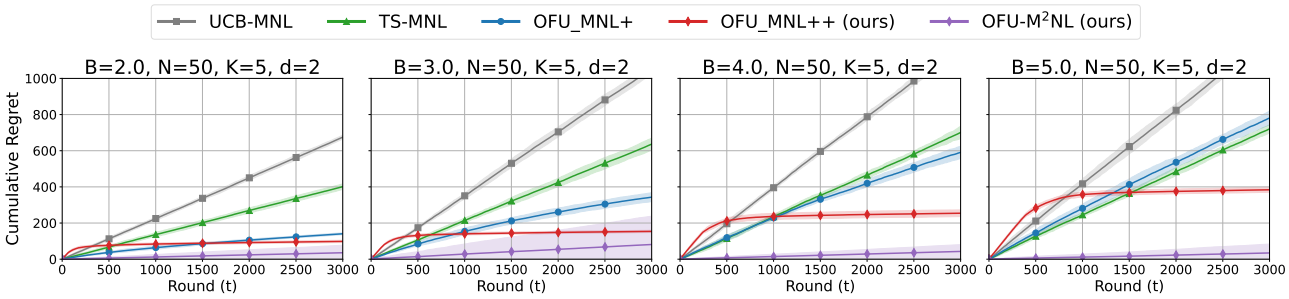

Figure 1: Cumulative regret for varying the norm-boundedness of the unknown parameter B.

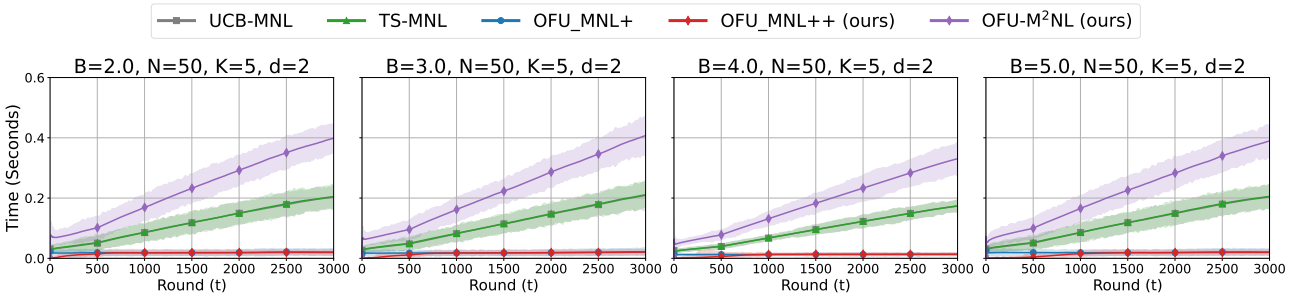

Figure 2: Runtime per round for varying the norm-boundedness of the unknown parameter B.

case, if the rewards are uniform (i.e., $r_{ti} = 1$), we can establish an instance-dependent regret bound.

Proposition 4.10. *Under the same conditions as Theorem 4.5 and assuming uniform rewards, for sufficiently large T,* `OFU-MNL++` *achieves a regret of $\tilde{\mathcal{O}}\left(d\sqrt{\sum_{t=1}^{T}\kappa_t^\star}\right)$, where $\kappa_t^\star := \sum_{i\in S_t^\star} p_t(i|S_t^\star, \mathbf{w}^\star)p_t(0|S_t^\star, \mathbf{w}^\star)$.*

This result improves upon the previous instance-dependent regret of $\tilde{\mathcal{O}}\left(e^B d\sqrt{\sum_{t=1}^{T}\kappa_t^\star}\right)$ (Proposition 2 of Lee & Oh (2024)), by a factor of e^B. The proof and further discussions are provided in Appendix E.

4.3. MLE-Based Approach

Inspired by Lee et al. (2024b), who proposed a $\mathrm{poly}(B)$-free confidence bound using the MLE for generalized linear models (GLM) (but not for MNL models), we introduce an MLE-based algorithm that achieves $\mathrm{poly}(B), K$-free regret. To this end, we first define the MLE estimator $\widehat{\mathbf{w}}_t$ as follows:

$$\widehat{\mathbf{w}}_t := \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_t(\mathbf{w}), \quad \text{where } \mathcal{L}_t(\mathbf{w}) = \sum_{s=1}^{t-1} \ell_s(\mathbf{w}).$$

Lemma 4.11 (Informal, Improved MLE confidence bound)**.** *Let $\mathcal{G}_t = \int_0^1 (1-v)\nabla^2\mathcal{L}_t(\widehat{\mathbf{w}}_t + v(\mathbf{w}^\star - \widehat{\mathbf{w}}_t))\mathrm{d}v + \frac{1}{8B^2}\mathbf{I}_d$. Then, for any $t \geqslant 1$, if Assumption 3.1 holds, then with probability at least $1-\delta$, we have:*

$$\|\mathbf{w}^\star - \widehat{\mathbf{w}}_t\|_{\mathcal{G}_t} = \mathcal{O}\left(\sqrt{d\log(Bt)}\right).$$

Note that $\mathcal{G}_t$ is used solely for analytical purposes. The algorithm and proofs are provided in Appendix F.

Theorem 4.12. *Let $\delta \in (0,1]$. Then, under Assumption 3.1, with probability at least $1-\delta$, the regret of* `OFU-M²NL` *(Algorithm F.1) is bounded as follows:*

$$\mathbf{Reg}_T(\mathbf{w}^\star) \lesssim d\log(BT)\sqrt{\sum_{t=1}^{T}\sigma_t^2 + \frac{1}{\kappa}d^2\left(\log(BT)\right)^2}.$$

Discussion of Theorem 4.12. Theorem 4.12 shows that `OFU-M²NL` enjoys a completely $\mathrm{poly}(B)$-free regret for any T, indicating that its regret is tighter than that of `OFU-MNL++` by a factor of $\mathcal{O}(\mathrm{poly}(B))$ in the non-leading term. However, its asymptotic regret still depends on $\log B$, whereas the asymptotic regret of `OFU-MNL++` remains entirely independent of B. Additionally, the per-round computational cost of `OFU-M²NL` increases linearly with t, while that of `OFU-MNL++` remains constant.

5. Numerical Experiments

We empirically evaluate the performance of our algorithms, `OFU-MNL++` and `OFU-M²NL`, by measuring cumulative regret over $T = 3000$ rounds. The algorithms are

tested on 20 independent instances, and we report the average performance along with a shaded area representing two standard deviations. In each instance, the true underlying parameter $\mathbf{w}^\star$ is uniformly sampled from the d-dimensional ball $\mathbb{B}^d(B)$ of radius B, and the context features x_{ti} are drawn from a $\mathbb{B}^d(1)$. The rewards are sampled from a uniform distribution in each round, i.e., $r_{ti} \sim \mathrm{Unif}(0, 1)$.

The baselines are the practical and state-of-the-art algorithms: the UCB-based algorithm, `UCB-MNL` (Oh & Iyengar, 2019), the Thompson Sampling-based algorithm, `TS-MNL` (Oh & Iyengar, 2019), and the constant-time algorithm, `OFU-MNL+` (Lee & Oh, 2024). Figure 1 shows that both of our algorithms significantly outperform the baseline algorithms. Although `OFU-MNL++` incurs high regret in the early rounds due to the adaptive warm-up phase (with the number of such rounds depending on B), its regret stabilizes after a certain point, exhibiting the lowest slope. Therefore, we believe that `OFU-MNL++` achieves the best asymptotic performance among all algorithms. This aligns with our theoretical results, which show that the asymptotic regret of `OFU-MNL++`, $\mathcal{O}(d \log T\sqrt{T})$, is entirely independent of B (even in logarithmic terms), whereas other algorithms exhibit B-dependence. Additionally, `OFU-M`^{2}`NL` demonstrates the most robust performance, maintaining its superiority even as B increases, particularly in the early rounds. For more details and additional results, refer to Appendix G.

Furthermore, Figure 2 shows that the online update methods (`OFU-MNL+` and `OFU-MNL++`) maintain a constant runtime per round, while the others exhibit a linear increase with t due to their use of MLE-based parameter estimation. Among them, our MLE-based approach, `OFU-M`^{2}`NL`, is the most computationally expensive, as it solves a convex optimization problem to compute the optimistic parameter—unlike the others, which rely on closed-form UCBs (see Line 6 in Algorithm F.1).

6. Conclusion

In this work, we construct the sharper online confidence bound for MNL models, with improvements in terms of $\log K$ and $\mathrm{poly}(B)$ dependencies. Leveraging this result, we propose a constant-time algorithm, `OFU-MNL++`, that achieves B, K-free regret in an asymptotic sense. Additionally, we introduce a MLE-based algorithm, `OFU-M`^{2}`NL`, which ensures $\mathrm{poly}(B), K$-free regret at every round.

Acknowledgments

We sincerely thank Yu-Jie Zhang and Jungyhun Lee for their valuable discussions. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-NR071853 and RS-2023-00222663), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263754), and by AI-Bio Research Grant through Seoul National University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abeille, M., Faury, L., and Calauzènes, C. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

Agrawal, P., Tulabandhula, T., and Avadhanula, V. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 310(2):737–750, 2023.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Thompson sampling for the mnl-bandit. In *Conference on learning theory*, pp. 76–78. PMLR, 2017.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Campolongo, N. and Orabona, F. Temporal variability in implicit online learning. *Advances in neural information processing systems*, 33:12377–12387, 2020.

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment optimization with changing contextual information. *The Journal of Machine Learning Research*, 21(1):8918–8961, 2020.

Cho, W., Hwang, T., Lee, J., and Oh, M.-h. Randomized exploration for reinforcement learning with multinomial logistic function approximation. *arXiv preprint arXiv:2405.20165*, 2024.

Davis, J. M., Gallego, G., and Topaloglu, H. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2):250–273, 2014.

Dong, S., Ma, T., and Van Roy, B. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pp. 1158–1160. PMLR, 2019.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Faury, L., Abeille, M., Jun, K.-S., and Calauzènes, C. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Logistic regression: The importance of being improper. In *Conference on learning theory*, pp. 167–208. PMLR, 2018.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hwang, T. and Oh, M.-h. Model-based reinforcement learning with multinomial logistic function approximation. *arXiv preprint arXiv:2212.13540*, 2022.

Jézéquel, R., Gaillard, P., and Rudi, A. Mixability made efficient: Fast online multiclass logistic regression. *Advances in Neural Information Processing Systems*, 34: 23692–23702, 2021.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Lee, J. and Oh, M.-h. Nearly minimax optimal regret for multinomial logistic bandit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Lee, J. and Oh, M.-h. Combinatorial reinforcement learning with preference feedback. *arXiv preprint arXiv:2502.10158*, 2025.

Lee, J., Yun, S.-Y., and Jun, K.-S. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pp. 4474–4482. PMLR, 2024a.

Lee, J., Yun, S.-Y., and Jun, K.-S. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024b.

Li, L.-F., Zhang, Y.-J., Zhao, P., and Zhou, Z.-H. Provably efficient reinforcement learning with multinomial logit function approximation. *arXiv preprint arXiv:2405.17061*, 2024.

McFadden, D. Modelling the choice of residential location. 1977.

Oh, M.-h. and Iyengar, G. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Oh, M.-h. and Iyengar, G. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9205–9213, 2021.

Ou, M., Li, N., Zhu, S., and Jin, R. Multinomial logit bandit with linear utility functions. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2602–2608. International Joint Conferences on Artificial Intelligence Organization, 2018.

Park, J., Kwon, J., and Lee, D. Infinite-horizon reinforcement learning with multinomial logistic function approximation. *arXiv preprint arXiv:2406.13633*, 2024.

Perivier, N. and Goyal, V. Dynamic pricing and assortment under a contextual mnl demand. *Advances in Neural Information Processing Systems*, 35:3461–3474, 2022.

Rusmevichientong, P., Shen, Z.-J. M., and Shmoys, D. B. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

Sauré, D. and Zeevi, A. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

Sawarni, A., Das, N., Barman, S., and Sinha, G. Generalized linear bandits with limited adaptivity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FTPDBQuT4G.

Tran-Dinh, Q., Li, Y.-H., and Cevher, V. Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences: Proceedings of the 3rd International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences-MCO 2015-Part I*, pp. 155–168. Springer, 2015.

Ville, J. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

Zhang, Y.-J. and Sugiyama, M. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix

A. Notation

Let T be the total number of rounds, with $t \in [T]$ representing the current round. We denote N as the total number of items, K as the maximum size of assortments, d as the dimension of feature vectors, and B as the upper bound on the norm of the unknown parameter. For ease of reference, we provide Table A.1.

Table A.1: Symbols

x_{ti}	feature vector for item i given at round t
r_{ti}	reward for item i given at round t
S_t	assortment chosen by an algorithm at round t
K_t	$:= \lvert S_t \rvert$, size of chosen assortment at round t
0	outside option
y_{ti}	choice response for each item $i \in S_t \cup \{0\}$ at round t
$R_t(S, \mathbf{w}^\star)$	$:= \sum_{i \in S} p_t(i \lvert S, \mathbf{w}^\star) r_{ti}$, expected revenue of the assortment S at round t
$\ell_t(\mathbf{w})$	$:= -\sum_{i \in S_t} y_{ti} \log\left(\frac{\exp(x_{ti}^\top \mathbf{w})}{1+\sum_{j \in S_t} \exp(x_{tj}^\top \mathbf{w})}\right)$, loss function at round t
$\bar{\ell}_t(\mathbf{z}_t)$	$:= -\sum_{i \in S_t} y_{ti} \log\left(\frac{\exp(z_{ti})}{1+\sum_{j \in S_t} \exp(z_{tj})}\right)$, loss function at round t, $z_{ti} = x_{ti}^\top \mathbf{w}$
$\mathcal{T}^w$	set of adaptive warm-up rounds
$\mathbf{w}_t$	online parameter estimate at round t
$\mathbf{w}'_t$	projection of $\mathbf{w}_t$ onto the current search space $\mathcal{W}_t$
$\mathbf{w}_t^w$	adaptive warm-up parameter estimate at round t
η	$:= 1$, step-size parameter for $\mathbf{w}_t$
η^w	$:= \frac{1}{2} + 3\sqrt{2}B$, step-size parameter for $\mathbf{w}_t^w$
λ	$:= 144d$, regularization parameter
λ^w	$:= \max\{12\sqrt{2}\eta^w B, 144\eta^w d, 2\}$ regularization parameter for adaptive warm-up
$\nabla^2 \ell_t(\mathbf{w})$	$= \sum_{i \in S_t} p_t(i \lvert S_t, \mathbf{w}) x_{ti} x_{ti}^\top - \sum_{i \in S_t} \sum_{j \in S_t} p_t(i \lvert S_t, \mathbf{w}) p_t(j \lvert S_t, \mathbf{w}) x_{ti} x_{tj}^\top$
H_t	$:= \lambda \mathbf{I}_d + \sum_{s \notin [t-1] \setminus \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w}_{s+1})$
$\tilde{H}_t$	$:= H_t + \eta \nabla^2 \ell_t(\mathbf{w}'_t) \mathbb{1}(t \notin \mathcal{T}^w)$
$H_t(\mathbf{w}^\star)$	$:= \frac{\lambda}{e} \mathbf{I}_d + \sum_{s \in [t-1] \setminus \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w}^\star)$
H_t^w	$:= \lambda^w \mathbf{I}_d + \sum_{s \in \mathcal{T}^w \setminus [t, \ldots, T]} \nabla^2 \ell_s(\mathbf{w}_{s+1})$
$\beta_t(\delta)$	$:= \mathcal{O}\left(\sqrt{d \log(t/\delta)} + B\sqrt{d}\right)$, confidence radius for $\mathbf{w}_t$ at round t
$\zeta_t(\delta)$	$:= \mathcal{O}\left(B\sqrt{d \log(t/\delta)} + B^{3/2}\sqrt{d} + B^2\right)$, confidence radius for $\mathbf{w}_t^w$ at round t
τ_t	$:= 6\sqrt{2}\zeta_t(\delta)$, threshold for determining whether to implement adaptive warm-up
UCB_{ti}	$:= x_{ti}^\top \mathbf{w}_t + \beta_t(\delta) \lVert x_{ti} \rVert_{H_t^{-1}}$, optimistic utility of item i at round t
$\tilde{R}_t(S)$	$:= \frac{\sum_{i \in S} \exp(\mathrm{UCB}_{ti}) r_{ti}}{1+\sum_{j \in S} \exp(\mathrm{UCB}_{tj})}$, optimistic expected revenue of assortment S at round t
σ_t^2	$:= \mathbb{E}_{i \sim p_t(\cdot \lvert S_t, \mathbf{w}^\star)}\left[\left(r_{ti} - \mathbb{E}_{j \sim p_t(\cdot \lvert S_t, \mathbf{w}^\star)}[r_{tj}]\right)^2\right]$, variance of rewards given S_t at round t

For notational simplicity, we express the loss function in two different forms throughout the proof, using them interchangeably

as needed:

$$\begin{aligned}
\ell_t(\mathbf{w}) &= -\sum_{i\in S_t} y_{ti} \log p_t(i|S_t, \mathbf{w}) = -\sum_{i\in S_t} y_{ti} \log\left(\frac{\exp(x_{ti}^\top \mathbf{w})}{1+\sum_{j\in S_t}\exp(x_{tj}^\top \mathbf{w})}\right),\\
\bar{\ell}_t(\mathbf{z}_t) &= -\sum_{i\in S_t} y_{ti} \log\left(\frac{\exp(z_{ti})}{1+\sum_{j\in S_t}\exp(z_{tj})}\right),\\
\nabla_{\mathbf{w}}\ell_t(\mathbf{w}) &= \sum_{i\in S_t}\left(p_t(i|S_t,\mathbf{w}) - y_{ti}\right)x_{ti},\\
\nabla_{\mathbf{z}}\bar{\ell}_t(\mathbf{z}_t) &= \boldsymbol{\sigma}_t(\mathbf{z}_t^\star) - \mathbf{y}_t,\\
\nabla^2_{\mathbf{w}}\ell_t(\mathbf{w}) &= \sum_{i\in S_t} p_t(i|S_t,\mathbf{w})x_{ti}x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_t(i|S_t,\mathbf{w})p_t(j|S_t,\mathbf{w})x_{ti}x_{tj}^\top,\\
\nabla^2_{\mathbf{z}}\bar{\ell}_t(\mathbf{z}_t) &= \operatorname{diag}(\boldsymbol{\sigma}_t(\mathbf{z}_t^\star)) - \boldsymbol{\sigma}_t(\mathbf{z}_t^\star)\boldsymbol{\sigma}_t(\mathbf{z}_t^\star)^\top,
\end{aligned} \tag{A.1}$$

where $z_{ti} = x_{ti}^\top \mathbf{w}$, $\mathbf{z}_t = (z_{ti})_{i\in S_t} \in \mathbb{R}^{|S_t|}$, and $\mathbf{y}_t = (y_{ti})_{i\in S_t} \in \mathbb{R}^{|S_t|}$. Hence, it is clear that $\ell_t(\mathbf{w}) = \bar{\ell}_t(\mathbf{z}_t)$.

B. Self-Concordant Properties of MNL Function

In this section, we present several key properties of self-concordant-like functions that are essential for proving the main theorems in this paper.

For simplicity, we will work with the MNL loss in the form of $\bar{\ell}$ rather than ℓ throughout this section. However, it is important to note that the properties introduced in this section also apply to ℓ. Whenever these properties are used in the proofs of other lemmas or theorems, we will explicitly demonstrate their applicability to ℓ.

We begin by revisiting the definition of self-concordant-like functions.

Definition B.1 (Self-concordant-like function, Tran-Dinh et al. 2015)**.** A convex function $f \in \mathcal{C}^3 : \mathbb{R}^K \to \mathbb{R}$ is M-self-concordant-like function with constant M if:

$$|\phi'''(s)| \leqslant M\|\mathbf{b}\|_2\phi''(s).$$

for $s \in \mathbb{R}$ and $M > 0$, where $\phi(s) := f(\mathbf{a} + s\mathbf{b})$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$.

To derive a tighter confidence bound in Theorem 4.2 and a tighter regret bound in Theorem 4.5, we redefine the concept of self-concordant-like functions specifically for the *MNL loss function* $\bar{\ell}$.

Definition B.2 (ℓ_∞-norm self-concordant-like MNL loss)**.** The MNL loss function $\bar{\ell}(\mathbf{z}) : \mathbb{R}^K \to \mathbb{R}$ is M-self-concordant-like function with constant M if:

$$|\phi'''(s)| \leqslant M\|\mathbf{b}\|_\infty\phi''(s).$$

for $s \in \mathbb{R}$ and $M > 0$, where $\phi(s) := \bar{\ell}(\mathbf{a} + s\mathbf{b})$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$.

Note that because $\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2$ for any vector $\mathbf{w} \in \mathbb{R}^K$, the new definition of a self-concordant-like function (Definition B.2), which is specifically designed for the MNL loss function, is tighter than the original definition (Definition B.1).

Using this new definition, we show that the MNL loss defined in (1) is a $3\sqrt{2}$-self-concordant-like function.

Proposition B.3 (Constant self-concordant-like MNL loss)**.** *For any $t \in [T]$, the multinomial logistic loss $\bar{\ell}_t$, defined in Equation (A.1), is $3\sqrt{2}$-self-concordant-like function under Definition B.2.*

Proof of Proposition B.3. Recall that the loss $\bar{\ell}_t$ is defined as:

$$\bar{\ell}_t(\mathbf{z}) = \underbrace{-\sum_{i\in S_t} y_{ti}z_{ti}}_{\text{linear}} + \underbrace{\log\left(1 + \sum_{i\in S_t} e^{z_{ti}}\right)}_{=:f(\mathbf{z})}$$

Since $\bar{\ell}_t$ consists of the linear term and $f(\mathbf{z}) : \mathbb{R}^{|S_t|} \to \mathbb{R}$, and the third derivatives of the linear term are zero, it suffices to show that $f(\mathbf{z})$ is a $3\sqrt{2}$-self-concordant-like function.

Fix any $t \in [T]$. For simplicity, let $K = |S_t|$. We define:

$$\phi(s) := f(\mathbf{a} + s\mathbf{b}) = \log\left(1 + \sum_{i=1}^{K} e^{a_i + sb_i}\right) = \log\left(\sum_{i=0}^{K} e^{a_i + sb_i}\right),$$

where $\mathbf{a} = [a_1, \ldots, a_K]^\top \in \mathbb{R}^K$ and $\mathbf{b} = [b_1, \ldots, b_K]^\top \in \mathbb{R}^K$, and $a_0 = b_0 = 0$. Then, by simple calculus, we have

$$\phi''(s) = \frac{\sum_{i<j}(b_i - b_j)^2 e^{a_i + sb_i} e^{a_j + sb_j}}{\left(\sum_{i=0}^{K} e^{a_i + sb_i}\right)^2} \geqslant 0,$$

and

$$\phi'''(s) = \frac{\sum_{i<j}(b_i - b_j)^2 e^{a_i + sb_i} e^{a_j + sb_j} \left[\sum_{k=0}^{K}(b_i + b_j - 2b_k) e^{a_k + sb_k}\right]}{\left(\sum_{i=0}^{K} e^{a_i + sb_i}\right)^3} \leqslant \left|\frac{\sum_{k=0}^{K}(b_i + b_j - 2a_k) e^{a_k + sb_k}}{\sum_{i=0}^{K} e^{a_i + sb_i}}\right| \phi''(s). \tag{B.1}$$

Note that for all $i, j, k = 0, \ldots, K$,

$$|b_i + b_j - 2b_k| \leqslant \sqrt{6}\sqrt{b_i^2 + b_j^2 + b_k^2} \leqslant 3\sqrt{2} \max_{i=0,\ldots,K} |b_i|.$$

Hence, we obtain

$$\left|\sum_{k=0}^{K}(b_i + b_j - 2b_k) e^{a_k + sb_k}\right| \leqslant \sum_{k=0}^{K} |b_i + b_j - 2b_k|\, e^{a_k + sb_k} \leqslant 3\sqrt{2} \max_{i=0,\ldots,K} |b_i| \sum_{i=0}^{K} e^{a_i + sb_i}. \tag{B.2}$$

Plugging in (B.2) into (B.1), we derive that

$$\phi'''(s) \leqslant 3\sqrt{2} \max_{i=0,\ldots,K} |b_i| \phi''(s) = 3\sqrt{2}\|\mathbf{b}\|_\infty \phi''(s).$$

By Definition B.2, we conclude that the MNL loss is $3\sqrt{2}$-self-concordant-like. □

Building on our new definition (Definition B.2), we establish several fundamental properties of the self-concordant-like MNL loss function. The following proposition is analogous to Theorem 3 of Tran-Dinh et al. (2015). However, Proposition B.4 provides a tighter result specifically tailored to the MNL loss function (though it may be extendable to other functions).

Proposition B.4. *For a convex function $f \in \mathcal{C}^3 : \mathbb{R}^K \to \mathbb{R}$, we define $D^3 f(\mathbf{x})[\mathbf{u}, \mathbf{u}, \mathbf{u}] := \langle D^3 f(\mathbf{x})[\mathbf{u}]\mathbf{u}, \mathbf{u}\rangle$. Then, if f is the MNL loss function, i.e., $f = \bar{\ell}$, then for any $\mathbf{x}, \mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^K$, we have:*

$$\left|D^3 f(\mathbf{x})[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_2]\right| \leqslant 3\sqrt{2}\, \|\mathbf{u}_1\|_\infty \|\mathbf{u}_2\|^2_{\nabla^2 f(\mathbf{x})}.$$

Proof of Proposition B.4. Let $\phi(s) = f(\mathbf{a} + s\mathbf{b})$. Then, we have

$$\phi''(s) = \nabla^2 f(\mathbf{a} + s\mathbf{b})\mathbf{b}^\top \mathbf{b}, \quad \phi'''(s) = D^3 f(\mathbf{a} + s\mathbf{b})[\mathbf{b}, \mathbf{b}, \mathbf{b}].$$

By Definition B.2 and Proposition B.3, we know that

$$\left|\phi'''(s)\right| \leqslant 3\sqrt{2}\|\mathbf{b}\|_\infty \phi'''(s).$$

By substituting $s = 0$, $\mathbf{a} = \mathbf{x}$, and $\mathbf{b} = \mathbf{u}_1$, we get

$$\left|D^3 f(\mathbf{x})[\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_1]\right| = \left|\phi'''(0)\right| \leqslant 3\sqrt{2}\|\mathbf{u}_1\|_\infty \phi'''(0) = 3\sqrt{2}\|\mathbf{u}_1\|_\infty \nabla^2 f(\mathbf{x})\mathbf{u}_1^\top \mathbf{u}_1,$$

which can be equivalently expressed as

$$-3\sqrt{2}\|\mathbf{u}_1\|_\infty \nabla^2 f(\mathbf{x}) \preceq D^3 f(\mathbf{x})[\mathbf{u}_1] \preceq 3\sqrt{2}\|\mathbf{u}_1\|_\infty \nabla^2 f(\mathbf{x}).$$

Therefore, for any $\mathbf{u}_2 \in \mathbb{R}^K$, we have

$$\begin{aligned} &\left|\mathbf{u}_2^\top D^3 f(\mathbf{x})[\mathbf{u}_1]\mathbf{u}_2\right| \leqslant 3\sqrt{2}\|\mathbf{u}_1\|_\infty \mathbf{u}_2^\top \nabla^2 f(\mathbf{x})\mathbf{u}_2 \\ \iff &\left|D^3 f(\mathbf{x})[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_2]\right| \leqslant 3\sqrt{2}\,\|\mathbf{u}_1\|_\infty \|\mathbf{u}_2\|^2_{\nabla^2 f(\mathbf{x})}. \end{aligned}$$

This concludes the proof. □

Proposition B.5, a variant of Theorem 4 in Tran-Dinh et al. (2015), establishes a key inequality for the Hessian of the MNL loss, which plays a crucial role in eliminating B-dependency.

Proposition B.5. *For any $t \in [T]$, the Hessian of the multinomial logistic loss $\bar{\ell}_t : \mathbb{R}^{|S_t|} \to \mathbb{R}$ satisfies that, for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{|S_t|}$, we have:*

$$e^{-3\sqrt{2}\|\mathbf{z}_1 - \mathbf{z}_2\|_\infty} \nabla^2 \bar{\ell}_t(\mathbf{z}_1) \preceq \nabla^2 \bar{\ell}_t(\mathbf{z}_2) \preceq e^{3\sqrt{2}\|\mathbf{z}_1 - \mathbf{z}_2\|_\infty} \nabla^2 \bar{\ell}_t(\mathbf{z}_1).$$

Proof of Proposition B.5. We denote $\mathbf{z}_s = \mathbf{z}_1 + s(\mathbf{z}_2 - \mathbf{z}_1)$ for notational convenience, where $s \in [0, 1]$. We define the function $\psi(s) := \mathbf{u}^\top \nabla^2 \bar{\ell}_t(\mathbf{z}_s)\mathbf{u} = \|\mathbf{u}\|^2_{\nabla^2 \bar{\ell}_t(\mathbf{z}_s)}$. Note that $\psi(0) = \|\mathbf{u}\|^2_{\nabla^2 \bar{\ell}_t(\mathbf{z}_1)}$ and $\psi(1) = \|\mathbf{u}\|^2_{\nabla^2 \bar{\ell}_t(\mathbf{z}_2)}$. Then, by Proposition B.4, we have

$$\left|\psi'(s)\right| = \left|D^3 \bar{\ell}_t(\mathbf{z}_s)[\mathbf{z}_2 - \mathbf{z}_1, \mathbf{u}, \mathbf{u}]\right| \leqslant 3\sqrt{2}\|\mathbf{z}_2 - \mathbf{z}_1\|_\infty \psi(s),$$

which can be equivalently written as follows:

$$\left|\frac{d \ln \psi(s)}{ds}\right| \leqslant 3\sqrt{2}\|\mathbf{z}_2 - \mathbf{z}_1\|_\infty.$$

By integrating both sides over $s \in [0, 1]$, we conclude the proof. □

Additionally, we introduce an improved version of Proposition 6 in Perivier & Goyal (2022), which serves as a useful tool for the subsequent proofs.

Proposition B.6. *For any $t \in [T]$, the Hessian of the multinomial logistic loss $\bar{\ell}_t : \mathbb{R}^{|S_t|} \to \mathbb{R}$ satisfies the following for any $\mathbf{u}, \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{|S_t|}$:*

$$\mathbf{u}^\top \left(\int_0^1 (1-s)\nabla^2 \bar{\ell}_t(\mathbf{z}_1 + s(\mathbf{z}_2 - \mathbf{z}_1))\mathrm{d}s\right)\mathbf{u} \geqslant \frac{1}{2 + 3\sqrt{2}\|\mathbf{z}_2 - \mathbf{z}_1\|_\infty} \mathbf{u}^\top \nabla^2 \bar{\ell}_t(\mathbf{z}_1)\mathbf{u}.$$

Proof of Proposition B.6. From Proposition B.5, we have

$$\begin{aligned} \mathbf{u}^\top \left(\int_0^1 (1-s)\nabla^2 \bar{\ell}_t(\mathbf{z}_1 + s(\mathbf{z}_2 - \mathbf{z}_1))\mathrm{d}s\right)\mathbf{u} &\geqslant \mathbf{u}^\top \nabla^2 \bar{\ell}_t(\mathbf{z}_1)\mathbf{u} \int_0^1 (1-s)e^{-3\sqrt{2}\|s(\mathbf{z}_2 - \mathbf{z}_1)\|_\infty}\mathrm{d}s \\ &\geqslant \mathbf{u}^\top \nabla^2 \bar{\ell}_t(\mathbf{z}_1)\mathbf{u} \left(\frac{1}{3\sqrt{2}\|(\mathbf{z}_2 - \mathbf{z}_1)\|_\infty} + \frac{e^{-3\sqrt{2}\|(\mathbf{z}_2 - \mathbf{z}_1)\|_\infty} - 1}{\left(3\sqrt{2}\|(\mathbf{z}_2 - \mathbf{z}_1)\|_\infty\right)^2}\right) \\ &\geqslant \mathbf{u}^\top \nabla^2 \bar{\ell}_t(\mathbf{z}_1)\mathbf{u} \left(\frac{1}{2 + 3\sqrt{2}\|(\mathbf{z}_2 - \mathbf{z}_1)\|_\infty}\right), \end{aligned}$$

where in the third inequality, we use the fact that $\frac{1}{x}\left(1 + \frac{e^{-x}-1}{x}\right) \geqslant \frac{1}{2+x}$ for all $x \geqslant 0$. □

C. Proof of Theorem 4.2

In this section, we provide the proof of Theorem 4.2. We begin with the main proof of the theorem, followed by the proof of the technical lemma that is used within the main argument.

C.1. Main Proof of Theorem 4.2

Proof of Theorem 4.2. The overall proof structure is similar to the analysis presented in Zhang & Sugiyama (2024); Lee & Oh (2024). However, as explained in the main paper, several novel analytical techniques are introduced to derive a B-improved, K-free confidence bound, including:

1. B, K-independent step size η by leveraging improved self-concordant properties,
2. $\mathcal{O}(\sqrt{\log t}\log K)$ via a sharper analysis using Ville's inequality (Ville, 1939).

Throughout the proof of Theorem 4.2, we denote $\mathcal{T} \subseteq [T]$ as the set of total update rounds. For any round $t \in [T]$, we denote $\mathcal{T}_t \subseteq \mathcal{T}$ as the set of update rounds that occur before t, i.e., $\mathcal{T}_t = \{s \in \mathcal{T} : s < t\} = \mathcal{T} \backslash \{t, t+1, \dots, T\}$. We assume the following conditions hold:

Condition C.1 (Update condition). For all $t \in \mathcal{T}$, we assume that

$$\sup_{\mathbf{w} \in \mathcal{W}_t} |x_{ti}^\top (\mathbf{w} - \mathbf{w}^\star)| \leqslant \alpha, \quad \forall i \in S_t,$$

where $\mathcal{W}_t$ is a compact convex set, and $\alpha > 0$.

We also denote the size of the assortment at round t as K_t, i.e., $K_t = |S_t| \leqslant K$.

Lemma C.1. *Suppose Condition C.1 holds. The update rule for the parameter at round $t \in \mathcal{T}$ is defined as:*

$$\mathbf{w}_t' = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}_t} \|\mathbf{w} - \mathbf{w}_t\|_{H_t},$$
$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}_t} \widetilde{\ell}_t(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t'\|_{H_t}^2,$$

where $\widetilde{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t') + \langle \mathbf{w} - \mathbf{w}_t', \nabla \ell_t(\mathbf{w}_t') \rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t'\|_{\nabla^2 \ell_t(\mathbf{w}_t')}^2$. Let $\eta = 1 + \frac{3\sqrt{2}}{2}\alpha$ and $\lambda \geqslant 12\sqrt{2}\eta\alpha$. Then, under Assumption 3.1, for any update round $t \in \mathcal{T}$, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}^\star\|_{H_{t+1}}^2 \leqslant 2\eta \left(\sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}^\star) - \sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}_{s+1}) \right) + 4B^2\lambda - \frac{1}{2} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}_{s+1} - \mathbf{w}_s'\|_{H_s}^2. \tag{C.1}$$

The proof is deferred to Appendix C.2.1. Following the approach of Faury et al. (2022); Zhang & Sugiyama (2024); Lee & Oh (2024), to bound the first term in Equation (C.1), we introduce an intermediary parameter that is $\mathcal{F}_s$-measurable. Note that $\mathbf{w}_{s+1}$ is $\mathcal{F}_s$-measurable.

To do so, we first define the softmax function at round t, denoted as $\boldsymbol{\sigma}_t(\mathbf{z}) : \mathbb{R}^{K_t} \to \mathbb{R}^{K_t}$, as follows:

$$[\boldsymbol{\sigma}_t(\mathbf{z})]_i = \frac{\exp([\mathbf{z}]_i)}{1 + \sum_{k=1}^{K_t} \exp([\mathbf{z}]_k)}, \quad \forall i \in [K_t], \tag{C.2}$$

where $[\cdot]_i$ denotes i'th element of the input vector. The probability of choosing the outside option is denoted as:

$$[\boldsymbol{\sigma}_t(\mathbf{z})]_0 = \frac{1}{1 + \sum_{k=1}^{K_t} \exp([\mathbf{z}]_k)}$$

Although $[\boldsymbol{\sigma}_t(\mathbf{z})]_0$ is not part of the output vector of the softmax function $\boldsymbol{\sigma}_t(\mathbf{z})$, it is expressed in a similar form to (C.2) for simplicity. Then, the MNL user choice model can be equivalently expressed as $p_t(i|S_t, \mathbf{w}) = \left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j \in S_t}\right)\right]_i$ for all $i \in [K_t]$ and $p_t(0|S_t, \mathbf{w}) = \left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j \in S_t}\right)\right]_0$. Furthermore, the loss function in (1) can also be expressed as $\ell(\mathbf{z}_t, \mathbf{y}_t) = \sum_{k=0}^{K_t} \mathbf{1}\{y_{ti} = 1\} \cdot \log\left(\frac{1}{[\boldsymbol{\sigma}_t(\mathbf{z}_t)]_k}\right)$.

We also define a pseudo-inverse function of $\boldsymbol{\sigma}_t(\cdot)$ as $\boldsymbol{\sigma}_t^+ : \mathbb{R}^{K_t} \to \mathbb{R}^{K_t}$, where $[\boldsymbol{\sigma}_t^+(\mathbf{q})]_i = \log\left(q_i/(1 - \|\mathbf{q}\|_1)\right)$ for any $\mathbf{q} \in \{\mathbf{p} \in [0,1]^{K_t} \mid \|\mathbf{p}\|_1 < 1\}$. Then, we define the intermediary parameter as follows:

$$\tilde{\mathbf{z}}_s := \boldsymbol{\sigma}_s^+\left(\mathbb{E}_{\mathbf{w} \sim P_s}\left[\boldsymbol{\sigma}_s\left((x_{sj}^\top \mathbf{w})_{j \in S_s}\right)\right]\right). \tag{C.3}$$

where $P_s := \mathcal{N}\left(\mathbf{w}_s, cH_s^{-1}\right)$ is a multivariate normal distribution with mean $\mathbf{w}_s$ and covariance cH_s^{-1}. Here, $c > 0$ is a positive constant to be specified later. Note that $\tilde{\mathbf{z}}_s$ is $\mathcal{F}_s$-measurable unlike $\mathbf{w}_{s+1}$ ($\mathbf{w}_{s+1}$ is $\mathcal{F}_{s+1}$-measurable). In general, $\tilde{\mathbf{z}}_s$ cannot be expressed as a linear function of the features $\{x_{sj}\}_{j \in S_s}$. Then, the first term in Equation (C.1) can be decomposed into two terms as follows:

$$\underbrace{\sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}^\star) - \sum_{s \in \mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s)}_{(a)} + \underbrace{\sum_{s \in \mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s) - \sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}_{s+1})}_{(b)}$$

First, we demonstrate that the term (a) is bounded by $\mathcal{O}\left(\log \frac{1}{\delta}\right)$ with high probability.

Lemma C.2. *Let $\delta \in (0, 1]$. Assume that Condition C.1 holds. Define the intermediary parameter as Equation (C.3) with and $c > 0$. Then, for any $t \in \mathcal{T}$, with probability at least $1 - \delta$, we have*

$$\sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}^\star) - \sum_{s \in \mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s) \leqslant \log \frac{1}{\delta}.$$

The proof is deferred to Appendix C.2.2. By setting $\frac{1}{\delta} = \mathcal{O}(t)$, we obtain the bound $\sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}^\star) - \sum_{s \in \mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s) = \mathcal{O}(\log t)$. Compared to Lemma F.2 of Lee & Oh (2024), which bound the similar term by $\mathcal{O}\left((\log t)^2 \log K\right)$, Lemma C.2 improves the bound by a factor of $\log t \log K$. This improvement is primarily due to the use of a more refined analysis based on Ville's inequality (Ville, 1939), rather than the Bernstein-type inequality with a *smoothed intermediate* term adopted in Zhang & Sugiyama (2024); Lee & Oh (2024). The latter approach incurs an additional $\log(Kt)$ factor, which ultimately leads to the looser bound of $\mathcal{O}\left((\log t)^2 \log K\right)$ for the term (a).

Now, we bound the term (b) by the following lemma:

Lemma C.3. *Let $c > 0$ and $\lambda \geqslant \max\{2, 72cd\}$. Then, for all $t \in \mathcal{T}$, we have*

$$\sum_{s \in \mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s) - \sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}_{s+1}) \leqslant \frac{1}{2c} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{H_s} + \sqrt{6}cd \log(t+2).$$

The proof is deferred to Appendix C.2.3.

Finally, by combining Lemma C.1, Lemma C.2, and Lemma C.3, we derive that

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}^\star\|^2_{H_{t+1}} \\
&\leqslant 2\eta \log \frac{1}{\delta} + \frac{\eta}{c} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{H_s} + 2\sqrt{6}\eta cd \log(t+2) + 4B^2\lambda - \frac{1}{2} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{H_s} \\
&\leqslant 2\eta \log \frac{1}{\delta} + 4\sqrt{6}\eta^2 d \log(t+2) + 4B^2\lambda && \text{(Set } c = 2\eta\text{)} \\
&= \mathcal{O}\left(\alpha^2 \cdot d \log(t/\delta) + B^2\lambda\right). && \left(\eta = \tfrac{1+3\sqrt{2}\alpha}{2}\right)
\end{aligned}$$

This implies that for all $t \in \mathcal{T} \backslash \{t_1\}$, where t_1 denote the first update round, we have

$$\|\mathbf{w}_t - \mathbf{w}^\star\|_{H_t} \leqslant \mathcal{O}\left(\alpha\sqrt{d \log(t/\delta)} + B\sqrt{\lambda}\right). \tag{C.4}$$

For $t_1 \in \mathcal{T}$, we know that $\|\mathbf{w}_{t_1} - \mathbf{w}^\star\|_{H_t} \leqslant B\sqrt{\lambda}$. Thus, Equation (C.4) holds for all $t \in \mathcal{T}$. This concludes the proof of Theorem 4.2. □

C.2. Proofs of Lemmas for Theorem 4.2

C.2.1. Proof of Lemma C.1

Proof of Lemma C.1. For any update round $s \in \mathcal{T}_t$ (an update round occurring before $t \in \mathcal{T}$), let $\tilde{\ell}_s(\mathbf{w}) = \ell_s(\mathbf{w}'_s) + \langle \nabla \ell_s(\mathbf{w}'_s), \mathbf{w} - \mathbf{w}'_s \rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}'_s\|^2_{\nabla^2 \ell_s(\mathbf{w}'_s)}$ be a second-order approximation of the original function $\ell_s(\mathbf{w})$ at the point $\mathbf{w}'_s$,

where $\mathbf{w}'_s = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}_s} \|\mathbf{w} - \mathbf{w}_s\|_{H_s}$ is the projection of $\mathbf{w}_s$ onto $\mathcal{W}_s$. Then, the update rule in (2) can be equivalently rewritten as follows:

$$\begin{aligned}
\mathbf{w}_{s+1} &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_s} \langle \nabla \ell_t(\mathbf{w}_s), \mathbf{w}\rangle + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}'_s\|^2_{\tilde{H}_s} \\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_s} \langle \nabla \ell_t(\mathbf{w}_s), \mathbf{w}\rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}'_s\|_{\nabla^2\ell_s(\mathbf{w}_s)} + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}'_s\|^2_{H_s} \\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}_s} \tilde{\ell}_s(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}'_s\|^2_{H_s}.
\end{aligned}$$

Then, by applying Lemma C.4, we get

$$\begin{aligned}
\langle \nabla\tilde{\ell}_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle &\leqslant \frac{1}{2\eta}\left(\|\mathbf{w}'_s - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{H_s}\right) \\
&\leqslant \frac{1}{2\eta}\left(\|\mathbf{w}_s - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{H_s}\right), \qquad \text{(C.5)}
\end{aligned}$$

where the last inequality holds due to the nonexpansive property of the projection mapping $P_{\mathcal{W}_s}$, i.e., $\|\mathbf{w}'_s - \mathbf{w}^\star\|^2_{H_s} = \|P_{\mathcal{W}_s}(\mathbf{w}_s) - P_{\mathcal{W}_s}(\mathbf{w}^\star)\|^2_{H_s} \leqslant \|\mathbf{w}_s - \mathbf{w}^\star\|^2_{H_s}$. On the other hand, by applying Lemma C.5, which is based on our improved self-concordant-like property (Proposition B.4), we obtain:

$$\ell_s(\mathbf{w}_{s+1}) - \ell_s(\mathbf{w}^\star) \leqslant \langle \nabla\ell_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle - \frac{1}{2 + 3\sqrt{2}\alpha}\|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}. \qquad \text{(C.6)}$$

Let $\eta = 1 + \frac{3\sqrt{2}}{2}\alpha$. Then, by combining (C.5) and (C.6), we obtain that

$$\begin{aligned}
\ell_s(\mathbf{w}_{s+1}) - \ell_s(\mathbf{w}^\star) &\leqslant \langle \nabla\ell_s(\mathbf{w}_{s+1}) - \nabla\tilde{\ell}_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle \\
&\quad + \frac{1}{2\eta}\left(\|\mathbf{w}_s - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{H_{s+1}} - \|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{H_s}\right). \qquad \text{(C.7)}
\end{aligned}$$

In the inequality above, the first term on the right-hand side can be further bounded as:

$$\begin{aligned}
&\langle \nabla\ell_s(\mathbf{w}_{s+1}) - \nabla\tilde{\ell}_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle \\
&\quad = \langle \nabla\ell_s(\mathbf{w}_{s+1}) - \nabla\ell_s(\mathbf{w}'_s) - \nabla^2\ell_s(\mathbf{w}'_s)(\mathbf{w}_{s+1} - \mathbf{w}'_s), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle \\
&\quad = \langle D^3\ell_s(\bar{\mathbf{w}}_s)[\mathbf{w}_{s+1} - \mathbf{w}'_s](\mathbf{w}_{s+1} - \mathbf{w}'_s), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle \qquad \text{(Taylor expansion)} \\
&\quad = D^3\ell_s(\bar{\mathbf{w}}_s)[\mathbf{w}_{s+1} - \mathbf{w}^\star, \mathbf{w}_{s+1} - \mathbf{w}'_s, \mathbf{w}_{s+1} - \mathbf{w}'_s], \qquad \text{(C.8)}
\end{aligned}$$

where in the second equality, we use the Taylor expansion by introducing $\bar{\mathbf{w}}_s$, which is a convex combination of $\mathbf{w}_{s+1}$ and $\mathbf{w}'_s$. Recall that by the definition of loss (see Equation (A.1)), the loss $\ell_s(\bar{\mathbf{w}}_s)$ can be expressed as $\bar{\ell}_s(\bar{\mathbf{z}}_s)$, where $\bar{\mathbf{z}}_s = (x_{si}^\top\bar{\mathbf{w}}_s)_{i\in S_s} \in \mathbb{R}^{|S_s|}$. Moreover, let $\mathbf{z}_{s+1} = (x_{si}^\top\mathbf{w}_{s+1})_{i\in S_s}$, $\mathbf{z}'_s = (x_{si}^\top\mathbf{w}'_s)_{i\in S_s}$, and $\mathbf{z}^\star = (x_{si}^\top\mathbf{w}^\star)_{i\in S_s}$. Then, by simple calculus, we get

$$\begin{aligned}
D^3\ell_s(\bar{\mathbf{w}}_s)[\mathbf{w}_{s+1} - \mathbf{w}^\star, \mathbf{w}_{s+1} - \mathbf{w}'_s, \mathbf{w}_{s+1} - \mathbf{w}'_s] &= D^3\bar{\ell}_s(\bar{\mathbf{z}}_s)[\mathbf{z}_{s+1} - \mathbf{z}^\star, \mathbf{z}_{s+1} - \mathbf{z}'_s, \mathbf{z}_{s+1} - \mathbf{z}'_s] \\
&\leqslant 3\sqrt{2}\|\mathbf{z}_{s+1} - \mathbf{z}^\star\|_\infty\|\mathbf{z}_{s+1} - \mathbf{z}'_s\|^2_{\nabla^2\bar{\ell}_s(\bar{\mathbf{z}}_s)} \qquad \text{(Proposition B.4)} \\
&\leqslant 3\sqrt{2}\max_{x\in\mathcal{X}_s}|x^\top(\mathbf{w}_{s+1} - \mathbf{w}^\star)|\|\mathbf{z}_{s+1} - \mathbf{z}'_s\|^2_{\nabla^2\bar{\ell}_s(\bar{\mathbf{z}}_s)} \\
&\leqslant 3\sqrt{2}\alpha\|\mathbf{z}_{s+1} - \mathbf{z}'_s\|^2_{\nabla^2\bar{\ell}_s(\bar{\mathbf{z}}_s)} \\
&= 3\sqrt{2}\alpha\|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_{\nabla^2\ell_s(\bar{\mathbf{w}}_s)} \qquad \text{(Definition of } \bar{\ell}) \\
&\leqslant 3\sqrt{2}\alpha\|\mathbf{w}_{s+1} - \mathbf{w}'_s\|^2_2, \qquad \text{(C.9)}
\end{aligned}$$

where the last inequality holds because

$$\begin{aligned}\nabla^2\ell_s(\bar{\mathbf{w}}_s) &= \sum_{i\in S_s} p_s(i|S_s,\bar{\mathbf{w}}_s)x_{si}x_{si}^\top - \sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s,\bar{\mathbf{w}}_s)p_s(j|S_s,\bar{\mathbf{w}}_s)x_{si}x_{sj}^\top \\ &= \sum_{i\in S_s} p_s(i|S_s,\bar{\mathbf{w}}_s)x_{si}x_{si}^\top - \left[\sum_{i\in S_s} p_s(i|S_s,\bar{\mathbf{w}}_s)x_{si}\right]\left[\sum_{i\in S_s} p_s(i|S_s,\bar{\mathbf{w}}_s)x_{si}\right]^\top \\ &\preceq \sum_{i\in S_s} p_s(i|S_s,\bar{\mathbf{w}}_s)x_{si}x_{si}^\top \preceq \mathbf{I}_d. \qquad (\|x_{si}\|_2 \leqslant 1, \text{Assumption 3.1})\end{aligned}$$

Hence, by plugging (C.8) and (C.9) into (C.7), and summing over $s \in \mathcal{T}_{t+1}$, we obtain

$$\begin{aligned}&\sum_{s\in\mathcal{T}_{t+1}} \ell_s(\mathbf{w}_{s+1}) - \sum_{s\in\mathcal{T}_{t+1}} \ell_s(\mathbf{w}^\star) \\ &\quad\leqslant 3\sqrt{2}\alpha \sum_{s\in\mathcal{T}_{t+1}} \|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_2^2 + \frac{1}{2\eta}\sum_{s\in\mathcal{T}_{t+1}}\left(\|\mathbf{w}_s-\mathbf{w}^\star\|_{H_s}^2 - \|\mathbf{w}_{s+1}-\mathbf{w}^\star\|_{H_{s+1}}^2 - \|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_{H_s}^2\right) \\ &\quad= 3\sqrt{2}\alpha \sum_{s\in\mathcal{T}_{t+1}} \|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_2^2 + \frac{1}{2\eta}\left(\|\mathbf{w}_{t_1}-\mathbf{w}^\star\|_{H_{t_1}}^2 - \|\mathbf{w}_{t+1}-\mathbf{w}^\star\|_{H_{t+1}}^2 - \sum_{s\in\mathcal{T}_{t+1}}\|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_{H_s}^2\right),\end{aligned}$$

where in the equality, $t_1 \in \mathcal{T}$ represents the first update round. Additionally, we use the fact that the parameter $\mathbf{w}_t$, and the matrices $\tilde{H}_t$ and H_t remain unchanged during non-update rounds. By rearranging the terms and using the fact that $\|\mathbf{w}_{t_1}-\mathbf{w}^\star\|_{H_{t_1}}^2 \leqslant \lambda\|\mathbf{w}_{t_1}-\mathbf{w}^\star\|_2^2 \leqslant 4B^2\lambda$, we get

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}^\star\|_{H_{t+1}}^2 \\ &\leqslant 2\eta\left(\sum_{s\in\mathcal{T}_{t+1}}\ell_s(\mathbf{w}^\star) - \sum_{s\in\mathcal{T}_{t+1}}\ell_s(\mathbf{w}_{s+1})\right) + 4B^2\lambda - \sum_{s\in\mathcal{T}_{t+1}}\|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_{H_s}^2 + 6\sqrt{2}\eta\alpha\sum_{s\in\mathcal{T}_{t+1}}\|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_2^2 \\ &\leqslant 2\eta\left(\sum_{s\in\mathcal{T}_{t+1}}\ell_s(\mathbf{w}^\star) - \sum_{s\in\mathcal{T}_{t+1}}\ell_s(\mathbf{w}_{s+1})\right) + 4B^2\lambda - \frac{1}{2}\sum_{s\in\mathcal{T}_{t+1}}\|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_{H_s}^2,\end{aligned}$$

where the last inequality holds because, by setting $\lambda \geqslant 12\sqrt{2}\eta\alpha$, we have $6\sqrt{2}\eta\alpha\|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_2^2 \leqslant \frac{1}{2}\|\mathbf{w}_{s+1}-\mathbf{w}'_s\|_{H_s}^2$. □

C.2.2. Proof of Lemma C.2

Proof of Lemma C.2. Recall the definition of $\tilde{\mathbf{z}}_t$ in Equation (C.3). Then, by definition of the pseudo-inverse function $\boldsymbol{\sigma}_t^+$, we have $\boldsymbol{\sigma}_t(\tilde{\mathbf{z}}_t) = \mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top\mathbf{w})_{j\in S_t}\right)\right]$. Let $i_t \in S_t \cup \{0\}$ denote the (random) index such that $y_{ti_t} = 1$; in other words, i_t is the item selected in round t. Then, we can express $\exp\left(\bar{\ell}_t(\tilde{\mathbf{z}}_t)\right)$ as follows:

$$\exp\left(-\bar{\ell}_t(\tilde{\mathbf{z}}_t)\right) = \exp\left(\log\left([\boldsymbol{\sigma}_t(\tilde{\mathbf{z}}_t)]_{i_t}\right)\right) = [\boldsymbol{\sigma}_t(\tilde{\mathbf{z}}_s)]_{i_t} = \left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top\mathbf{w})_{j\in S_t}\right)\right]\right]_{i_t}. \tag{C.10}$$

Similarly, denoting $\mathbf{z}_t^\star = \left(x_{tj}^\top\mathbf{w}^\star\right)_{j\in S_t} \in \mathbb{R}^{K_t}$, we can also express $\exp\left(-\ell_t(\mathbf{w}^\star)\right)$ as follows:

$$\exp\left(-\ell_t(\mathbf{w}^\star)\right) = \exp\left(-\bar{\ell}_t(\mathbf{z}_t^\star)\right) = \exp\left(\log\left([\boldsymbol{\sigma}_t(\mathbf{z}_t^\star)]_{i_t}\right)\right) = [\boldsymbol{\sigma}_t(\mathbf{z}_t^\star)]_{i_t}. \tag{C.11}$$

Here, note that $[\boldsymbol{\sigma}_t(\mathbf{z}_t^\star)]_{i_t} = p_t(i_t|S_t,\mathbf{w}^\star)$.

Now, we define

$$A_t := \exp\left(\sum_{s\in\mathcal{T}_{t+1}}\ell_s(\mathbf{w}^\star) - \sum_{s\in\mathcal{T}_{t+1}}\bar{\ell}_s(\tilde{\mathbf{z}}_s)\right).$$

For completeness, we define $A_0 := 1$. First, we show that $(A_t)_{t\geqslant 0}$ is a nonnegative supermartingale with respect to the filtration $\mathcal{F}_t = \sigma\left(S_1, \mathbf{y}_1, \ldots, S_t\right)$.

If t is an update round, we have:

$$A_t = \frac{\prod_{s\in\mathcal{T}_{t+1}} \exp\left(-\bar{\ell}_s(\tilde{\mathbf{z}}_s)\right)}{\prod_{s\in\mathcal{T}_{t+1}} \exp\left(-\ell_s(\mathbf{w}^\star)\right)} = A_{t-1}\frac{\exp\left(-\bar{\ell}_t(\tilde{\mathbf{z}}_t)\right)}{\exp\left(-\ell_t(\mathbf{w}^\star)\right)} = A_{t-1}\frac{\left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]\right]_{i_t}}{\left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_{i_t}}.$$

(Eqn. (C.10) and (C.11))

Note that the term $\left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]\right]_{i_t} / \left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_{i_t}$ is $\mathcal{F}_{t+1}$-measurable. Further, we get

$$\begin{aligned}
\mathbb{E}\left[A_t \mid \mathcal{F}_t\right] &= A_{t-1}\cdot \mathbb{E}\left[\frac{\left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]\right]_{i_t}}{\left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_{i_t}} \,\middle|\, \mathcal{F}_t\right] \\
&= A_{t-1}\cdot \sum_{i\in S_t\cup\{0\}} p_t(i|S_t,\mathbf{w}^\star)\frac{\left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]\right]_{i}}{\left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_{i}} \\
&= A_{t-1}\cdot \sum_{i\in S_t\cup\{0\}} \left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_i \frac{\left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]\right]_{i}}{\left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_{i}} && \left(\left[\boldsymbol{\sigma}_t\left(\mathbf{z}_t^\star\right)\right]_i = p_t(i|S_t,\mathbf{w}^\star)\right) \\
&= A_{t-1}\cdot \sum_{i\in S_t\cup\{0\}} \left[\mathbb{E}_{\mathbf{w}\sim P_t}\left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]\right]_{i} = A_{t-1}.
\end{aligned}$$

Moreover, if t is not an update round, we simply set $A_t = A_{t-1}$. It follows that $(A_t)_{t\geqslant 0}$ is indeed a martingale, and hence also a supermartingale. Since $(A_t)_{t\geqslant 0}$ is a nonnegative supermartingale, we can apply Ville's inequality (Ville, 1939) to obtain:

$$\begin{aligned}
\mathbb{P}\left(\sum_{s\in\mathcal{T}_{t+1}} \ell_s(\mathbf{w}^\star) - \sum_{s\in\mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s) \geqslant \log\frac{1}{\delta}\right) &= \mathbb{P}\left(A_t \geqslant \frac{1}{\delta}\right) \\
&\leqslant \mathbb{P}\left(\sup_{t\geqslant 0} A_t \geqslant \frac{1}{\delta}\right) \\
&\leqslant \mathbb{E}[A_0]\delta && \text{(Ville's inequality)} \\
&= \delta, && (A_0 = 1)
\end{aligned}$$

which concludes the proof. □

C.2.3. Proof of Lemma C.3

Proof of Lemma C.3. The proof closely follows Lemma F.3 of Lee & Oh (2024) (or Lemma 14 of Zhang & Sugiyama (2024)), with the only difference being that the summation is taken over the subset of rounds $\mathcal{T}_{t+1} \subseteq [t]$, rather than the full set of rounds $[t]$. For completeness, we include the full proof below.

To begin with, the proof builds on an observation from Proposition 2 of Foster et al. (2018), which states that $\tilde{\mathbf{z}}_s$ serves as an aggregation forecaster for the logistic function. Accordingly, for any $s \in \mathcal{T}_{t+1}$, the following holds:

$$\bar{\ell}_s(\tilde{\mathbf{z}}_s) \leqslant -\log\left(\mathbb{E}_{\mathbf{w}\sim P_s}\left[e^{-\ell_s(\mathbf{w})}\right]\right) = -\log\left(\frac{1}{Z_s}\int_{\mathbb{R}^d} e^{-L_s(\mathbf{w})}\mathrm{d}\mathbf{w}\right), \tag{C.12}$$

where $L_s(\mathbf{w}) := \ell_s(\mathbf{w}) + \frac{1}{2c}\|\mathbf{w}-\mathbf{w}'_s\|^2_{H_s}$ and $Z_s := \int_{\mathbb{R}^d} e^{-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}'_s\|^2_{H_s}}\mathrm{d}\mathbf{w}$.

We define the the quadratic approximation of $L_s(\mathbf{w})$ as follows:

$$\tilde{L}_s(\mathbf{w}) := L_s(\mathbf{w}_{s+1}) + \langle \nabla L_s(\mathbf{w}_{s+1}), \mathbf{w}-\mathbf{w}_{s+1}\rangle + \frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{H_s}.$$

Then, by Lemma C.6 and considering the fact that ℓ_s is $3\sqrt{2}$-self-concordant-like function by Proposition B.3, we get

$$L_s(\mathbf{w}) \leqslant \tilde{L}_s(\mathbf{w}) + e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}. \tag{C.13}$$

Furthermore, we define the function $\tilde{f}_{s+1}: \mathcal{W} \to \mathbb{R}$ as

$$\tilde{f}_{s+1}(\mathbf{w}) = \exp\left(-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{H_s} - e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right).$$

Then, we can then derive a lower bound for the expectation in Equation (C.12) as follows:

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}\sim P_s}\left[e^{-\ell_s(\mathbf{w})}\right] &= \frac{1}{Z_s}\int_{\mathbb{R}^d}\exp(-L_s(\mathbf{w}))\mathrm{d}\mathbf{w} \\
&\geqslant \frac{1}{Z_s}\int_{\mathbb{R}^d}\exp(-\tilde{L}_s(\mathbf{w}) - e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla\ell_s(\mathbf{w}_{s+1})})\mathrm{d}\mathbf{w} && \text{(Eqn. (C.13))} \\
&= \frac{\exp(-L_s(\mathbf{w}_{s+1}))}{Z_s}\int_{\mathbb{R}^d}\tilde{f}_{s+1}(\mathbf{w})\cdot\exp(-\langle\nabla L_s(\mathbf{w}_{s+1}), \mathbf{w}-\mathbf{w}_{s+1}\rangle)\mathrm{d}\mathbf{w}. && \text{(Definition of } \tilde{f}_{s+1}(\mathbf{w}))
\end{aligned}$$

Moreover, we define $\tilde{Z}_{s+1} = \int_{\mathbb{R}^d}\tilde{f}_{s+1}(\mathbf{w})\mathrm{d}\mathbf{w} < +\infty$, and denote the distribution whose density function is $\tilde{f}_{s+1}(\mathbf{w})/\tilde{Z}_{s+1}$ as $\tilde{P}_{s+1}$. Then, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}\sim P_s}\left[e^{-\ell_s(\mathbf{w})}\right] &\geqslant \frac{\exp(-L_s(\mathbf{w}_{s+1}))\tilde{Z}_{s+1}}{Z_s}\mathbb{E}_{\mathbf{w}\sim\tilde{P}_{s+1}}\left[\exp(-\langle\nabla L_s(\mathbf{w}_{s+1}), \mathbf{w}-\mathbf{w}_{s+1}\rangle)\right] \\
&\geqslant \frac{\exp(-L_s(\mathbf{w}_{s+1}))\tilde{Z}_{s+1}}{Z_s}\exp\Bigg(-\underbrace{\mathbb{E}_{\mathbf{w}\sim\tilde{P}_{s+1}}\left[\langle\nabla L_s(\mathbf{w}_{s+1}), \mathbf{w}-\mathbf{w}_{s+1}\rangle\right]}_{=0}\Bigg) && \text{(Jensen's inequality)} \\
&= \frac{\exp(-L_s(\mathbf{w}_{s+1}))\tilde{Z}_{s+1}}{Z_s}, && \text{(C.14)}
\end{aligned}$$

where the equality holds because $\tilde{P}_{s+1}$ is symmetric around $\mathbf{w}_{s+1}$. Plugging (C.14) into (C.12), we get

$$\bar{\ell}_s(\tilde{\mathbf{z}}_s) \leqslant \ell_s(\mathbf{w}_{s+1}) + \frac{1}{2c}\|\mathbf{w}'_s-\mathbf{w}_{s+1}\|^2_{H_s} + \log Z_s - \log\tilde{Z}_{s+1}.$$

We can further bound the term $-\log\tilde{Z}_{s+1}$ as follows:

$$\begin{aligned}
-\log\tilde{Z}_{s+1} &= -\log\left(\int_{\mathbb{R}^d}\exp\left(-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{H_s} - e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right)\mathrm{d}\mathbf{w}\right) \\
&= -\log\left(\widehat{Z}_{s+1}\cdot\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\exp\left(-e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right)\right]\right) \\
&\leqslant -\log\widehat{Z}_{s+1} + \mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right], && \text{(Jensen's inequality)}
\end{aligned}$$

where in the second equality, we define $\widehat{P}_{s+1} = \mathcal{N}(\mathbf{w}_{s+1}, cH_s^{-1})$ and $\widehat{Z}_{s+1} := \int_{\mathbb{R}^d} e^{-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{H_s}}\mathrm{d}\mathbf{w}$. Hence, we get

$$\bar{\ell}_s(\tilde{\mathbf{z}}_s) \leqslant \ell_s(\mathbf{w}_{s+1}) + \frac{1}{2c}\|\mathbf{w}'_s-\mathbf{w}_{s+1}\|^2_{H_s} + \log\frac{Z_s}{\widehat{Z}_{s+1}} + \mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right] \tag{C.15}$$

To bound $\frac{Z_s}{\widehat{Z}_{s+1}}$ in Equation (C.15), we have

$$\frac{Z_s}{\widehat{Z}_{s+1}} = \frac{\int_{\mathbb{R}^d} e^{-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}'_s\|^2_{H_s}}\mathrm{d}\mathbf{w}}{\int_{\mathbb{R}^d} e^{-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{H_s}}\mathrm{d}\mathbf{w}} = \frac{\int_{\mathbb{R}^d} e^{-\frac{1}{2c}\|\mathbf{w}\|^2_{H_s}}\mathrm{d}\mathbf{w}}{\int_{\mathbb{R}^d} e^{-\frac{1}{2c}\|\mathbf{w}\|^2_{H_s}}\mathrm{d}\mathbf{w}} = 1,$$

which indicates that

$$\log\frac{Z_s}{\widehat{Z}_{s+1}} = 0. \tag{C.16}$$

Now, we bound the last term in Equation (C.15). Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right] \\
&\qquad\leqslant \underbrace{\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{36\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\right]}}_{\texttt{(b)-1}}\underbrace{\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\|\mathbf{w}-\mathbf{w}_{s+1}\|^4_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right]}}_{\texttt{(b)-2}}. \qquad \text{(C.17)}
\end{aligned}$$

Then, there exist orthogonal bases $\mathbf{e}_1,\ldots,\mathbf{e}_d\in\mathbb{R}^d$ such that $\mathbf{w}-\mathbf{w}_{s+1}$ follows the same distribution as $\widehat{P}_{s+1}$, and can be expressed as:

$$\sum_{j=1}^{d}\sqrt{c\lambda_j\left(H_s^{-1}\right)}X_j\mathbf{e}_j, \quad \text{where } X_j\overset{i.i.d.}{\sim}\mathcal{N}(0,1),\forall j\in[d], \qquad \text{(C.18)}$$

and $\lambda_j\left(H_s^{-1}\right)$ denotes the j-th largest eigenvalue of H_s^{-1}. Now, we bound the term `(b)-1` in (C.17) as follows:

$$\begin{aligned}
\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{36\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\right]} &= \sqrt{\mathbb{E}_{X_j}\left[\prod_{j=1}^{d}e^{36c\lambda_j\left(H_s^{-1}\right)X_j^2}\right]} \\
&\leqslant \sqrt{\prod_{j=1}^{d}\mathbb{E}_{X_j}\left[e^{36c/\lambda X_j^2}\right]} && (c\lambda_j\left(H_s^{-1}\right)\leqslant c/\lambda) \\
&= \left(\mathbb{E}_{X\sim\chi^2}\left[e^{36c/\lambda X}\right]\right)^{\frac{d}{2}} \leqslant \mathbb{E}_{X\sim\chi^2}\left[e^{18cd/\lambda X}\right], && \text{(Jensen's inequality)}
\end{aligned}$$

where χ^2 represents the chi-square distribution. By setting $\lambda\geqslant 72cd$, we get

$$\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{36\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\right]}\leqslant\mathbb{E}_{X\sim\chi^2}\left[e^{\frac{X}{4}}\right]\leqslant\sqrt{2}, \qquad \text{(C.19)}$$

where the last inequality holds because the moment-generating function of the χ^2-distribution satisfies $\mathbb{E}_{X\sim\chi^2}[e^{tX}]\leqslant 1/\sqrt{1-2t}$ for all $t\leqslant 1/2$.

To bound the term `(b)-2` in (C.17), let $M_s=(\nabla^2\ell_s(\mathbf{w}_{s+1}))^{-1/2}H_s(\nabla^2\ell_s(\mathbf{w}_{s+1}))^{-1/2}$ and $\lambda'_j=\lambda_j\left(cM_s^{-1}\right)$ be the j-th largest eigenvalue of the matrix cM_s^{-1}. Then, we have

$$\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\|\mathbf{w}-\mathbf{w}_{s+1}\|^4_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right]}=\sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,cH_s^{-1})}\left[\|\mathbf{w}\|^4_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right]}=\sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,cM_s^{-1})}\left[\|\mathbf{w}\|_2^4\right]}.$$

Furthermore, by performing an analysis similar to that in Equation (C.18), we obtain

$$\begin{aligned}
\sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,cM_s^{-1})}\left[\|\mathbf{w}\|_2^4\right]} &= \sqrt{\mathbb{E}_{X_j\sim\mathcal{N}(0,1)}\left[\left\|\sum_{j=1}^{d}\sqrt{\lambda'_j}X_j\mathbf{e}_j\right\|_2^4\right]} = \sqrt{\mathbb{E}_{X_j\sim\mathcal{N}(0,1)}\left[\left(\sum_{j=1}^{d}\lambda'_jX_j^2\right)^2\right]} \\
&= \sqrt{\sum_{j=1}^{d}\sum_{j'=1}^{d}\lambda'_j\lambda'_{j'}\mathbb{E}_{X_j,X_{j'}\sim\mathcal{N}(0,1)}\left[X_j^2X_{j'}^2\right]} \\
&\leqslant \sqrt{3\sum_{j=1}^{d}\sum_{j'=1}^{d}\lambda'_j\lambda'_{j'}} && (\mathbb{E}_{X_j,X_{j'}\sim\mathcal{N}(0,1)}[X_j^2X_{j'}^2]\leqslant 3,\forall j,j'\in[d]) \\
&= \sqrt{3}c\,\mathrm{Tr}\left(M_s^{-1}\right). && (\textstyle\sum_{j=1}^{d}\lambda'_j=\mathrm{Tr}\left(cM_s^{-1}\right))
\end{aligned}$$

Here, $\mathrm{Tr}(A)$ denotes the trace of the matrix A.

Define the matrix $Q_{s+1} := \frac{\lambda}{2}\mathbf{I}_d + \sum_{s' \in \mathcal{T}_{s+1}} \nabla^2 \ell_{s'}(\mathbf{w}_{s'+1})$. By setting $\lambda \geqslant 2$, we can ensure that $\nabla^2 \ell_s(\mathbf{w}_{s+1}) \preceq \mathbf{I}_d \leqslant \frac{\lambda}{2}\mathbf{I}_d$. As a result, we have $Q_{s+1} \preceq \lambda \mathbf{I}_d + \sum_{s' \in \mathcal{T}_s} \nabla^2 \ell_{s'}(\mathbf{w}_{s'+1}) = H_s$. Using this relationship, we can bound the trace as follows:

$$\begin{aligned}\operatorname{Tr}\left(M_s^{-1}\right) &= \operatorname{Tr}\left(H_s^{-1}\nabla^2 \ell_s(\mathbf{w}_{s+1})\right) \leqslant \operatorname{Tr}\left(Q_{s+1}^{-1}\nabla^2 \ell_s(\mathbf{w}_{s+1})\right) \\ &= \operatorname{Tr}\left(Q_{s+1}^{-1}\left(Q_{s+1} - Q_s\right)\right) \leqslant \log \frac{\det(Q_{s+1})}{\det(Q_s)},\end{aligned}$$

where in the last inequality, we apply Lemma 4.5 of Hazan et al. (2016). Hence, we get

$$\sqrt{\mathbb{E}_{\mathbf{w} \sim \widehat{P}_{s+1}}\left[\|\mathbf{w} - \mathbf{w}_{s+1}\|^4_{\nabla^2 \ell_s(\mathbf{w}_{s+1})}\right]} \leqslant \sqrt{3}c \log \frac{\det(Q_{s+1})}{\det(Q_s)}. \tag{C.20}$$

By substituting (C.19) and (C.20) into (C.17), combining the result with (C.15) and (C.16), and summing over $s \in \mathcal{T}_{t+1}$, we obtain

$$\begin{aligned}\sum_{s \in \mathcal{T}_{t+1}} \bar{\ell}_s(\tilde{\mathbf{z}}_s) - \sum_{s \in \mathcal{T}_{t+1}} \ell_s(\mathbf{w}_{s+1}) &\leqslant \frac{1}{2c} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}'_s - \mathbf{w}_{s+1}\|^2_{H_s} + \sqrt{6}c \sum_{s \in \mathcal{T}_{t+1}} \log \frac{\det(Q_{s+1})}{\det(Q_s)} \\ &= \frac{1}{2c} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}'_s - \mathbf{w}_{s+1}\|^2_{H_s} + \sqrt{6}c \log \frac{\det(Q_{t+1})}{\det\left(\frac{\lambda}{2}\mathbf{I}_d\right)} \\ &\leqslant \frac{1}{2c} \sum_{s \in \mathcal{T}_{t+1}} \|\mathbf{w}'_s - \mathbf{w}_{s+1}\|^2_{H_s} + \sqrt{6}cd \log\left(t+2\right).\end{aligned}$$

This concludes the proof. □

C.3. Technical Lemmas for Theorem 4.2

Lemma C.4 (Proposition 4.1 of Campolongo & Orabona 2020)**.** *Let the $\mathbf{w}_{t+1}$ be the solution of the update rule*

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \mathcal{V}} \eta_t \ell_t(\mathbf{w}) + D_\psi(\mathbf{w}, \mathbf{w}_t),$$

where $\mathcal{V} \subseteq \mathcal{W} \subseteq \mathbb{R}^d$ is a non-empty convex set and $D_\psi(\mathbf{w}_1, \mathbf{w}_2) = \psi(\mathbf{w}_1) - \psi(\mathbf{w}_2) - \langle \nabla\psi(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle$ is the Bregman Divergence w.r.t. a strictly convex and continuously differentiable function $\psi : \mathcal{W} \to \mathbb{R}$. Further supposing $\psi(\mathbf{w})$ is 1-strongly convex w.r.t. a certain norm $\|\cdot\|$ in $\mathcal{W}$, then there exists a $\mathbf{g}'_t \in \partial \ell_t(\mathbf{w}_{t+1})$ such that

$$\langle \eta_t \mathbf{g}'_t, \mathbf{w}_{t+1} - \mathbf{u}\rangle \leqslant \langle \nabla\psi(\mathbf{w}_t) - \nabla\psi(\mathbf{w}_{t+1}), \mathbf{w}_{t+1} - \mathbf{u}\rangle$$

for any $\mathbf{u} \in \mathcal{W}$.

Lemma C.5. *For any $t \in [T]$ and $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ such that $\max_{i \in S_t} |x_{ti}^\top(\mathbf{w} - \mathbf{w}')| \leqslant \alpha$, the multinomial logistic loss $\ell_t : \mathbb{R}^d \to \mathbb{R}$, defined in (1), satisfies the following property:*

$$\ell_t(\mathbf{w}) \geqslant \ell_t(\mathbf{w}') + \nabla\ell_t(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}') + \frac{1}{2 + 3\sqrt{2}\alpha}(\mathbf{w} - \mathbf{w}')^\top \nabla^2 \ell_t(\mathbf{w}')(\mathbf{w} - \mathbf{w}').$$

Proof of Lemma C.5. Recall that by definition (see Equation (A.1)), the loss $\ell_t(\mathbf{w})$ can be rewritten as $\bar{\ell}_t(\mathbf{z}_t)$, where $\mathbf{z}_t = (x_{ti}^\top \mathbf{w})_{i \in S_t} \in \mathbb{R}^{|S_t|}$. Similarly, $\ell_t(\mathbf{w}') = \bar{\ell}_t(\mathbf{z}'_t)$. Then, by a second order Taylor expansion, we have

$$\begin{aligned}\bar{\ell}_t(\mathbf{z}_t) &= \bar{\ell}_t(\mathbf{z}'_t) + \nabla\bar{\ell}_t(\mathbf{z}'_t)^\top\left(\mathbf{z}_t - \mathbf{z}'_t\right) + \left(\mathbf{z}_t - \mathbf{z}'_t\right)^\top \left(\int_0^1 (1-s)\nabla^2 \bar{\ell}_t(\mathbf{z}'_t + s(\mathbf{z}_t - \mathbf{z}'_t))\mathrm{d}s\right)\left(\mathbf{z}_t - \mathbf{z}'_t\right) \\ &\geqslant \bar{\ell}_t(\mathbf{z}'_t) + \nabla\bar{\ell}_t(\mathbf{z}'_t)^\top\left(\mathbf{z}_t - \mathbf{z}'_t\right) + \frac{1}{2 + 3\sqrt{2}\|\mathbf{z}_t - \mathbf{z}'_t\|_\infty}\left(\mathbf{z}_t - \mathbf{z}'_t\right)^\top \nabla^2 \bar{\ell}_t(\mathbf{z}'_t)\left(\mathbf{z}_t - \mathbf{z}'_t\right),\end{aligned} \tag{C.21}$$

where the inequality holds by Proposition B.6. Moreover, by definition, we know that

$$\begin{aligned}\nabla\bar{\ell}_t(\mathbf{z}'_t)^\top\left(\mathbf{z}_t - \mathbf{z}'_t\right) &= \nabla\ell_t(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}'), \\ \text{and } \left(\mathbf{z}_t - \mathbf{z}'_t\right)^\top \nabla^2 \bar{\ell}_t(\mathbf{z}'_t)\left(\mathbf{z}_t - \mathbf{z}'_t\right) &= (\mathbf{w} - \mathbf{w}')^\top \nabla^2 \ell_t(\mathbf{w}')(\mathbf{w} - \mathbf{w}').\end{aligned}$$

Hence, we can rewrite Equation (C.21) equivalently as follows:

$$\begin{aligned}\ell_t(\mathbf{w}) &\geqslant \ell_t(\mathbf{w}') + \nabla\ell_t(\mathbf{w}')^\top(\mathbf{w}-\mathbf{w}') + \frac{1}{2+3\sqrt{2}\max_{i\in S_t}|x_{ti}^\top(\mathbf{w}-\mathbf{w}')|}(\mathbf{w}-\mathbf{w}')^\top\nabla^2\ell_t(\mathbf{w}')(\mathbf{w}-\mathbf{w}') \\ &\geqslant \ell_t(\mathbf{w}') + \nabla\ell_t(\mathbf{w}')^\top(\mathbf{w}-\mathbf{w}') + \frac{1}{2+3\sqrt{2}\alpha}(\mathbf{w}-\mathbf{w}')^\top\nabla^2\ell_t(\mathbf{w}')(\mathbf{w}-\mathbf{w}'),\end{aligned}$$

which concludes the proof. □

Lemma C.6 (Lemma 18 of Zhang & Sugiyama 2024)**.** *For any $H_t \succeq 0$, let $L_t(\mathbf{w}) = \ell_t(\mathbf{w}) + \frac{1}{2c}\|\mathbf{w}-\mathbf{w}_t\|_{H_t}^2$. Assume that ℓ_t is a M-self-concordant-like function. Then, for any $\mathbf{w}, \mathbf{w}_t \in \mathcal{W}$, the quadratic approximation $\tilde{L}_t(\mathbf{w}) = L_t(\mathbf{w}_{t+1}) + \langle\nabla L_t(\mathbf{w}_{t+1}), \mathbf{w}-\mathbf{w}_{t+1}\rangle + \frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{t+1}\|_{H_t}^2$ satisfies*

$$L_t(\mathbf{w}) \leqslant \tilde{L}_t(\mathbf{w}) + e^{M^2\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_{\nabla\ell_t(\mathbf{w}_{t+1})}^2.$$

D. Proof of Theorem 4.5

In this section, we present the proof of Theorem 4.5. To begin, we define a set of adaptive warm-up rounds as follows:

$$\mathcal{T}^w := \left\{t \in [T] : \max_{x\in\mathcal{X}_t}\|x\|^2_{(H_t^w)^{-1}} \geqslant 1/\tau_t^2\right\}, \tag{D.1}$$

where we define the threshold τ_t as:

$$\tau_t := 6\sqrt{2}\zeta_t(\delta) = \mathcal{O}\left(B\sqrt{d\log(t/\delta)} + B^{3/2}\sqrt{d} + B^2\right).$$

Moreover, we define the following two confidence sets for all $t \in [T]$:

$$\begin{aligned}\mathcal{W}_t^w(\delta) &:= \left\{\mathbf{w}\in\mathbb{R}^d \mid \|\mathbf{w}-\mathbf{w}_t^w\|_{H_t^w} \leqslant \zeta_t(\delta)\right\}, \quad \text{and} \\ \mathcal{C}_t(\delta) &:= \left\{\mathbf{w}\in\mathbb{R}^d \mid \|\mathbf{w}-\mathbf{w}_t\|_{H_t} \leqslant \beta_t(\delta)\right\},\end{aligned} \tag{D.2}$$

where

$$\begin{aligned}\zeta_t(\delta) &:= \sqrt{2\eta^w\log\frac{1}{\delta} + 4\sqrt{6}(\eta^w)^2 d\log(t+2) + 4B^2\lambda^w} \\ &= \mathcal{O}\left(B\sqrt{d\log(t/\delta)} + B^{3/2}\sqrt{d} + B^2\right), \quad (\text{set } \alpha = 2B,\ \eta^w = \tfrac{1}{2}(1+3\sqrt{2}\alpha),\ \lambda^w = \max\{12\sqrt{2}\eta^w\alpha, 144\eta^w d, 2\})\end{aligned}$$

and

$$\begin{aligned}\beta_t(\delta) &:= \sqrt{2\eta\log\frac{1}{\delta} + 4\sqrt{6}\eta^2 d\log(t+2) + 4B^2\lambda} \\ &= \mathcal{O}\left(\sqrt{d\log(t/\delta)} + B\sqrt{d}\right). \qquad (\text{set } \alpha = \tfrac{1}{3\sqrt{2}},\ \eta = 1,\ \lambda = \max\{12\sqrt{2}\eta\alpha, 144\eta d, 2\} = 144d)\end{aligned}$$

Then, the true parameter $\mathbf{w}^\star$ lies within both confidence sets with high probability.

Corollary D.1 (Confidence set for adaptive warm-up)**.** *Let $\delta \in (0,1]$. We set $\eta^w = \frac{1}{2}(1+3\sqrt{2}B)$ and $\lambda^w = \max\{12\sqrt{2}\eta^w\alpha, 144\eta^w d, 2\}$. Then, we have*

$$\Pr[\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{W}_t^w(\delta)] \geqslant 1-\delta.$$

The proof can be found in Appendix D.2.1.

Corollary D.2 (Restatement of Corollary 4.4, Confidence set for planning & learning)**.** *Let $\delta \in (0,1]$. We set $\eta = 1$, $\lambda = 144d$, and $\tau_t = 6\sqrt{2}\zeta_t(\delta) = \mathcal{O}\left(B\sqrt{d\log(t/\delta)} + B^{3/2}\sqrt{d} + B^2\right)$. Then, if $\mathbf{w}^\star \in \mathcal{W}_t^w(\delta)$ for all $t \geqslant 1$, we have*

$$\Pr\left[\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{C}_t(\delta)\right] \geqslant 1-\delta.$$

The proof is provided in Appendix D.2.2.

Furthermore, we introduce several useful lemmas. Lemma D.3 shows that UCB_{ti} provides an optimistic estimate of the true utility.

Lemma D.3 (Lemma E.1 of Lee & Oh 2024)**.** *Let* $UCB_{ti} = x_{ti}^\top \mathbf{w}_t + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$. *Assume that* $\mathbf{w}^\star \in \mathcal{C}_t(\delta)$, *where* $\mathcal{C}_t(\delta) := \{\mathbf{w} \in \mathcal{W} \mid \|\mathbf{w}_t - \mathbf{w}\|_{H_t} \leqslant \beta_t(\delta)\}$. *Then, we have*

$$0 \leqslant UCB_{ti} - x_{ti}^\top \mathbf{w}^\star \leqslant 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}.$$

Lemma D.4 shows that $\tilde{R}_t(S_t)$, defined in (8), is an upper bound of the true expected revenue of the optimal assortment, $R_t(S_t^\star, \mathbf{w}^\star)$.

Lemma D.4 (Optimism, Lemma 4 of Oh & Iyengar 2021)**.** *Let* $\tilde{R}_t(S) = \frac{\sum_{i\in S}\exp(UCB_{ti})r_{ti}}{1+\sum_{j\in S}\exp(UCB_{tj})}$. *And suppose* $S_t = \operatorname{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$. *If for every item* $i \in S_t^\star$, $UCB_{ti} \geqslant x_{ti}^\top \mathbf{w}^\star$, *then for all* $t \geqslant 1$, *the following inequalities hold:*

$$R_t(S_t^\star, \mathbf{w}^\star) \leqslant \tilde{R}_t(S_t^\star) \leqslant \tilde{R}_t(S_t).$$

It is important to note that Lemma D.4 does not assert that the expected revenue is a monotonic function in general. Rather, it specifically states that the expected revenue associated with the "optimal" assortment increases as the MNL parameters increase (Agrawal et al., 2019; Oh & Iyengar, 2021; Lee & Oh, 2024).

Lemma D.5 shows that $\tilde{R}_t(S_t)$ increases as the utility values of the items in S_t further grow.

Lemma D.5 (Overly optimism, Lemma H.2 of Lee & Oh 2024)**.** *We define* $\tilde{R}_t(S) := \frac{\sum_{i\in S}\exp(UCB_{ti})r_{ti}}{1+\sum_{j\in S}\exp(UCB_{tj})}$ *and* $S_t = \operatorname{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$. *Assume* $\overline{UCB}_{ti} \geqslant UCB_{ti} \geqslant 0$ *for all* $i \in [N]$. *Then, we have*

$$\tilde{R}_t(S_t) \leqslant \frac{\sum_{i\in S_t}\exp(\overline{UCB}_{ti})r_{ti}}{1+\sum_{j\in S_t}\exp(\overline{UCB}_{tj})}.$$

Moreover, we demonstrate that the rewards for the chosen assortment, r_{ti} for all $i \in S_t$ satisfy the condition $R_t(S_t, \mathbf{w}^\star)$.

Lemma D.6. *For all round* $t \in [T]$, *we have*

$$r_{ti} \geqslant R_t(S_t, \mathbf{w}^\star), \quad \forall i \in S_t.$$

The proof is provided in Appendix D.2.3.

We introduce an elliptical potential lemma that will be used in our proof.

Lemma D.7 (Elliptical potential lemma)**.** *Define* $H_t(\mathbf{w}) := \lambda\mathbf{I}_d + \sum_{s\in[t-1]\setminus\mathcal{T}^w} \nabla^2\ell_s(\mathbf{w})$. *If* $\|x_{si}\|^2_{H_s(\mathbf{w})^{-1}} \leqslant \frac{1}{2}$ *for all* $i \in S_s$ *and* $s \in [t]\setminus\mathcal{T}^w$, *then we have*

$$\sum_{s\in[t]\setminus\mathcal{T}^w}\sum_{i\in S_s\cup\{0\}} p_s(i|S_s,\mathbf{w})\left\|x_{si} - \mathbb{E}_{j\sim p_s(\cdot|S_s,\mathbf{w})}[x_{sj}]\right\|^2_{(H_s(\mathbf{w}))^{-1}} \leqslant 2d\log\left(1+\frac{t}{d\lambda}\right).$$

The proof is deferred to Appendix D.2.4.

Lemma D.8 shows that H_t and $H_t(\mathbf{w}^\star)$ remain similar when updated only for $t \notin \mathcal{T}^w$.

Lemma D.8. *Let* $H_t = \lambda\mathbf{I}_d + \sum_{s\in[t-1]\setminus\mathcal{T}^w} \nabla^2\ell_s(\mathbf{w}_{s+1})$ *and* $H_t(\mathbf{w}^\star) = \frac{\lambda}{e}\mathbf{I}_d + \sum_{s\in[t-1]\setminus\mathcal{T}^w} \nabla^2\ell_s(\mathbf{w}^\star)$. *Then, we have*

$$\frac{1}{e}H_t(\mathbf{w}^\star) \preceq H_t \preceq eH_t(\mathbf{w}^\star).$$

The proof is provided in Appendix D.2.5.

Additionally, we present a useful lemma that will be employed to bound the second-order term of the regret.

Lemma D.9 (Lemma E.3 of Lee & Oh 2024)**.** *Define* $Q : \mathbb{R}^K \to \mathbb{R}$*, such that for any* $\mathbf{u} = (u_1, \ldots, u_K) \in \mathbb{R}^K$*,* $Q(\mathbf{u}) = \sum_{i=1}^K \frac{\exp(u_i)}{1+\sum_{k=1}^K \exp(u_k)}$*. Let* $p_i(\mathbf{u}) = \frac{\exp(u_i)}{1+\sum_{k=1}^K \exp(u_k)}$*. Then, for all* $i \in [K]$*, we have*

$$\left|\frac{\partial^2 Q}{\partial i \partial j}\right| \leqslant \begin{cases} 3p_i(\mathbf{u}) & \text{if } i = j, \\ 2p_i(\mathbf{u})p_j(\mathbf{u}) & \text{if } i \neq j. \end{cases}$$

The size of the set $\mathcal{T}^w$ is bounded as described in the following lemma:

Lemma D.10. *The size of the set* $\mathcal{T}^w$*, defined in Equation (D.1), is bounded as follows:*

$$|\mathcal{T}^w| \leqslant \frac{2}{\kappa}\tau_T^2 d \log\left(1 + \frac{T}{d\lambda}\right).$$

The proof is deferred to Appendix D.2.6.

We are now ready to provide the proof of Theorem 4.5.

D.1. Main Proof of Theorem 4.5

Proof of Theorem 4.5. Throughout the proof of the theorem, assume the following event holds:

$$\{\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{W}_t^w(\delta)\} \bigcup \{\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{C}_t(\delta)\}, \tag{D.3}$$

which occurs with a probability of at least $1 - 2\delta$ by Corollary D.1 and D.2.

From the definition of $\mathcal{T}^w$ (see Equation (D.1)), we decompose the regret as follows:

$$\begin{aligned}
\mathbf{Reg}_T(\mathbf{w}^\star) &= \sum_{t=1}^T R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) \\
&= \sum_{t \in \mathcal{T}^w} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) + \sum_{t \notin \mathcal{T}^w} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) \\
&\leqslant |\mathcal{T}^w| + \sum_{t \notin \mathcal{T}^w} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) \qquad (R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) \leqslant 1) \\
&\leqslant \frac{2}{\kappa}\tau_T^2 d \log\left(1 + \frac{T}{d\lambda}\right) + \sum_{t \notin \mathcal{T}^w} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star), \qquad \text{(D.4)}
\end{aligned}$$

where the last inequality holds by Lemma D.10. Next, we concentrate on deriving a bound for the last term. We define $\overline{\text{UCB}}_{ti}$ as $\overline{\text{UCB}}_{ti} := x_{ti}^\top \mathbf{w}^\star + 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$. Under the event in Equation (D.3), by Lemma D.3, we have

$$\text{UCB}_{ti} \leqslant x_{ti}^\top \mathbf{w}^\star + 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}} =: \overline{\text{UCB}}_{ti}.$$

Then, we define the *overly optimistic* expected revenue, $\tilde{\tilde{R}}_t(S_t)$, as

$$\tilde{\tilde{R}}_t(S_t) := \frac{\sum_{i \in S_t} \exp(\overline{\text{UCB}}_{ti}) r_{ti}}{1 + \sum_{j \in S_t} \exp(\overline{\text{UCB}}_{tj})}.$$

Using this definition and applying the optimism lemmas, we can derive an upper bound for the regret as follows:

$$\begin{aligned}
\sum_{t \notin \mathcal{T}^w} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) &\leqslant \sum_{t \notin \mathcal{T}^w} \tilde{R}_t(S_t) - R_t(S_t, \mathbf{w}^\star) && \text{(Lemma D.4)} \\
&\leqslant \sum_{t \notin \mathcal{T}^w} \tilde{\tilde{R}}_t(S_t) - R_t(S_t, \mathbf{w}^\star). && \text{(Lemma D.5)}
\end{aligned}$$

Now, we define a function $\tilde{Q} : \mathbb{R}^{|S_t|} \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1, \ldots, u_{|S_t|})^\top \in \mathbb{R}^{|S_t|}$, $\tilde{Q}(\mathbf{u}) = \sum_{k=1}^{|S_t|} \frac{\exp(u_k) r_{ti_k}}{1+\sum_{j=1}^{|S_t|} \exp(u_j)}$. Here, we denote $S_t = \{i_1, \ldots, i_{|S_t|}\}$ for simplicity. Additionally, let $\mathbf{u}_t = (u_{ti_1}, \ldots u_{ti_{|S_t|}})^\top = (\overline{\text{UCB}}_{ti_1}, \ldots, \overline{\text{UCB}}_{ti_{|S_t|}})^\top$ and $\mathbf{u}_t^\star = (u_{ti_1}^\star, \ldots u_{ti_{|S_t|}}^\star)^\top = (x_{ti_1}^\top \mathbf{w}^\star, \ldots, x_{ti_{|S_t|}}^\top \mathbf{w}^\star)^\top$.

Then, by a second order Taylor expansion, we derive

$$\sum_{t\notin\mathcal{T}^w}\tilde{\bar{R}}_t(S_t)-R_t(S_t,\mathbf{w}^\star)=\sum_{t\notin\mathcal{T}^w}\tilde{Q}(\mathbf{u}_t)-\tilde{Q}(\mathbf{u}_t^\star)$$
$$=\underbrace{\sum_{t\notin\mathcal{T}^w}\nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t-\mathbf{u}_t^\star)}_{I_1}+\underbrace{\frac{1}{2}\sum_{t\notin\mathcal{T}^w}(\mathbf{u}_t-\mathbf{u}_t^\star)^\top\nabla^2\tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t-\mathbf{u}_t^\star)}_{I_2},\tag{D.5}$$

where $\bar{\mathbf{u}}_t=(\bar{u}_{ti_1},\dots,\bar{u}_{ti_{|S_t|}})^\top\in\mathbb{R}^{|S_t|}$ is the convex combination of $\mathbf{u}_t$ and $\mathbf{u}_t^\star$.

First, we bound the term I_1.

$$\begin{aligned}
&\sum_{t\notin\mathcal{T}^w}\nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t-\mathbf{u}_t^\star)\\
&=\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t}\frac{\exp(x_{ti}^\top\mathbf{w}^\star)r_{ti}}{1+\sum_{k\in S_t}\exp(x_{tk}^\top\mathbf{w}^\star)}(u_{ti}-u_{ti}^\star)-\sum_{j\in S_t}\frac{\exp(x_{tj}^\top\mathbf{w}^\star)r_{tj}\sum_{i\in S_t}\exp(x_{ti}^\top\mathbf{w}^\star)}{(1+\sum_{k\in S_t}\exp(x_{tk}^\top\mathbf{w}^\star))^2}(u_{ti}-u_{ti}^\star)\\
&=\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}(u_{ti}-u_{ti}^\star)-\sum_{i\in S_t}\sum_{j\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}p_t(j|S_t,\mathbf{w}^\star)(u_{tj}-u_{tj}^\star)\\
&=\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left((u_{ti}-u_{ti}^\star)-\sum_{j\in S_t}p_t(j|S_t,\mathbf{w}^\star)(u_{tj}-u_{tj}^\star)\right)\\
&=\sum_{t\notin\mathcal{T}^w}2\beta_t(\delta)\underbrace{\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}}-\sum_{j\in S_t}p_t(j|S_t,\mathbf{w}^\star)\|x_{tj}\|_{H_t^{-1}}\right)}_{\geqslant 0}\\
&\leqslant 2\beta_T(\delta)\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}}-\sum_{j\in S_t}p_t(j|S_t,\mathbf{w}^\star)\|x_{tj}\|_{H_t^{-1}}\right),\qquad(\beta_t(\delta)\text{ is non-decreasing})
\end{aligned}$$

where in the last inequality, we use the fact that $\beta_t(\delta)$ is non-decreasing and that the following holds:

$$\begin{aligned}
&\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}}-\sum_{j\in S_t}p_t(j|S_t,\mathbf{w}^\star)\|x_{tj}\|_{H_t^{-1}}\right)\\
&\quad=\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}}\Bigg(r_{ti}-\underbrace{\sum_{j\in S_t}p_t(j|S_t,\mathbf{w}^\star)r_{tj}}_{=R_t(S_t,\mathbf{w}^\star)}\Bigg)\geqslant 0.\qquad(\text{Lemma D.6})
\end{aligned}$$

Let $x_{t0}=\mathbf{0}$ and $r_{t0}=0$. For simplicity, we denote $\mathbb{E}_t^{\mathbf{w}}[x_{ti}]=\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w})}[x_{ti}]$, and $\mathbb{E}_t^{\mathbf{w}}[r_{ti}]=\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w})}[r_{ti}]$. Here, $\mathbb{E}_t^{\mathbf{w}}$ represents the expectation taken with respect to the distribution $p_t(\cdot|S_t,\mathbf{w})$. Note that $\mathbb{E}_t^{\mathbf{w}}[r_{ti}]=R_t(S_t,\mathbf{w})$. Then, we can rewrite the above inequality in the following form:

$$\begin{aligned}
&\sum_{t\notin\mathcal{T}^w}\nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t-\mathbf{u}_t^\star)\\
&\leqslant 2\beta_T(\delta)\sum_{t\notin\mathcal{T}^w}\left(\mathbb{E}_t^{\mathbf{w}^\star}\left[r_{ti}\|x_{ti}\|_{H_t^{-1}}\right]-\mathbb{E}_t^{\mathbf{w}^\star}\left[r_{ti}\right]\mathbb{E}_t^{\mathbf{w}^\star}\left[\|x_{tj}\|_{H_t^{-1}}\right]\right)\qquad(x_{t0}=\mathbf{0},\,r_{t0}=0)\\
&=2\beta_T(\delta)\sum_{t\notin\mathcal{T}^w}\underbrace{\mathbb{E}_t^{\mathbf{w}^\star}\left[\left(r_{ti}-\mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}]\right)\left(\|x_{ti}\|_{H_t^{-1}}-\mathbb{E}_t^{\mathbf{w}^\star}\left[\|x_{tj}\|_{H_t^{-1}}\right]\right)\right]}_{\text{Covariance between }r_{ti}\text{ and }\|x_{ti}\|_{H_t^{-1}}\text{ given }S_t}.\qquad(\text{D.6})
\end{aligned}$$

By Lemma D.6, we know that $r_{ti}\geqslant R_t(S_t,\mathbf{w}^\star)=\mathbb{E}_t^{\mathbf{w}^\star}[r_{ti}]$ for all $i\in S_t$. Therefore, we can bound the term inside the

expectation in (D.6) as follows:

$$\Big(\underbrace{r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}]}_{\geqslant 0 \text{ by Lemma D.6}}\Big)\left(\|x_{ti}\|_{H_t^{-1}} - \mathbb{E}_t^{\mathbf{w}^\star}\left[\|x_{tj}\|_{H_t^{-1}}\right]\right) \leqslant \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}]\right)\left(\|x_{ti}\|_{H_t^{-1}} - \|\mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\|_{H_t^{-1}}\right)$$

(Jensen's inequality)

$$\leqslant \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}]\right)\left\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\right\|_{H_t^{-1}}, \tag{D.7}$$

where the last inequality holds due to the fact that $\|\mathbf{a}\| = \|\mathbf{a} - \mathbf{b} + \mathbf{b}\| \leqslant \|\mathbf{a} - \mathbf{b}\| + \|\mathbf{b}\|$ for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Plugging (D.7) into (D.6), we obtain

$$\begin{aligned}
&\sum_{t\notin\mathcal{T}^w} \nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t - \mathbf{u}_t^\star) \\
&\leqslant 2\beta_T(\delta)\sum_{t\notin\mathcal{T}^w}\mathbb{E}_t^{\mathbf{w}^\star}\left[(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}])\left\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\right\|_{H_t^{-1}}\right] \\
&\leqslant 2\beta_T(\delta)\sqrt{\sum_{t\notin\mathcal{T}^w}\underbrace{\mathbb{E}_t^{\mathbf{w}^\star}\left[(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}])^2\right]}_{=\mathbb{V}_t^{\mathbf{w}^\star}[r_{ti}]=:\sigma_t^2}}\sqrt{\sum_{t\notin\mathcal{T}^w}\mathbb{E}_t^{\mathbf{w}^\star}\left[\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\|_{H_t^{-1}}^2\right]} \quad \text{(Cauchy-Schwartz inequality)} \\
&= 2\beta_T(\delta)\sqrt{\sum_{t\notin\mathcal{T}^w}\sigma_t^2}\sqrt{\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t\cup\{0\}} p_t(i|S_t,\mathbf{w}^\star)\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\|_{H_t^{-1}}^2}
\end{aligned} \tag{D.8}$$

where in the last equality, we define the variance of the rewards under $\mathbf{w}^\star$, given S_t, as $\sigma_t^2 := \mathbb{V}_t^{\mathbf{w}^\star}[r_{ti}] = \mathbb{E}_t^{\mathbf{w}^\star}\left[\left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}]\right)^2\right]$. We define $H_t(\mathbf{w}^\star) = \frac{\lambda}{e}\mathbf{I}_d + \sum_{s\in[t-1]\setminus\mathcal{T}^w}\nabla^2\ell_s(\mathbf{w}^\star)$. Using this definition, we further bound the right-hand side of Equation (D.8) as follows:

$$\begin{aligned}
\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t\cup\{0\}} p_t(i|S_t,\mathbf{w}^\star)\left\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\right\|_{H_t^{-1}}^2 &\leqslant e\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t\cup\{0\}} p_t(i|S_t,\mathbf{w}^\star)\left\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\right\|_{H_t(\mathbf{w}^\star)^{-1}}^2 \quad \text{(Lemma D.8)} \\
&\leqslant e\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t\cup\{0\}} p_t(i|S_t,\mathbf{w}^\star)\left\|x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}]\right\|_{H_t(\mathbf{w}^\star)^{-1}}^2 \\
&\leqslant 2ed\log\left(1 + \frac{eT}{d\lambda}\right), \quad \text{(Lemma D.7)}
\end{aligned}$$

where when applying the elliptical potential lemma (Lemma D.7), we verify the condition $\|x_{ti}\|_{H_t(\mathbf{w}^\star)^{-1}}^2 \leqslant \frac{1}{2}$ for all $i \in S_t$ as follows:

$$\|x_{ti}\|_{H_t(\mathbf{w}^\star)^{-1}}^2 \leqslant e\|x_{ti}\|_{H_t^{-1}}^2 \leqslant \frac{e}{\lambda} = \frac{e}{144d} \leqslant \frac{1}{2}. \quad \text{(Lemma D.8, } \lambda = 144d\text{)}$$

Therefore, we can bound the term I_1 in Equation (D.5) as follows:

$$\sum_{t\notin\mathcal{T}^w}\nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t - \mathbf{u}_t^\star) \leqslant 2\beta_T(\delta)\sqrt{\sum_{t=1}^T\sigma_t^2}\sqrt{2ed\log\left(1 + \frac{eT}{d\lambda}\right)}. \tag{D.9}$$

Now, we bound the term I_2 in (D.5). We define a function $Q : \mathbb{R}^{|S_t|} \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1, \ldots, u_{|S_t|}) \in \mathbb{R}^{|S_t|}$, $Q(\mathbf{u}) = \sum_{i=1}^{|S_t|}\frac{\exp(u_i)}{1+\sum_{j=1}^{|S_t|}\exp(u_j)}$. Then, it is clear that $\left|\frac{\partial^2\tilde{Q}}{\partial i\partial j}\right| \leqslant \left|\frac{\partial^2 Q}{\partial i\partial j}\right|$ since $r_{ti} \in [0,1]$. Hence, we get

$$\begin{aligned}
\frac{1}{2}\sum_{t\notin\mathcal{T}^w}(\mathbf{u}_t - \mathbf{u}_t^\star)^\top\nabla^2\tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t - \mathbf{u}_t^\star) &\leqslant \frac{1}{2}\sum_{t\notin\mathcal{T}^w}\sum_{i\in S_t}\sum_{j\in S_t}(u_{ti} - u_{ti}^\star)\frac{\partial^2\tilde{Q}}{\partial i\partial j}(u_{tj} - u_{tj}^\star) \\
&\leqslant \frac{1}{2}\sum_{t=1}^T\sum_{i\in S_t}\sum_{j\in S_t}|u_{ti} - u_{ti}^\star|\left|\frac{\partial^2 Q}{\partial i\partial j}\right||u_{tj} - u_{tj}^\star|. \quad (r_{ti}\in[0,1])
\end{aligned}$$

Furthermore, we denote $p_i(\bar{\mathbf{u}}_t) = \frac{\exp(\bar{u}_{ti})}{1+\sum_{k=1}^{|S_t|}\exp(\bar{u}_{tk})}$. Then, we have

$$
\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}|u_{ti}-u_{ti}^\star|\left|\frac{\partial^2 Q}{\partial i\partial j}\right||u_{tj}-u_{tj}^\star| \\
&\quad=\frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t,j\neq i}|u_{ti}-u_{ti}^\star|\left|\frac{\partial^2 Q}{\partial i\partial j}\right||u_{tj}-u_{tj}^\star|+\frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}|u_{ti}-u_{ti}^\star|\left|\frac{\partial^2 Q}{\partial i\partial i}\right||u_{ti}-u_{ti}^\star| \\
&\quad\leqslant\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t,j\neq i}|u_{ti}-u_{ti}^\star|p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t)|u_{tj}-u_{tj}^\star|+\frac{3}{2}\sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2p_i(\bar{\mathbf{u}}_t) && \text{(Lemma D.9)} \\
&\quad\leqslant\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}|u_{ti}-u_{ti}^\star|p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t)|u_{tj}-u_{tj}^\star|+\frac{3}{2}\sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2p_i(\bar{\mathbf{u}}_t) \\
&\quad\leqslant\frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}(u_{ti}-u_{ti}^\star)^2p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t)+\frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t}(u_{tj}-u_{tj}^\star)^2p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t) && \text{(AM-GM inequality)} \\
&\quad+\frac{3}{2}\sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2p_i(\bar{\mathbf{u}}_t) \\
&\quad\leqslant\frac{5}{2}\sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2p_i(\bar{\mathbf{u}}_t).
\end{aligned}
$$

Therefore, we can bound the term I_2 in Equation (D.5) as follows:

$$
\begin{aligned}
\frac{1}{2}\sum_{t\notin\mathcal{T}^w}(\mathbf{u}_t-\mathbf{u}_t^\star)^\top\nabla^2\tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t-\mathbf{u}_t^\star)&\leqslant\frac{5}{2}\sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2p_i(\bar{\mathbf{u}}_t) \\
&=10\sum_{t=1}^{T}\sum_{i\in S_t}p_i(\bar{\mathbf{u}}_t)\beta_t(\delta)^2\|x_{ti}\|^2_{H_t^{-1}} && \text{(definitions of } \mathbf{u}_t \text{ and } \mathbf{u}_t^\star\text{)} \\
&\leqslant 10\sum_{t=1}^{T}\max_{i\in S_t}\beta_t(\delta)^2\|x_{ti}\|^2_{H_t^{-1}} \\
&\leqslant 10\beta_T(\delta)^2\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} \\
&\leqslant\frac{20}{\kappa}\beta_T(\delta)^2d\log\left(1+\frac{T}{d\lambda}\right). && \text{(D.10)}
\end{aligned}
$$

Finally, by substituting (D.9) and (D.10) into (D.4), and setting $\lambda = 144d$, $\tau_T = \mathcal{O}\left(B\sqrt{d\log(T/\delta)}+B^{3/2}\sqrt{d}+B^2\right)$, and $\beta_T(\delta) = \mathcal{O}\left(\sqrt{d\log(T/\delta)}+B\sqrt{d}\right)$, we obtain

$$
\begin{aligned}
\mathbf{Reg}_T(\mathbf{w}^\star)&\leqslant\frac{2}{\kappa}\tau_T^2d\log\left(1+\frac{T}{d\lambda}\right)+2\beta_T(\delta)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\sqrt{2ed\log\left(1+\frac{eT}{d\lambda}\right)}+\frac{20}{\kappa}\beta_T(\delta)^2d\log\left(1+\frac{T}{d\lambda}\right) \\
&=\mathcal{O}\left(\left(d\log T+Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^{T}\sigma_t^2}+\frac{1}{\kappa}B^3d^2\left(\log T\right)^2+\frac{1}{\kappa}B^4d\log T\right).
\end{aligned}
$$

By setting $\delta\leftarrow\delta/2$, we complete the proof of Theorem 4.5. □

D.2. Proofs of Corollaries and Lemmas for Theorem 4.5

D.2.1. Proof of Corollary D.1

Proof of Corollary D.1. In Theorem 4.2, we consider the case where $\mathcal{W}_t = \mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\|_2 \leqslant B\}$ for all $t \geqslant 1$ and the Hessian matrix is H_t^w.

Condition: $\sup_{\mathbf{w}\in\mathcal{W}_t} |x_{ti}^\top(\mathbf{w} - \mathbf{w}^\star)| \leqslant \alpha$ for all $i \in S_t \implies \alpha = 2B.$

We set $\alpha = 2B$, as shown below:

$$\sup_{\mathbf{w}\in\mathcal{W}_t} |x_{ti}^\top(\mathbf{w} - \mathbf{w}^\star)| = \sup_{\mathbf{w}\in\mathcal{W}} |x_{ti}^\top(\mathbf{w} - \mathbf{w}^\star)| \leqslant \sup_{\mathbf{w}\in\mathcal{W}} \|x_{ti}\|_2 \|\mathbf{w} - \mathbf{w}^\star\|_2 \leqslant 2B.$$

Substituting $\alpha = 2B$ (which gives $\eta^w = \frac{1}{2} + 3\sqrt{2}B$) into Theorem 4.2, while setting $\lambda^w = \max\{12\sqrt{2}\eta^w\alpha, 144\eta^w d, 2\}$, for $t \in \mathcal{T}^w$, we obtain

$$\|\mathbf{w}^\star - \mathbf{w}_t^w\|_{H_t^w} \leqslant \zeta_t(\delta) = \mathcal{O}\left(B\sqrt{d\log(t/\delta)} + B^{3/2}\sqrt{d} + B^2\right).$$

For $t \notin \mathcal{T}^w$, the confidence set $\mathcal{W}_t^w(\delta)$, along with $\mathbf{w}_t^w$ and H_t^w, remains unchanged. Thus, the proof is complete. □

D.2.2. Proof of Corollary D.2

Proof of Corollary D.2. As with Corollary D.1, we prove this using Theorem 4.2. Let $\mathcal{W}_t = \mathcal{W}_t^w(\delta)$ and let the Hessian matrix be H_t.

Condition: $\sup_{\mathbf{w}\in\mathcal{W}_t} |x_{ti}^\top(\mathbf{w} - \mathbf{w}^\star)| \leqslant \alpha$ for all $i \in S_t \implies \alpha = \frac{1}{3\sqrt{2}}.$

We set $\alpha = \frac{1}{3\sqrt{2}}$, as shown below:

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{W}_t} |x_{ti}^\top(\mathbf{w} - \mathbf{w}^\star)| &= \sup_{\mathbf{w}\in\mathcal{W}_t^w(\delta)} |x_{ti}^\top(\mathbf{w} - \mathbf{w}^\star)| \\
&\leqslant \max_{x\in\mathcal{X}_t} \|x\|_{(H_t^w)^{-1}} \left(\max_{\mathbf{w}\in\mathcal{W}_t^w(\delta)} \|\mathbf{w} - \mathbf{w}_t^w\|_{H_t^w} + \|\mathbf{w}_t^w - \mathbf{w}^\star\|_{H_t^w}\right) && \text{(Hölder's inequality)} \\
&\leqslant \frac{1}{\tau_t}\left(\max_{\mathbf{w}\in\mathcal{W}_t^w(\delta)} \|\mathbf{w} - \mathbf{w}_t^w\|_{H_t^w} + \|\mathbf{w}_t^w - \mathbf{w}^\star\|_{H_t^w}\right) && (t \notin \mathcal{T}^w) \\
&\leqslant \frac{1}{6\sqrt{2}\zeta_t(\delta)}\left(\zeta_t(\delta) + \|\mathbf{w}_t^w - \mathbf{w}^\star\|_{H_t^w}\right) && \text{(Definitions of } \tau_t \text{ and } \mathcal{W}_t^w(\delta)) \\
&\leqslant \frac{\zeta_t(\delta)}{3\sqrt{2}\zeta_t(\delta)} && \text{(Corollary D.1)} \\
&= \frac{1}{3\sqrt{2}}.
\end{aligned}$$

Plugging $\alpha = \frac{1}{3\sqrt{2}}$ (which implies $\eta = \frac{1}{2}(1 + 3\sqrt{2}\alpha) = 1$) into Theorem 4.2, while setting $\lambda = \max\{12\sqrt{2}\eta\alpha, 144\eta d, 2\} = 144d$, for $t \notin \mathcal{T}^w$, we derive

$$\|\mathbf{w}^\star - \mathbf{w}\|_{H_t} \leqslant \beta_t(\delta) = \mathcal{O}\left(\sqrt{d\log(t/\delta)} + B\sqrt{d}\right).$$

For $t \in \mathcal{T}^w$, the confidence set $\mathcal{C}_t(\delta)$, along with $\mathbf{w}_t$ and H_t, remains the same. This conclude the proof of Corollary D.2. □

D.2.3. Proof of Lemma D.6

Proof of Lemma D.6. By the definition of the optimal assortment and Lemma D.4, we have

$$R_t(S_t, \mathbf{w}^\star) \leqslant R_t(S_t^\star, \mathbf{w}^\star) \leqslant \tilde{R}_t(S_t^\star) \leqslant \tilde{R}_t(S_t).$$

Thus, it is sufficient to show that $r_{ti} \geqslant \tilde{R}_t(S_t)$ for all $i \in S_t$.

We prove this by contradiction. Suppose there exists an item $i \in S_t$ such that $r_{ti} < \tilde{R}_t(S_t)$. If we remove item i from the assortment S_t, it would result in higher expected revenue. This contradicts the optimality of $S_t = \operatorname{argmax}_{S \in \mathcal{S}} \tilde{R}_t(S)$. Hence, we conclude

$$r_{ti} \geqslant \tilde{R}_t(S_t), \quad \forall i \in S_t,$$

which completes the proof. □

D.2.4. Proof of Lemma D.7

Proof of Lemma D.7. For simplicity, let $\mathbb{E}_t^{\mathbf{w}}[x_{tj}] = \mathbb{E}_{j \sim p_t(\cdot|S_t,\mathbf{w})}[x_{tj}]$ and $x_{t0} = \mathbf{0}$. Then, we can express $\nabla^2 \ell_s(\mathbf{w})$ as follows:

$$\begin{aligned}
\nabla^2 \ell_s(\mathbf{w}) &= \sum_{i \in S_s} p_s(i|S_s,\mathbf{w}) x_{si} x_{si}^\top - \sum_{i \in S_s} \sum_{j \in S_s} p_s(i|S_s,\mathbf{w}) p_s(j|S_s,\mathbf{w}) x_{si} x_{sj}^\top \\
&= \sum_{i \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w}) x_{si} x_{si}^\top - \sum_{i \in S_s \cup \{0\}} \sum_{j \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w}) p_s(j|S_s,\mathbf{w}) x_{si} x_{sj}^\top \\
&= \mathbb{E}_s^{\mathbf{w}}[x_{si} x_{si}^\top] - \mathbb{E}_s^{\mathbf{w}}[x_{si}] \left(\mathbb{E}_s^{\mathbf{w}}[x_{si}]\right)^\top \\
&= \mathbb{E}_s^{\mathbf{w}}\left[(x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}])(x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}])^\top\right] \\
&= \sum_{i \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w})(x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}])(x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}])^\top.
\end{aligned}$$

Using the definition $H_t(\mathbf{w}) = \lambda \mathbf{I}_d + \sum_{s \in [t-1] \backslash \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w})$, it follows that for any update round $s \in [t] \backslash \mathcal{T}^w$, we have

$$\det\left(H_{s+1}\right) = \det\left(H_s\right) \left(1 + \sum_{i \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w}) \left\|x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}]\right\|^2_{H_s(\mathbf{w})^{-1}}\right).$$

By the assumption that $\|x_{si}\|^2_{H_s(\mathbf{w})^{-1}} \leqslant \frac{1}{2}$ for all $i \in S_s$, we know that $\sum_{i \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w}) \left\|x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}]\right\|^2_{H_s(\mathbf{w})^{-1}} \leqslant 1$. Then, using the fact that $z \leqslant 2\log(1+z)$ for any $z \in [0,1]$, we obtain

$$\begin{aligned}
&\sum_{s \in [t] \backslash \mathcal{T}^w} \sum_{i \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w}) \left\|x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}]\right\|^2_{H_s(\mathbf{w})^{-1}} \\
&\qquad \leqslant 2 \sum_{s \in [t] \backslash \mathcal{T}^w} \log\left(1 + \sum_{i \in S_s \cup \{0\}} p_s(i|S_s,\mathbf{w}) \left\|x_{si} - \mathbb{E}_s^{\mathbf{w}}[x_{sj}]\right\|^2_{H_s(\mathbf{w})^{-1}}\right) \\
&\qquad = 2 \sum_{s \in [t] \backslash \mathcal{T}^w} \log\left(\frac{\det(H_{s+1})}{\det(H_1)}\right) \\
&\qquad = 2 \log\left(\frac{\det(H_{t+1})}{\det(H_1)}\right) \\
&\qquad \leqslant 2d \log\left(\frac{\operatorname{tr}(H_{t+1})}{d\lambda}\right) \leqslant 2d \log\left(1 + \frac{t}{d\lambda}\right),
\end{aligned}$$

which concludes the proof. □

D.2.5. Proof of Lemma D.8

Proof of Lemma D.8. For any $s \in [t-1] \backslash \mathcal{T}^w$, let $X_s \in \mathbb{R}^{|S_t| \times d}$ be the matrix whose i'th row is $x_{si}^\top$. Then, by the equivalent notation of the loss (see Equation (A.1)), we have

$$\begin{aligned}
\nabla^2 \ell_s(\mathbf{w}_{s+1}) &= X_s^\top \nabla^2_{\mathbf{z}} \bar{\ell}_s(\mathbf{z}_{s+1}) X_s && \text{(Eqn. (A.1))} \\
&\leq e^{3\sqrt{2}\|\mathbf{z}_{s+1} - \mathbf{z}_s^\star\|_\infty} X_s^\top \nabla^2_{\mathbf{z}} \bar{\ell}_s(\mathbf{z}_s^\star) X_s && \text{(Proposition B.5)} \\
&\leq e X_s^\top \nabla^2_{\mathbf{z}} \bar{\ell}_s(\mathbf{z}_s^\star) X_s && (\|\mathbf{z}_{s+1} - \mathbf{z}_s^\star\|_\infty \leqslant \tfrac{1}{3\sqrt{2}}) \\
&= e \nabla^2 \ell_s(\mathbf{w}^\star),
\end{aligned}$$

where the last inequality holds because, for any $s \notin \mathcal{T}^w$, the following holds:

$$\begin{aligned}
\|\mathbf{z}_{s+1} - \mathbf{z}_s^\star\|_\infty &= \max_{i \in S_s} |x_{si}^\top (\mathbf{w}_{s+1} - \mathbf{w}^\star)| \\
&= \max_{i \in S_s} \|x_{si}\|_{(H_s^w)^{-1}} \left(\|\mathbf{w}_{s+1} - \mathbf{w}_s^w\|_{H_s^w} + \|\mathbf{w}_s^w - \mathbf{w}^\star\|_{H_s^w} \right) && \text{(Hölder's inequality)} \\
&\leqslant \frac{1}{\tau_s} \left(\|\mathbf{w}_{s+1} - \mathbf{w}_s^w\|_{H_s^w} + \|\mathbf{w}_s^w - \mathbf{w}^\star\|_{H_s^w} \right) && (s \notin \mathcal{T}^w) \\
&\leqslant \frac{1}{6\sqrt{2}\zeta_s(\delta)} \left(\zeta_s(\delta) + \|\mathbf{w}_s^w - \mathbf{w}^\star\|_{H_s^w} \right) && \text{(Definitions of } \tau_s \text{ and } \mathbf{w}_{s+1} \in \mathcal{W}_s^w(\delta)) \\
&\leqslant \frac{\zeta_s(\delta)}{3\sqrt{2}\zeta_s(\delta)} && \text{(Corollary D.1)} \\
&= \frac{1}{3\sqrt{2}}.
\end{aligned}$$

Thus, we get

$$H_t = \lambda \mathbf{I}_d + \sum_{s \in [t-1] \setminus \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w}_{s+1}) \preceq \lambda \mathbf{I}_d + e \sum_{s \in [t-1] \setminus \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w}^\star) = e H_t(\mathbf{w}^\star).$$

To prove the other inequality, we use a similar line of reasoning:

$$\nabla^2 \ell_s(\mathbf{w}^\star) = X_s^\top \nabla^2_{\mathbf{z}} \bar{\ell}_s(\mathbf{z}_s^\star) X_s \preceq e^{3\sqrt{2}\|\mathbf{z}_{s+1} - \mathbf{z}_s^\star\|_\infty} X_s^\top \nabla^2_{\mathbf{z}} \bar{\ell}_s(\mathbf{z}_{s+1}) X_s \preceq e X_s^\top \nabla^2_{\mathbf{z}} \bar{\ell}_s(\mathbf{z}_{s+1}) X_s = e \nabla^2 \ell_s(\mathbf{w}_{s+1}),$$

which implies that

$$H_t = \lambda \mathbf{I}_d + \sum_{s \in [t-1] \setminus \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w}_{s+1}) \succeq \lambda \mathbf{I}_d + \frac{1}{e} \sum_{s \in [t-1] \setminus \mathcal{T}^w} \nabla^2 \ell_s(\mathbf{w}^\star) \succeq \frac{1}{e} H_t(\mathbf{w}^\star).$$

This concludes the proof. □

D.2.6. Proof of Lemma D.10

Proof of Lemma D.10. Recall that by the definition of H_t^w, we have

$$\begin{aligned}
H_t^w &= \lambda \mathbf{I}_d + \sum_{s \in \mathcal{T}^w \setminus \{t, \dots, T\}} \nabla^2 \ell_s(\mathbf{w}_{s+1}) \\
&= \lambda \mathbf{I}_d + \sum_{s \in \mathcal{T}^w \setminus \{t, \dots, T\}} p_s(i_s | \{i_s\}, \mathbf{w}_{s+1}) x_{si_s} x_{si_s}^\top - p_s(i_s | \{i_s\}, \mathbf{w}_{s+1}) p_s(i_s | \{i_s\}, \mathbf{w}_{s+1}) x_{si_s} x_{si_s}^\top \\
&= \lambda \mathbf{I}_d + \sum_{s \in \mathcal{T}^w \setminus \{t, \dots, T\}} p_s(i_s | \{i_s\}, \mathbf{w}_{s+1}) p_s(0 | \{i_s\}, \mathbf{w}_{s+1}) x_{si_s} x_{si_s}^\top,
\end{aligned}$$

where i_s is the index of the item such that $x_{si_s} = \operatorname{argmax}_{x \in \mathcal{X}_s} \|x\|^2_{(H_s^w)^{-1}}$. Then, we get

$$\begin{aligned}
\sum_{t \in \mathcal{T}^w} \max_{x \in \mathcal{X}_t} \|x\|^2_{(H_t^w)^{-1}} &= \sum_{t \in \mathcal{T}^w} \|x_{ti_t}\|^2_{(H_t^w)^{-1}} && (x_{ti_t} = \operatorname{argmax}_{x \in \mathcal{X}_t} \|x\|^2_{(H_t^w)^{-1}} \text{ for } t \in \mathcal{T}^w) \\
&\leqslant \frac{1}{\kappa} \sum_{t \in \mathcal{T}^w} p_t(i_t | \{i_t\}, \mathbf{w}_{t+1}) p_t(0 | \{i_t\}, \mathbf{w}_{t+1}) \|x_{ti_t}\|^2_{(H_t^w)^{-1}} && \text{(Definition of } \kappa) \\
&= \frac{1}{\kappa} \sum_{t \in \mathcal{T}^w} \min \left\{ \frac{1}{2}, p_t(i_t | \{i_t\}, \mathbf{w}_{t+1}) p_t(0 | \{i_t\}, \mathbf{w}_{t+1}) \|x_{ti_t}\|^2_{(H_t^w)^{-1}} \right\}, \\
& && (\|x_{ti_t}\|^2_{(H_t^w)^{-1}} \leqslant \tfrac{1}{\lambda^w} \|x_{ti_t}\|_2 \leqslant \tfrac{1}{2}, \lambda^w \geqslant 2) \\
&\leqslant \frac{2}{\kappa} d \log \left(1 + \frac{T}{d\lambda} \right). && \text{(Lemma D.11)}
\end{aligned}$$

On the other hand, by the definition of the update rule in Algorithm 1, we have

$$\sum_{t\in\mathcal{T}^w}\max_{x\in\mathcal{X}_t}\|x\|^2_{(H_t^w)^{-1}} \geqslant \sum_{t\in\mathcal{T}^w}\frac{1}{\tau_t^2} \geqslant \frac{1}{\tau_T^2}\left|\mathcal{T}^w\right|. \qquad (\tau_t \text{ is non-decreasing})$$

By combining the two results above, we obtain

$$|\mathcal{T}^w| \leqslant \frac{2}{\kappa}\tau_T^2 d\log\left(1+\frac{T}{d\lambda}\right),$$

which concludes the proof. □

D.3. Technical Lemmas

Lemma D.11 (Lemma F.2 and H.3 of Lee & Oh 2024)**.** *Let* $H_t = \lambda\mathbf{I}_d + \sum_{s=1}^{t-1}\nabla^2\ell_s(\mathbf{w}_{s+1})$. *Define* $\tilde{x}_{si} := x_{si} - \mathbb{E}_{j\sim p_s(\cdot|S_s,\mathbf{w}_{s+1})}[x_{sj}]$. *If* $\|x_{si}\|^2_{H_s^{-1}} \leqslant \frac{1}{2}$ *for all* $i\in S_t$ *and* $s\in[t]$, *then the following statements hold true:*

(1) $\sum_{s=1}^t\sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})p_s(0|S_s,\mathbf{w}_{s+1})\|x_{si}\|^2_{H_s^{-1}} \leqslant 2d\log\left(1+\frac{t}{d\lambda}\right)$,

(2) $\sum_{s=1}^t\sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})\|\tilde{x}_{si}\|^2_{H_s^{-1}} \leqslant 2d\log\left(1+\frac{t}{d\lambda}\right)$,

(3) $\sum_{s=1}^t\max_{i\in S_s}\|x_{si}\|^2_{H_s^{-1}} \leqslant \frac{2}{\kappa}d\log\left(1+\frac{t}{d\lambda}\right)$,

(4) $\sum_{s=1}^t\max_{i\in S_s}\|\tilde{x}_{si}\|^2_{H_s^{-1}} \leqslant \frac{2}{\kappa}d\log\left(1+\frac{t}{d\lambda}\right)$.

E. Instance-Dependent Regret

As a special case, if the rewards are uniform (i.e., $r_{ti}=1$), we can establish an instance-dependent regret bound.

Proposition E.1 (Restatement of Proposition 4.10 Instance-dependent regret under uniform rewards)**.** *Define* $\kappa_t^\star := \sum_{i\in S_t^\star} p_t(i|S_t^\star,\mathbf{w}^\star)p_t(0|S_t^\star,\mathbf{w}^\star)$. *Under the same conditions as Theorem 4.5 and assuming uniform rewards, the regret of* `OFU-MNL++` *is upper bounded by*

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^T\kappa_t^\star + \frac{1}{\kappa}B^3d^2\left(\log T\right)^2} + \frac{1}{\kappa}B^4 d\log T\right).$$

E.1. Proof of Proposition 4.10

In this section, we present the proof of Proposition 4.10. In the case of uniform rewards, where $r_{ti=1}$ for all $i\in[N]$, the σ_t^2 term can be upper-bounded by $\kappa_t^\star$ plus an additive term.

Proof of Proposition 4.10. From Theorem 4.5, we have

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^T\sigma_t^2 + \frac{1}{\kappa}B^3d^2\left(\log T\right)^2} + \frac{1}{\kappa}B^4 d\log T\right).$$

When rewards are uniform, we can rewrite the $\sum_{t=1}^{T} \sigma_t^2$ term as follows:

$$\begin{aligned}
\sum_{t=1}^{T} \sigma_t^2 &= \sum_{t=1}^{T} \mathbb{E}_{i\sim p_t(\cdot|S_t,\mathbf{w}^\star)} \left[\left(r_{ti} - \mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}[r_{tj}] \right)^2 \right] \\
&= \sum_{t=1}^{T} \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star) r_{ti}^2 - \left(\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star) r_{ti} \right)^2 \\
&= \sum_{t=1}^{T} \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star) p_t(0|S_t,\mathbf{w}^\star) \qquad (r_{ti}=1) \\
&\leqslant \sum_{t=1}^{T} \kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star).
\end{aligned}$$

where the last inequality holds by the following lemma:

Lemma E.2 (Lemma 11 of Perivier & Goyal 2022)**.** *Let* $\kappa_t^\star := \sum_{i\in S_t^\star} p_t(i|S_t^\star,\mathbf{w}^\star) p_t(0|S_t^\star,\mathbf{w}^\star)$*. Then, we have*

$$\sum_{t=1}^{T} \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star) p_t(0|S_t,\mathbf{w}^\star) \leqslant \sum_{t=1}^{T} \kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star).$$

Hence, we get

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T} \right) \sqrt{ \sum_{t=1}^{T} \kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star) + \frac{1}{\kappa} B^3 d^2 (\log T)^2 + \frac{1}{\kappa} B^4 d \log T } \right).$$

Solving the above equation completes the proof of Proposition 4.10. □

E.2. Discussion on Instance-Dependent Regret

In this subsection, we further discuss about the instance-dependent parameter $\kappa_t^\star$ and the variance of rewards σ_t.

True meaning of $\kappa_t^\star$. The instance-dependent parameter $\kappa_t^\star$, which appears in the regret bounds of many existing (multinomial) logistic and GLM bandits (Abeille et al., 2021; Faury et al., 2022; Perivier & Goyal, 2022; Lee & Oh, 2024; Sawarni et al., 2024), is indeed the *variance of the uniform rewards given* $S_t^\star$. In contrast, σ_t^2 denotes the variance of general rewards (including both uniform and non-uniform) for the offered assortment S_t. Under uniform rewards, as shown in the analysis of Proposition 4.10, $\kappa_t^\star$ and σ_t^2 are closely related, as the assortment size remains the same and the rewards are identical.

Possibility of Instance-Dependent Regret under Non-Uniform Rewards. Readers might expect an instance-dependent regret bound for general non-uniform rewards. However, we cautiously argue that establishing such a bound in the non-uniform case is non-trivial using existing analytical approaches. Unlike prior works on binary logistic bandits (Abeille et al., 2021; Faury et al., 2022), uniform rewards MNL bandits (Perivier & Goyal, 2022; Lee & Oh, 2024), and generalized linear bandits (Sawarni et al., 2024), the size and rewards of the offered assortment $S^\star$ and the optimal assortment $S_t^\star$ are different. This fundamental discrepancy makes it impossible to bound quantities related to S_t using those related to $S_t^\star$.

F. Proof of Theorem 4.12

In this section, we provide the proof of Theorem 4.12. For ease of reference, Table F.1 summarizes the notations used for `OFU-M`^{2}`NL`.

We define $\mathcal{L}_t(\mathbf{w})$ as the negative log-likelihood of $\mathbf{w}$ with respect to data collected up to $t-1$, and $\widehat{\mathbf{w}}_t$ as the corresponding maximum likelihood estimate (MLE) estimate:

$$\mathcal{L}_t(\mathbf{w}) := \sum_{s=1}^{t-1} \ell_s(\mathbf{w}) = -\sum_{s=1}^{t-1} \sum_{i\in S_s} y_{ts} \log p_s(i|S_s,\mathbf{w}), \quad \widehat{\mathbf{w}}_t := \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_t(\mathbf{w}).$$

Table F.1: Symbols for OFU-M^{2}NL

$\mathcal{L}_t(\mathbf{w})$	$:= -\sum_{s=1}^{t-1}\sum_{i\in S_s} y_{ts}\log p_s(i\|S_s,\mathbf{w})$.
$\hat{\mathbf{w}}_t$	maximum likelihood estimate (MLE) estimate at round t
λ^{MLE}	$:= \frac{1}{8B^2}$, regularization parameter for MLE
$H_t^{\text{MLE}}(\mathbf{w})$	$:= \sum_{s=1}^{t-1}\nabla^2\ell_s(\mathbf{w}) + \lambda^{\text{MLE}}\mathbf{I}_d$
$\boldsymbol{\nu}_t^\star$	parameter that satisfies $\frac{1}{2}\left\|\mathbf{w}^\star - \widehat{\mathbf{w}}_t\right\|^2_{\nabla^2\mathcal{L}_t(\boldsymbol{\nu}_t^\star)} = \left\|\mathbf{w}^\star - \widehat{\mathbf{w}}_t\right\|^2_{\int_0^1(1-v)\nabla^2\mathcal{L}_t(\widehat{\mathbf{w}}_t+v(\mathbf{w}^\star-\widehat{\mathbf{w}}_t))\mathrm{d}v}$
$\beta_t^{\text{MLE}}(\delta)$	$:= \mathcal{O}\left(\sqrt{d\log(Bt/\delta)}\right)$
$\gamma_t^{\text{MLE}}(\delta)$	$:= \sqrt{2\beta_t^{\text{MLE}}(\delta)^2+1}$
$\mathcal{C}_t^{\text{MLE}}(\delta)$	$:= \left\{\mathbf{w}\in\mathcal{W} : \mathcal{L}_t(\mathbf{w}) - \mathcal{L}_t(\widehat{\mathbf{w}}_t) \leqslant \beta_t^{\text{MLE}}(\delta)^2\right\}$
$\tilde{\mathbf{w}}_{ti}$	$:= \operatorname{argmax}_{\mathbf{w}\in\mathcal{C}_t^{\text{MLE}}(\delta)} x_{ti}^\top\mathbf{w}$, optimistic utility of item i at round t
$\tilde{R}_t^{\text{MLE}}(S)$	$:= \frac{\sum_{i\in S}\exp(x_{ti}^\top\tilde{\mathbf{w}}_{ti})r_{ti}}{1+\sum_{j\in S}\exp(x_{tj}^\top\tilde{\mathbf{w}}_{tj})}$, optimistic expected revenue of assortment S at round t
σ_t^2	$:= \mathbb{E}_{i\sim p_t(\cdot\|S_t,\mathbf{w}^\star)}\left[\left(r_{ti} - \mathbb{E}_{j\sim p_t(\cdot\|S_t,\mathbf{w}^\star)}[r_{tj}]\right)^2\right]$, variance of rewards given S_t at round t

And the confidence set is defined as follows:

$$\mathcal{C}_t^{\text{MLE}}(\delta) := \left\{\mathbf{w}\in\mathcal{W} : \mathcal{L}_t(\mathbf{w}) - \mathcal{L}_t(\widehat{\mathbf{w}}_t) \leqslant \beta_t^{\text{MLE}}(\delta)^2\right\},$$

where

$$\beta_t^{\text{MLE}}(\delta) := \sqrt{\log\frac{1}{\delta} + d\log\left(\max\left\{e, \frac{4eB(t-1)}{d}\right\}\right)}.$$

The confidence radius $\beta_t^{\text{MLE}}(\delta)$ follows directly from Theorem 3.1 in Lee et al. (2024b). This result is derived by incorporating the Lipschitz constant for the MNL loss, i.e., $L_t = \max_{\mathbf{w}\in\mathcal{W}}\|\nabla\mathcal{L}_t(\mathbf{w})\|_2 \leqslant (t-1)\|\nabla\ell_t(\mathbf{w})\|_2 \leqslant 2(t-1)$ (under Assumption 3.1).

Lemma F.1 (Unified CS for generalized linear models (GLMs), Theorem 3.1 of Lee et al. 2024b)**.** *Let $L_t := \max_{\mathbf{w}\in\mathcal{W}}\|\nabla\mathcal{L}_t(\mathbf{w})\|_2$ be the Lipschitz constant of $\mathcal{L}_t(\cdot)$, which may depend on $\{(x_s,r_s)\}_{s=1}^{t-1}$. Then, we have $\Pr[\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{C}_t^{MLE}(\delta)] \geqslant 1-\delta$, where*

$$\mathcal{C}_t^{MLE}(\delta) := \left\{\mathbf{w}\in\mathcal{W} : \mathcal{L}_t(\mathbf{w}) - \mathcal{L}_t(\widehat{\mathbf{w}}_t) \leqslant \beta_t^{MLE}(\delta)^2 = \log\frac{1}{\delta} + d\log\left(\max\left\{e, \frac{2eBL_t}{d}\right\}\right)\right\}.$$

Then, we offer an assortment S_t that maximizes the optimistic expected revenue $\tilde{R}_t^{\text{MLE}}(S)$ as follows:

$$S_t = \operatorname*{argmax}_{S\in\mathcal{S}}\tilde{R}_t^{\text{MLE}}(S) = \operatorname*{argmax}_{S\in\mathcal{S}}\frac{\sum_{i\in S}\exp(x_{ti}^\top\tilde{\mathbf{w}}_{ti})r_{ti}}{1+\sum_{j\in S}\exp(x_{tj}^\top\tilde{\mathbf{w}}_{tj})}, \quad \text{where } \tilde{\mathbf{w}}_{ti} = \operatorname*{argmax}_{\mathbf{w}\in\mathcal{C}_t^{\text{MLE}}(\delta)} x_{ti}^\top\mathbf{w}. \tag{F.1}$$

Additionally, we define the Hessian of the regularized loss at $\mathbf{w}$ as:

$$H_t^{\text{MLE}}(\mathbf{w}) := \sum_{s=1}^{t-1}\nabla^2\ell_s(\mathbf{w}) + \lambda^{\text{MLE}}\mathbf{I}_d, \quad \text{where } \lambda^{\text{MLE}} = \frac{1}{8B^2}.$$

Now, we present useful lemmas that will be used in the proof of Theorem 4.12.

Lemma F.2 (Restatement of Lemma 4.11, Improved MLE confidence bound)**.** *For any $t\in[T]$, we define $\boldsymbol{\nu}_t^\star$ such that $\frac{1}{2}\left\|\mathbf{w}^\star - \widehat{\mathbf{w}}_t\right\|^2_{\nabla^2\mathcal{L}_t(\boldsymbol{\nu}_t^\star)} = \left\|\mathbf{w}^\star - \widehat{\mathbf{w}}_t\right\|^2_{\int_0^1(1-v)\nabla^2\mathcal{L}_t(\widehat{\mathbf{w}}_t+v(\mathbf{w}^\star-\widehat{\mathbf{w}}_t))\mathrm{d}v}$ and $H_t^{MLE}(\boldsymbol{\nu}_t^\star) := \nabla^2\mathcal{L}_t(\boldsymbol{\nu}_t^\star) + \lambda^{MLE}\mathbf{I}_d = \sum_{s=1}^{t-1}\nabla^2\ell_s(\boldsymbol{\nu}_t^\star) + \lambda^{MLE}\mathbf{I}_d$. Let $\lambda^{MLE} = \frac{1}{8B^2}$. Then, for any $t\geqslant 1$, if $\mathbf{w}^\star\in\mathcal{C}_t^{MLE}(\delta)$ and Assumption 3.1 holds, then we have*

$$\left\|\mathbf{w}^\star - \widehat{\mathbf{w}}_t\right\|^2_{H_t^{MLE}(\boldsymbol{\nu}_t^\star)} \leqslant \underbrace{2\beta_t^{MLE}(\delta)^2+1}_{=:\gamma_t^{MLE}(\delta)^2} = \mathcal{O}\left(d\log(Bt)\right).$$

Algorithm F.1 OFU-M^{2}NL, **OFU**-**M**aximum Likelihood Estimation **MNL**

Input: failure level δ, confidence radius $\beta_t^{\text{MLE}}(\delta)$.
for round $t = 1, \dots, T$ **do**
  Observe feature set $\mathcal{X}_t$.
  Calculate the norm-constrained MLE: $\widehat{\mathbf{w}}_t \leftarrow \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \mathcal{L}_t(\mathbf{w})$.
  Update $\mathcal{C}_t^{\text{MLE}}(\delta) \leftarrow \left\{\mathbf{w} \in \mathcal{W} : \mathcal{L}_t(\mathbf{w}) - \mathcal{L}_t(\widehat{\mathbf{w}}_t) \leqslant \beta_t^{\text{MLE}}(\delta)^2\right\}$.
  Set $\tilde{\mathbf{w}}_{ti} \leftarrow \operatorname{argmax}_{\mathbf{w} \in \mathcal{C}_t^{\text{MLE}}(\delta)} x_{ti}^\top \mathbf{w}$ for all $i \in [N]$.
  Offer $S_t = \operatorname{argmax}_{S \in \mathcal{S}} \tilde{R}_t^{\text{MLE}}(S)$ and observe $\mathbf{y}_t$.
end for

The proof is provided in Appendix F.2. By Lemma F.2, we define the ellipsoidal version of the confidence radius as follows:

$$\gamma_t^{\text{MLE}}(\delta) := \sqrt{2 \log \frac{1}{\delta} + 2d \log \left(\max \left\{ e, \frac{4eB(t-1)}{d} \right\} \right) + 1} = \mathcal{O}\left(\sqrt{d \log(Bt)} \right).$$

Additionally, we present useful technical lemmas.

Lemma F.3. *For any $t \in [T]$, $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{C}_t^{MLE}(\delta)$, and $\omega_{ti} \geqslant 0$, we have*

$$\sum_{i \in S_t} |p_t(j|S_t, \mathbf{w}_1) - p_t(j|S_t, \mathbf{w}_2)| \, \omega_{ti} \leqslant 4\gamma_t^{MLE}(\delta) \max_{i \in S_t} \omega_{ti} \max_{i \in S_t} \|x_{ti}\|_{H_t^{MLE}(\boldsymbol{\nu}_t^\star)^{-1}}.$$

The proof is deferred to Appendix F.2.2.

F.1. Main Proof of Theorem 4.12

Proof of Theorem 4.12. We follow a reasoning process similar to that used in the proof of Theorem 4.5.

First, we define the set of large elliptical potential rounds as follows:

$$\mathcal{T}_0^{\text{MLE}} := \left\{ t \in [T] : \|x_{ti}\|^2_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \geqslant \frac{1}{2}, \quad \forall i \in S_t \right\}.$$

Furthermore, we define $V_t := \sum_{s \in \mathcal{T}_0^{\text{MLE}}, s<t} \tilde{x}_{si_s} \tilde{x}_{si_s}^\top + \lambda^{\text{MLE}} \mathbf{I}_d$, where i_s is any item in S_s (for example $i_s := \operatorname{argmax}_{j \in S_s} \|x_{sj}\|^2_{H_s^{\text{MLE}}(\boldsymbol{\nu}_s^\star)^{-1}}$), and $\tilde{x}_{si} := \sqrt{p_s(i|S_s, \boldsymbol{\nu}_s^\star) p_t(0|S_s, \boldsymbol{\nu}_s^\star)} x_{si}$. Then, the following inequality holds:

$$\begin{aligned}
\left(\frac{d\lambda^{\text{MLE}} + |\mathcal{T}_0^{\text{MLE}}|}{d} \right)^d &\geqslant \left(\frac{\operatorname{Tr}(V_T)}{d} \right)^d && (\|\tilde{x}_{ti}\|_2 \leqslant 1) \\
&\geqslant \det(V_T) && \text{(AM-GM inequality)} \\
&= \det(V_0) \prod_{t \in \mathcal{T}_0^{\text{MLE}}} \left(1 + \|\tilde{x}_{ti_t}\|^2_{V_t^{-1}} \right) && \text{(rank-1 update equality for det.)} \\
&\geqslant \det(V_0) \prod_{t \in \mathcal{T}_0^{\text{MLE}}} \left(1 + \|\tilde{x}_{ti_t}\|^2_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right) && (H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star) \succeq V_t) \\
&\geqslant \det(V_0) \prod_{t \in \mathcal{T}_0^{\text{MLE}}} \left(1 + \kappa \|x_{ti_t}\|^2_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right) && \text{(Definition of } \kappa) \\
&\geqslant \left(\lambda^{\text{MLE}} \right)^d \left(1 + \frac{\kappa}{2} \right)^{|\mathcal{T}_0^{\text{MLE}}|}, && (t \in \mathcal{T}_0^{\text{MLE}})
\end{aligned}$$

where the third inequality holds because

$$
\begin{aligned}
H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star) &:= \sum_{s=1}^{t-1} \nabla^2 \ell_s(\boldsymbol{\nu}_t^\star) + \lambda^{\mathrm{MLE}} \mathbf{I}_d \\
&= \sum_{s=1}^{t-1} \left(\sum_{i \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) x_{si} x_{si}^\top - \sum_{i \in S_s} \sum_{j \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) p_s(j|S_s, \boldsymbol{\nu}_t^\star) x_{si} x_{sj}^\top \right) + \lambda^{\mathrm{MLE}} \mathbf{I}_d \\
&= \sum_{s=1}^{t-1} \left(\sum_{i \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) x_{si} x_{si}^\top - \frac{1}{2} \sum_{i \in S_s} \sum_{j \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) p_s(j|S_s, \boldsymbol{\nu}_t^\star) \left(x_{si} x_{sj}^\top + x_{sj} x_{si}^\top \right) \right) + \lambda^{\mathrm{MLE}} \mathbf{I}_d \\
&\geq \sum_{s=1}^{t-1} \left(\sum_{i \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) x_{si} x_{si}^\top - \frac{1}{2} \sum_{i \in S_s} \sum_{j \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) p_s(j|S_s, \boldsymbol{\nu}_t^\star) \left(x_{si} x_{si}^\top + x_{sj} x_{sj}^\top \right) \right) + \lambda^{\mathrm{MLE}} \mathbf{I}_d \\
&\qquad\qquad (xx^\top + yy^\top \succeq xy^\top yx^\top \text{ for any } x, y \in \mathbb{R}^d) \\
&= \sum_{s=1}^{t-1} \sum_{i \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) \left(1 - \sum_{j \in S_s} p_s(j|S_s, \boldsymbol{\nu}_t^\star) \right) x_{si} x_{si}^\top + \lambda^{\mathrm{MLE}} \mathbf{I}_d \\
&= \sum_{s=1}^{t-1} \sum_{i \in S_s} p_s(i|S_s, \boldsymbol{\nu}_t^\star) p_s(0|S_s, \boldsymbol{\nu}_t^\star) x_{si} x_{si}^\top + \lambda^{\mathrm{MLE}} \mathbf{I}_d \\
&\geq \sum_{s=1}^{t-1} \tilde{x}_{si_s} \tilde{x}_{si_s}^\top + \lambda^{\mathrm{MLE}} \mathbf{I}_d \succeq V_t.
\end{aligned}
$$

Thus, we have

$$
|\mathcal{T}_0^{\mathrm{MLE}}| \leqslant \frac{d}{\log(1+\kappa/2)} \log\left(1 + \frac{T}{d\lambda^{\mathrm{MLE}}}\right). \tag{F.2}
$$

Hence, by Equation (F.2), we can decompose the regret as follows:

$$
\begin{aligned}
\mathbf{Reg}_T(\mathbf{w}^\star) = \sum_{t=1}^{T} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) &\leqslant \left|\mathcal{T}_0^{\mathrm{MLE}}\right| + \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) \\
&\leqslant \frac{d}{\log(1+\kappa/2)} \log\left(1 + \frac{T}{d\lambda^{\mathrm{MLE}}}\right) + \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star),
\end{aligned} \tag{F.3}
$$

Next, we bound the second term of Equation (F.3). Let $\mathrm{UCB}_{ti} = x_{ti}^\top \widehat{\mathbf{w}}_t + \gamma_t^{\mathrm{MLE}}(\delta) \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}$ and $\overline{\mathrm{UCB}}_{ti}$ as $\overline{\mathrm{UCB}}_{ti} := x_{ti}^\top \mathbf{w}^\star + 2\gamma_t^{\mathrm{MLE}}(\delta) \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}$. Then, for all $i \in [N]$ and $t \geqslant 1$, we have

$$
x_{ti}^\top \tilde{\mathbf{w}}_{ti} - x_{ti}^\top \mathbf{w}^\star \leqslant \mathrm{UCB}_{ti} - x_{ti}^\top \mathbf{w}^\star \leqslant 2\gamma_t^{\mathrm{MLE}}(\delta) \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}, \qquad \text{(Definition of } \tilde{\mathbf{w}}_{ti}\text{, Lemma F.2)}
$$

which implies $x_{ti}^\top \tilde{\mathbf{w}}_{ti} \leqslant \overline{\mathrm{UCB}}_{ti}$. Thus, by Lemma D.5, we get

$$
\tilde{R}_t^{\mathrm{MLE}}(S_t) \leqslant \tilde{\bar{R}}_t^{\mathrm{MLE}}(S_t), \tag{F.4}
$$

where $\tilde{\bar{R}}_t^{\mathrm{MLE}}(S_t) := \frac{\sum_{i \in S_t} \exp(\overline{\mathrm{UCB}}_{ti}) r_{ti}}{1 + \sum_{j \in S_t} \exp(\mathrm{UCB}_{tj})}$.

Define a function $\tilde{Q} : \mathbb{R}^{|S_t|} \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1, \ldots, u_{|S_t|})^\top \in \mathbb{R}^{|S_t|}$, $\tilde{Q}(\mathbf{u}) = \sum_{k=1}^{|S_t|} \frac{\exp(u_k) r_{ti_k}}{1 + \sum_{j=1}^{|S_t|} \exp(u_j)}$. For simplicity, we write $S_t = \{i_1, \ldots, i_{|S_t|}\}$. Furthermore, we denote $\mathbf{u}_t = (u_{ti_1}, \ldots u_{ti_{|S_t|}})^\top = (\overline{\mathrm{UCB}}_{ti_1}, \ldots, \overline{\mathrm{UCB}}_{ti_{|S_t|}})^\top$

and $\mathbf{u}_t^\star = (u_{ti_1}^\star, \dots u_{ti_{|S_t|}}^\star)^\top = (x_{ti_1}^\top \mathbf{w}^\star, \dots, x_{ti_{|S_t|}}^\top \mathbf{w}^\star)^\top$. Then, we have

$$\begin{aligned}
\sum_{t \notin \mathcal{T}_0^{\text{MLE}}} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) &\leqslant \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \tilde{R}_t^{\text{MLE}}(S_t) - R_t(S_t, \mathbf{w}^\star) && \text{(Lemma D.4)} \\
&\leqslant \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \tilde{\tilde{R}}_t^{\text{MLE}}(S_t) - R_t(S_t, \mathbf{w}^\star) && \text{(Eqn. (F.4))} \\
&= \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \tilde{Q}(\mathbf{u}_t) - \tilde{Q}(\mathbf{u}_t^\star) \\
&= \underbrace{\sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \nabla \tilde{Q}(\mathbf{u}_t^\star)^\top (\mathbf{u}_t - \mathbf{u}_t^\star)}_{I_3} + \underbrace{\frac{1}{2} \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} (\mathbf{u}_t - \mathbf{u}_t^\star)^\top \nabla^2 \tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t - \mathbf{u}_t^\star)}_{I_4}, && \text{(F.5)}
\end{aligned}$$

where $\bar{\mathbf{u}}_t = (\bar{u}_{ti_1}, \dots, \bar{u}_{ti_{|S_t|}})^\top \in \mathbb{R}^{|S_t|}$ is the convex combination of $\mathbf{u}_t$ and $\mathbf{u}_t^\star$.

We first bound the term I_3. For simplicity, let $\mathbb{E}_t^{\mathbf{w}}[x_{ti}] = \mathbb{E}_{j \sim p_t(\cdot|S_t, \mathbf{w})}[x_{ti}]$, and $\mathbb{E}_t^{\mathbf{w}}[r_{ti}] = \mathbb{E}_{j \sim p_t(\cdot|S_t, \mathbf{w})}[r_{ti}]$. Then, we get

$$\begin{aligned}
&\sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \nabla \tilde{Q}(\mathbf{u}_t^\star)^\top (\mathbf{u}_t - \mathbf{u}_t^\star) \\
&= \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \sum_{i \in S_t} \frac{\exp(x_{ti}^\top \mathbf{w}^\star) r_{ti}}{1 + \sum_{k \in S_t} \exp(x_{tk}^\top \mathbf{w}^\star)} (u_{ti} - u_{ti}^\star) - \sum_{j \in S_t} \frac{\exp(x_{tj}^\top \mathbf{w}^\star) r_{tj} \sum_{i \in S_t} \exp(x_{ti}^\top \mathbf{w}^\star)}{(1 + \sum_{k \in S_t} \exp(x_{tk}^\top \mathbf{w}^\star))^2} (u_{ti} - u_{ti}^\star) \\
&= \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) r_{ti} \left(2\gamma_t^{\text{MLE}}(\delta) \|x_{ti}\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} - \sum_{j \in S_t} p_t(j|S_t, \mathbf{w}^\star) 2\gamma_t^{\text{MLE}}(\delta) \|x_{tj}\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right) \\
&\leqslant 2\gamma_T^{\text{MLE}}(\delta) \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) r_{ti} \left(\|x_{ti}\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} - \sum_{j \in S_t} p_t(j|S_t, \mathbf{w}^\star) \|x_{tj}\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right) \\
&= 2\gamma_T^{\text{MLE}}(\delta) \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \mathbb{E}_t^{\mathbf{w}^\star} \left[\left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left(\|x_{ti}\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} - \mathbb{E}_t^{\mathbf{w}^\star} \left[\|x_{tj}\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right] \right) \right] && (x_{t0} = \mathbf{0}, r_{t0} = 0) \\
&\leqslant 2\gamma_T^{\text{MLE}}(\delta) \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \mathbb{E}_t^{\mathbf{w}^\star} \left[\left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left\| x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}] \right\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right] && \text{(Similar to Eqn.(D.7))}
\end{aligned}$$

We further decompose the last term as follows:

$$\begin{aligned}
&\sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \mathbb{E}_t^{\mathbf{w}^\star} \left[\left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left\| x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}] \right\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right] \\
&= \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \sum_{i \in S_t} \sqrt{p_t(i|S_t, \mathbf{w}^\star) p_t(i|S_t, \boldsymbol{\nu}_t^\star)} \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&+ \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \sum_{i \in S_t} \left(\sqrt{p_t(i|S_t, \mathbf{w}^\star)} - \sqrt{p_t(i|S_t, \boldsymbol{\nu}_t^\star)} \right) \sqrt{p_t(i|S_t, \mathbf{w}^\star)} \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&+ \sum_{t \notin \mathcal{T}_0^{\text{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left(\left\| x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}] \right\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} - \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right). && \text{(F.6)}
\end{aligned}$$

Then, the first term in (F.6) can be bounded as follows:

$$\begin{aligned}
&\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} \sqrt{p_t(i|S_t, \mathbf{w}^\star) p_t(i|S_t, \boldsymbol{\nu}_t^\star)} \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\leqslant \sqrt{\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \underbrace{\sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right)^2}_{=:\sigma_t^2}} \sqrt{\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\nu}_t^\star) \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|^2_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}} \\
&\hfill \text{(Cauchy-Schwarz inequality)} \\
&\leqslant \sqrt{\sum_{t=1}^T \sigma_t^2} \sqrt{2d \log\left(1 + \frac{T}{d\lambda^{\mathrm{MLE}}}\right)}. \qquad \text{(Lemma D.7)}
\end{aligned}$$

Note that when applying the elliptical potential lemma (Lemma D.7), we ensure that the condition $\|x_{ti}\|^2_{H_t(\boldsymbol{\nu}_t^\star)^{-1}} \leqslant \frac{1}{2}$ holds for all $t \notin \mathcal{T}_0^{\mathrm{MLE}}$. Additionally, the second term in (F.6) can be bounded as follows:

$$\begin{aligned}
&\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} \left(\sqrt{p_t(i|S_t, \mathbf{w}^\star)} - \sqrt{p_t(i|S_t, \boldsymbol{\nu}_t^\star)} \right) \sqrt{p_t(i|S_t, \mathbf{w}^\star)} \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\leqslant \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} \frac{|p_t(i|S_t, \mathbf{w}^\star) - p_t(i|S_t, \boldsymbol{\nu}_t^\star)|}{\sqrt{p_t(i|S_t, \mathbf{w}^\star)} + \sqrt{p_t(i|S_t, \boldsymbol{\nu}_t^\star)}} \sqrt{p_t(i|S_t, \mathbf{w}^\star)} \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\leqslant \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} |p_t(i|S_t, \mathbf{w}^\star) - p_t(i|S_t, \boldsymbol{\nu}_t^\star)| \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\leqslant 4\gamma_T^{\mathrm{MLE}}(\delta) \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \max_{i \in S_t} \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \max_{i \in S_t} \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \qquad \text{(Lemma F.3)} \\
&\leqslant 4\gamma_T^{\mathrm{MLE}}(\delta) \sqrt{\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \max_{i \in S_t} \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|^2_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}} \sqrt{\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \max_{i \in S_t} \|x_{ti}\|^2_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}} \qquad \text{(Cauchy-Schwarz inequality)} \\
&\leqslant \frac{4}{\sqrt{\kappa}} \gamma_T^{\mathrm{MLE}}(\delta) \sqrt{\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\nu}_t^\star) \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|^2_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}} \sqrt{\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \max_{i \in S_t} \|x_{ti}\|^2_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}}} \qquad \text{(Definition of } \kappa) \\
&\leqslant \frac{8}{\kappa} \gamma_T^{\mathrm{MLE}}(\delta) d \log\left(1 + \frac{T}{d\lambda^{\mathrm{MLE}}}\right). \qquad \text{(Lemma D.7 and D.11)}
\end{aligned}$$

Finally, we bound the last term in (F.6). Using the inequality $\|\mathbf{a}\| - \|\mathbf{b}\| \leqslant \|\mathbf{a} - \mathbf{b}\|$ for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we get

$$\begin{aligned}
&\sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \left(r_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[r_{tj}] \right) \left(\left\| x_{ti} - \mathbb{E}_t^{\mathbf{w}^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} - \left\| x_{ti} - \mathbb{E}_t^{\boldsymbol{\nu}_t^\star}[x_{tj}] \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \right) \\
&\leqslant \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \left\| \sum_{j \in S_t} (p_t(j|S_t, \boldsymbol{\nu}_t^\star) - p_t(j|S_t, \mathbf{w}^\star)) x_{tj} \right\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\leqslant 4\gamma_T^{\mathrm{MLE}}(\delta) \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \sum_{i \in S_t} |p_t(i|S_t, \boldsymbol{\nu}_t^\star) - p_t(i|S_t, \mathbf{w}^\star)| \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\leqslant 4\gamma_T^{\mathrm{MLE}}(\delta) \sum_{t \notin \mathcal{T}_0^{\mathrm{MLE}}} \max_{i \in S_t} \|x_{ti}\|^2_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \qquad \text{(Lemma F.3)} \\
&\leqslant \frac{8}{\kappa} \gamma_T^{\mathrm{MLE}}(\delta) d \log\left(1 + \frac{T}{d\lambda^{\mathrm{MLE}}}\right). \qquad \text{(Lemma D.11)}
\end{aligned}$$

By combining the three results above, we can establish a bound for I_3.

$$\sum_{t\notin\mathcal{T}_0^{\text{MLE}}} \nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t-\mathbf{u}_t^\star) \leqslant 2\gamma_T^{\text{MLE}}(\delta)\sqrt{\sum_{t=1}^T\sigma_t^2}\sqrt{2d\log\left(1+\frac{T}{d\lambda^{\text{MLE}}}\right)} + \frac{36}{\kappa}\gamma_T^{\text{MLE}}(\delta)^2 d\log\left(1+\frac{T}{d\lambda^{\text{MLE}}}\right). \tag{F.7}$$

On the other hand, the term I_4 in Equation (F.5) can be bounded by following the same process as in Equation (D.10) in Appendix D.

$$\frac{1}{2}\sum_{t\notin\mathcal{T}_0^{\text{MLE}}} (\mathbf{u}_t-\mathbf{u}_t^\star)^\top\nabla^2\tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t-\mathbf{u}_t^\star) \leqslant \frac{20}{\kappa}\gamma_T^{\text{MLE}}(\delta)^2 d\log\left(1+\frac{T}{d\lambda^{\text{MLE}}}\right). \tag{F.8}$$

Plugging (F.7) and (F.8) into (F.3), and setting $\lambda^{\text{MLE}}=\frac{1}{8B^2}$ and $\gamma_T^{\text{MLE}}(\delta)=\mathcal{O}(\sqrt{d\log(BT)})$, we obtain

$$\begin{aligned}\sum_{t=1}^T R_t(S_t^\star,\mathbf{w}^\star)-R_t(S_t,\mathbf{w}^\star) &\leqslant 2\gamma_T^{\text{MLE}}(\delta)\sqrt{\sum_{t=1}^T\sigma_t^2}\sqrt{2d\log\left(1+\frac{T}{d\lambda^{\text{MLE}}}\right)}\\ &\quad+\frac{d}{\log(1+\kappa/2)}\log\left(1+\frac{T}{d\lambda^{\text{MLE}}}\right)+\frac{56}{\kappa}\gamma_T^{\text{MLE}}(\delta)^2 d\log\left(1+\frac{T}{d\lambda^{\text{MLE}}}\right)\\ &= \mathcal{O}\left(d\log(BT)\sqrt{\sum_{t=1}^T\sigma_t^2+\frac{1}{\kappa}d^2\left(\log(BT)\right)^2}\right).\end{aligned}$$

□

F.2. Proof of Lemmas for Theorem 4.12

F.2.1. Proof of Lemma F.2

Proof of Lemma F.2. By using a Taylor expansion and applying the first-order optimality condition for a convex function, we obtain

$$\begin{aligned}\mathcal{L}_t(\mathbf{w}^\star)-\mathcal{L}_t(\widehat{\mathbf{w}}_t) &= \underbrace{\langle\nabla\mathcal{L}_t(\widehat{\mathbf{w}}_t),\mathbf{w}^\star-\widehat{\mathbf{w}}_t\rangle}_{\geqslant 0,\text{ first order optimality condition}} + (\mathbf{w}^\star-\widehat{\mathbf{w}}_t)^\top\left(\int_0^1(1-v)\nabla^2\mathcal{L}_t(\widehat{\mathbf{w}}_t+v(\mathbf{w}^\star-\widehat{\mathbf{w}}_t))\mathrm{d}v\right)(\mathbf{w}^\star-\widehat{\mathbf{w}}_t)\\ &\geqslant \|\mathbf{w}^\star-\widehat{\mathbf{w}}_t\|^2_{\int_0^1(1-v)\nabla^2\mathcal{L}_t(\widehat{\mathbf{w}}_t+v(\mathbf{w}^\star-\widehat{\mathbf{w}}_t))\mathrm{d}v}\\ &= \frac{1}{2}\|\mathbf{w}^\star-\widehat{\mathbf{w}}_t\|^2_{\nabla^2\mathcal{L}_t(\boldsymbol{\nu}_t^\star)} && \text{(Definition of } \boldsymbol{\nu}_t^\star)\\ &\geqslant \frac{1}{2}\|\mathbf{w}^\star-\widehat{\mathbf{w}}_t\|^2_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)}-4B^2\lambda^{\text{MLE}}. && \text{(Definition of } H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star))\end{aligned}$$

By setting $L_t=2(t-1)$ and $\lambda^{\text{MLE}}=\frac{1}{8B^2}$, and applying Lemma F.1, we derive

$$\|\mathbf{w}^\star-\widehat{\mathbf{w}}_t\|^2_{H_t^{\text{MLE}}(\boldsymbol{\nu}_t^\star)} \leqslant 2\log\frac{1}{\delta}+2d\log\left(\max\left\{e,\frac{4eB(t-1)}{d}\right\}\right)+1 = 2\beta_t^{\text{MLE}}(\delta)^2+1,$$

which concludes the proof. □

F.2.2. PROOF OF LEMMA F.3

Proof of Lemma F.3. By the mean value theorem, there exists $\boldsymbol{\xi} = (1-c)\mathbf{w}_1 + c\mathbf{w}_2$ for some $c \in (0,1)$ such that

$$
\begin{aligned}
&\sum_{i \in S_t} \left| p_t(i|S_t, \mathbf{w}_1) - p_t(i|S_t, \mathbf{w}_2) \right| \omega_{ti} \\
&\quad = \sum_{i \in S_t} \left| \nabla p_t(i|S_t, \boldsymbol{\xi})^\top (\mathbf{w}_1 - \mathbf{w}_2) \right| \omega_{ti} \\
&\quad = \sum_{i \in S_t} \left| \left(p_t(i|S_t, \boldsymbol{\xi}) x_{ti} - p_t(i|S_t, \boldsymbol{\xi}) \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}) x_{tj} \right)^\top (\mathbf{w}_1 - \mathbf{w}_2) \right| \omega_{ti} \\
&\quad \leqslant \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}) \left| x_{ti}^\top (\mathbf{w}_1 - \mathbf{w}_2) \right| \omega_{ti} + \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}) \omega_{ti} \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}) \left| x_{tj}^\top (\mathbf{w}_1 - \mathbf{w}_2) \right| \\
&\quad \leqslant 2\gamma_t^{\mathrm{MLE}}(\delta) \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}) \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \omega_{ti} + 2\gamma_t^{\mathrm{MLE}}(\delta) \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}) \omega_{ti} \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}) \|x_{tj}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{C}_t^{\mathrm{MLE}}(\delta), \text{Lemma F.2}) \\
&\quad \leqslant 4\gamma_t^{\mathrm{MLE}}(\delta) \max_{i \in S_t} \omega_{ti} \max_{i \in S_t} \|x_{ti}\|_{H_t^{\mathrm{MLE}}(\boldsymbol{\nu}_t^\star)^{-1}},
\end{aligned}
$$

which concludes the proof. □

G. Experiment Details and Additional Results

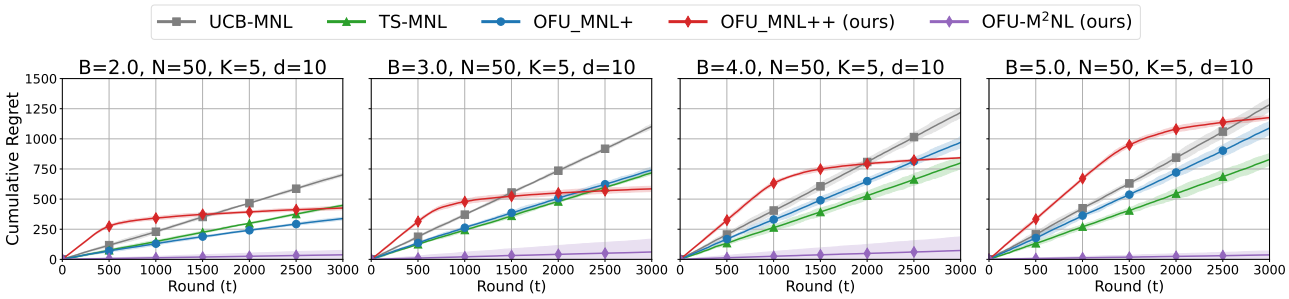

Figure G.1: Cumulative regret for different values of B when $d = 10$.

For each instance, we sample the true parameter $\mathbf{w}^\star$ uniformly from a d-dimensional Euclidean ball of radius B, denoted by $\mathbb{B}^d(B)$. Similarly, each context feature x_{ti} is independently and identically distributed (i.i.d.) from a unit ball, denoted as $\mathbb{B}^d(1)$. This ensures that $\|\mathbf{w}^\star\|_2 \leqslant B$ and $\|x_{ti}\|_2 \leqslant 1$, satisfying Assumption 3.1. The rewards are sampled independently in each round from a uniform distribution, i.e., $r_{ti} \sim \mathrm{Unif}(0,1)$. We set the number of items to $N = 50$ and the maximum assortment size to $K = 5$. For each instance, we conducted 20 independent runs and reported the average cumulative regret (Figures 1 and G.1) as well as the average runtime per round (Figure 2) for each algorithm. In our experiments, since the threshold τ_t is too conservative in practice, we empirically tuned the hyperparameter τ_t for `OFU-MNL++` by searching over a certain range of values while maintaining its inverse relationship with α (i.e., a higher τ_t corresponds to a lower α).

As an additional experiment, Figure G.1 presents results for a larger value of d, specifically $d = 10$. Our algorithms continue to outperform other baselines. While the performance of `OFU-MNL++` is somewhat sensitive to the values of B and d, primarily due to the adaptive warm-up rounds, its asymptotic performance appears to be the best. Notably, the slope of the regret curve is the smallest for large t. Additionally, `OFU-MNL++` enjoys a constant computational cost, similar to `OFU-MNL+`. In contrast, `OFU-M`^{2}`NL` is the slowest among the algorithms, as it requires solving a convex optimization problem to compute the optimistic parameter $\tilde{\mathbf{w}}_{ti}$, as described in Equation (F.1).