PASTA: Pessimistic Assortment Optimization

Juncheng Dong [*1] **Weibin Mo** [*2] **Zhengling Qi** [3] **Cong Shi** [4] **Ethan X. Fang** [5] **Vahid Tarokh** [1]

Abstract

We consider a class of assortment optimization problems in an *offline* data-driven setting. A firm does not know the underlying customer choice model but has access to an offline dataset consisting of the historically offered assortment set, customer choice, and revenue. The objective is to use the offline dataset to find an optimal assortment. Due to the combinatorial nature of assortment optimization, the problem of insufficient data coverage is likely to occur in the offline dataset. Therefore, designing a provably efficient offline learning algorithm becomes a significant challenge. To this end, we propose an algorithm referred to as *Pessimistic ASsortment opTimizAtion* (PASTA for short) designed based on the principle of pessimism, that can correctly identify the optimal assortment by only requiring the offline data to cover the optimal assortment under general settings. In particular, we establish a regret bound for the offline assortment optimization problem under the celebrated multinomial logit model. We also propose an efficient computational procedure to solve our pessimistic assortment optimization problem. Numerical studies demonstrate the superiority of the proposed method over the existing baseline method.

1. Introduction

One of the most critical problems faced by a seller is to select products for presentation to potential buyers. Often faced with limited display spaces and storage costs in both brick-and-mortar and online retailing, the seller needs to carefully choose a set of products from the vast collection of all available products for displaying to its customers. In this line, the problem of selecting an *assortment*, i.e., a collection of products from all available products, in order to maximize the seller's revenue is the *assortment optimization problem*. Obviously, the choice behavior of customers (McFadden, 1981) is of great importance in the problem of assortment optimization. Without loss of generality, we assume the choice of each customer can be described by a preference vector θ^*. This subsumes the seminal multinomial logit (MNL) model (McFadden, 1973) which is arguably the most well-studied and widespread models in assortment optimization literature (Please see Section 6 for more details) (Talluri & van Ryzin, 2004; Caro & Gallien, 2007; Rusmevichientong et al., 2010; Davis et al., 2013; Chen et al., 2021a; Aouad et al., 2022).

In practice, θ^* is often unknown and needs to be estimated. Assuming no historical data of customers, *dynamic assortment optimization* adaptively learns θ^* in a trial-and-error fashion by updating the assortment and observing the subsequent choices of customers sequentially (Caro & Gallien, 2007; Chen et al., 2020; Rusmevichientong et al., 2020; Chen et al., 2021b; Li et al., 2022). Meanwhile, in our era of Big Data, companies often collect abundant customer data. Therefore, it is often in companies' best interest to learn from the existing (potentially massive) offline datasets rather than starting from scratch. Moreover, offline learning is beneficial since online exploration can sometimes be expensive or infeasible. Hence, we take the first stab to formally study the following important question faced by every seller.

Research Question: *Given a pre-collected offline dataset of historically offered assortment, customers choices, and revenue, how can we find an efficient and theoretically justified offline algorithm to estimate the optimal assortment set without unrealistic assumptions on the offline dataset?*

When the dataset is not adaptively collected, it is not uncommon to encounter the challenge of insufficient coverage of data. For estimators such as the maximum likelihood estimator (MLE) to approximate θ^* accurately, the offline dataset must include sufficiently many assortments and customer choices. In other words, the data-collecting process needs to sufficiently explore different assortments (by the

[*] Equal contribution [1] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705, United States. [2] Mitchell E. Daniels, Jr. School of Business, Purdue University, West Lafayette, IN 47907, United States. [3] Department of Decision Sciences, George Washington University, Washington, DC 20052, United States. [4] Herbert Business School, University of Miami, Coral Gables, FL 33146, United States. [5] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, United States.. Correspondence to: Zhengling Qi <qizhengling@gwu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

seller) and different choices (by the customers). This is unlikely to happen for offline datasets because the seller would not choose unreasonable assortments whose expected revenues are obviously suboptimal, and the customers would not choose products against their preferences.

Major Contributions. The main contribution of this work is two-fold. First, based on the principle of pessimism, we propose the *Pessimistic ASsortment opTimizAtion* (PASTA for short) framework, which correctly identifies the optimal assortment. In particular, our framework only requires that the offline dataset covers the optimal assortment set instead of all possible (combinatorially many) assortment sets. Second, we derive the first finite-sample regret bound for offline assortment optimization under the multinomial logit (MNL) model (Please see Section 6), one of the most widely used models for modeling customers' choices. We subsequently propose an algorithm, also with the name *PASTA*, that can efficiently solve the pessimistic assortment optimization problems. Experiments on the simulated datasets (so that θ^* is known) corroborate the efficacy of pessimistic assortment optimization.

Paper Organization. We briefly review the related work in Section 2 and the preliminaries in Section 3. We propose the pessimistic assortment optimization in Section 4. In Section 5, we present the theoretical results. In Section 6, we study pessimistic assortment optimization under the MNL model as a concrete example. In Section 7, we propose an algorithm that can solve the problem efficiently. We provide experimental results in Section 8, after which we conclude.

2. Related Work

Assortment Optimization. The assortment optimization problem under the MNL model without any constraints was first studied in (Talluri & van Ryzin, 2004). Then more complicated assortment optimization problems under various types of constraints, including space requirement (Rusmevichientong et al., 2009) and cardinality (Rusmevichientong et al., 2010), were considered. (Davis et al., 2013) proposed a linear programming (LP) formulation of the assortment optimization problem that includes several previous works as special cases corresponding to different constraints in the formulation of LP. This line of work assumes that the true parameters of the customer models are known (or at least can be accurately estimated) (Gallego & Topaloglu, 2014; Feldman & Topaloglu, 2015; Flores et al., 2019; Désir et al., 2020; Liu et al., 2020; Aouad et al., 2021). Another closely related line of work is *dynamic assortment optimization* (Caro & Gallien, 2007). In the setting of dynamic assortment optimization, the seller without any prior information about the customers, has finite selling horizons in which it observes the choices of customers and, based on the observed behaviors, optimize their assortments in an adaptive, trial-and-error fashion (Sauré & Zeevi, 2013; Wang et al., 2018; Chen et al., 2021a; Rusmevichientong et al., 2020; Chen et al., 2020; 2021b). In comparison with the online setting used in dynamic assortment optimization, our work departs from the existing literature by focusing on the offline setting where the seller only has collected datasets but not any control on the data-collecting process.

Pessimism in Offline Learning. The principle of pessimism has been successfully used in reinforcement learning (RL) for finding an optimal policy with pre-collected datasets. On the empirical side, it has helped with improving the performance of both the model-based approach and value-based approach in offline setting (e.g., Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020). The importance of pessimism has been analyzed and verified theoretically in the setting of RL (Jin et al., 2021; Fu et al., 2022). Our work main contribution is to take a pessimistic approach to assortment optimization problems and demonstrate its empirical and theoretical values. Moreover, our work differs from the above works by focusing on a decision-making problem with exponentially many choices.

3. Preliminary

Let $[N] \doteq \{1, 2, .., N\}$ denote the set of N distinct items. For each item i, a feature vector $x_i \in \mathbb{R}^d$ is available. Assume that $\{x_i\}_{i\in[N]}$ are fixed vectors. Denote the collection of all possible assortments under *consideration* by $\mathbb{S} \subseteq 2^{[N]}\backslash\{\varnothing\}$. For the offline data, we define a random vector (S, A, R) from each customer, where $S \subseteq [N]$ denotes an assortment presented to the customer, $A \in S \cup \{0\}$ denotes the item purchased by the customer for $A \in S$ ($A = 0$ where no purchase is made), and R denotes the corresponding revenue. The ultimate goal of assortment optimization is to find an optimal set of items $s^* \in \mathbb{S}$ for all customers to maximize the expected revenue. A specific goal of this work is to study how to leverage the offline data, which consists of i.i.d. samples of the random triplet (S, A, R) in order to learn an optimal assortment.

For the assortment optimization with offline data, a fundamental question is to estimate the expected revenue for an unexplored assortment $s \in \mathbb{S}$. This amounts to addressing the causal relationship between assortment and revenue. Under the celebrated potential outcome framework (Rubin, 1974), let the random variable $R(s)$ be the potential revenue under an intervention that the assortment is set to be $s \in \mathbb{S}$. Our goal is to find an optimal assortment

$$s^* \in \arg\max_{s\in\mathbb{S}} \mathbb{E}[R(s)].$$

Note that the expected potential revenue $\mathbb{E}[R(s)]$ defined in the counterfactual world may not be identifiable from the observed data without additional assumptions. Throughout this paper, we make the following standard consistency and un-confoundedness assumptions in causal inference.

Assumption 3.1. [CONSISTENCY] With probability one, the observed revenue coincides with the potential revenue of the observed assortment. That is, $R = R(S)$ almost surely.

Assumption 3.2. [UN-CONFOUNDEDNESS] The potential revenues are independent variables of the observed assortment, i.e., $\{R(s)\}_{s\in\mathbb{S}} \perp\!\!\!\perp S$.

Assumption 3.1 ensures that the observed revenue is consistent with the potential revenue of purchasing item A $(\neq 0)$, or no purchase if $A = 0$, under the observed assortment S. Assumption 3.2 rules out possible unobserved factors that could confound the causal effect of assortment on revenue[1].

Denote $\pi_S(s) \doteq \mathbb{P}(S = s)$ as the probability of observing assortment s in the offline data. To non-parametrically identify $\mathbb{E}[R(s)]$ for every $s \in \mathbb{S}$, we further require the following positivity assumption (Imbens & Rubin, 2015).

Assumption 3.3. [POSITIVITY EVERYWHERE] For all $s \in \mathbb{S}$, the probability $\pi(s)$ of observing assortment s is positive (i.e. $\pi(s) > 0$).

Assumption 3.3 requires that every assortment can be observed with a positive chance in the offline data. *This is a strong assumption that will be later relaxed it Assumption 5.1 (I), i.e., requiring positivity only at optimum.* With Assumptions 3.1–3.3, we can identify the effect of an assortment set via inverse propensity score weighting (Rosenbaum & Rubin, 1983): for any $s \in \mathbb{S}$,

$$\mathbb{E}[R(s)] = \mathbb{E}\left\{\frac{\mathbb{I}(S = s)R}{\pi_S(s)}\right\}, \tag{1}$$

where the expectation in the right-hand-side is taken with respect to the data distribution of (S, A, R). However, when the number of possible assortments $|\mathbb{S}|$ grows exponentially in N, Assumption 3.3 rarely holds for all $s \in \mathbb{S}$ in practice, given potentially limited offline data particularly when N is large. Moreover, when an assortment s corresponds to an inferior expected revenue $\mathbb{E}[R(s)]$, it may not be considered by the seller at all. As a consequence, the probability of observing such an assortment $\pi_S(s)$ is zero. These may prevent us from estimating (1) for every assortment $s \in \mathbb{S}$.

We may tempt to use the following identification strategy:

$$\begin{aligned}\mathbb{E}[R(s)] &= \mathbb{E}(R|S = s) \quad \text{(by Assumptions 3.1-3.2)}\\ &= \mathbb{E}[\mathbb{E}(R|S = s, A)] = \sum_{i\in s\cup\{0\}} \pi_A(i|s;\boldsymbol{x})r_{s,i}, \end{aligned} \tag{2}$$

where $\boldsymbol{x} \doteq \{x_j\}_{j\in[N]}$ are the features across items, $\pi_A(i|s;\boldsymbol{x}) \doteq \mathbb{P}(A = i|S = s)$ is the customer's choice probability (McFadden, 1973) of purchasing the i-th item given an assortment s, $r_{s,i} \doteq \mathbb{E}(R|S = s, A = i)$ is the conditional expected revenue given the assortment s with the i-th item being purchased. For ease of notation, we omit the features $\boldsymbol{x}$ in π_A when there is no confusion. Identifying $\mathbb{E}[R(s)]$ as above requires the knowledge of $\pi_A(i|s)$ and $r_{s,i}$, which can be learned from data. Although such an identification approach does not explicitly depend on $\pi_S(s)$, full identification of $\pi_A(i|s)$ and $r_{s,i}$ requires the positivity of $\pi_S(s)$ for every $s \in \mathbb{S}$ as assumed above.

Despite the aforementioned challenge of insufficient coverage over assortments, we argue that finding an optimal assortment s^* may not necessarily require $\pi_S(s) > 0$ everywhere but only at the optimal assortment s^*. In particular, based on (2), when computing

$$s^* \in \arg\max_{s\in\mathbb{S}} \sum_{i\in s\cup\{0\}} \pi_A(i|s)r_{s,i}, \tag{3}$$

we may not necessarily need to estimate $\pi_A(i|s)$ and $r_{s,i}$ well for $s \neq s^*$, as long as *sub-optimal assortments can be safely ruled out during the optimization*. Our insight is that the estimation of $\pi_A(i|s)$ and $r_{s,i}$ for the less seen assortment s in the data often incurs large errors. Deploying pessimism by taking the estimation error into consideration can rule out those assortments (Jin et al., 2021), while standard predict-then-optimize (Bertsimas & Kallus, 2020) or empirical maximization approaches (Zhao et al., 2012) may suffer from an overestimation of $\mathbb{E}[R(s)]$. Hence, in our proposed pessimistic assortment optimization framework, we only require the positivity at optimum $\pi_S(s^*) > 0$, which is a much weaker assumption than that of Assumption 3.3.

In this paper, we focus on handling the estimation error from $\pi_A(i|s)$ while assuming that $r_{s,i}$ is known. This is a typical assumption in the literature of assortment optimization (Talluri & van Ryzin, 2004; Davis et al., 2013; Flores et al., 2019; Aouad et al., 2021). Our framework can be naturally extended to the scenario where we need to estimate $r_{s,i}$'s. For optimization tractability, we further assume that $r_{s,i} = r_i$ that the expected revenue depends only on the purchased item but not on the underlying assortment. This assumption is reasonable in many applications where the revenue is a deterministic consequence of a purchased item. This can also be easily extended under our pessimism framework but could result in a more complicated assortment optimization problem.

Below, without loss of generality, we assume that $r_i \geqslant 0$ for $i \in [N]$, while $r_0 = 0$ (no purchase incurs zero revenue). For any vector x, let $x^\top$ and $||x||_2$ respectively denote the transpose and ℓ_2-norm of x. For any set A, let $|A|$ denote the cardinal number of A. For any two sequences $\{\varpi(n)\}_{n\geqslant 1}$

[1] With the observed features $\{x_j\}_{j\in[N]}$, it can be possible to relax Assumption 3.2 to a more plausible condition: the independence holds conditional on the observed features. However, for notation simplicity, without loss of generality, we consider Assumption 3.2.

and $\{\gamma(n)\}_{n\geqslant 1}$, we write $\varpi(n) \gtrsim \gamma(n)$ (resp. $\varpi(N) \lesssim \gamma(n)$) whenever there exist constants $c_1 > 0$ (respectively $c_2 > 0$ such that $\varpi(n) \geqslant c_1\gamma(n)$ (resp. $\varpi(n) \leqslant c_2\gamma(n)$). Moreover, we write $\varpi(n) \eqsim \gamma(n)$ whenever $\varpi(n) \gtrsim \gamma(n)$ and $\varpi(n) \lesssim \gamma(n)$.

4. Pessimistic Offline Assortment Optimization

In this section, we introduce our pessimistic offline assortment optimization framework. To this end, based on Eq. (2), we first estimate the choice probability $\pi_A(i|s)$ from offline data. Subsequently we calculate optimizing values in optimization problem (3) using a plug-in estimator of $\pi_A(\cdot)$. Consider a generic form of model $\pi_A(a|s;\theta^*,\boldsymbol{x})$ with the unknown true parameter θ^*. Again, for ease of notation, we omit the features $\boldsymbol{x}$ in π_A when there is no confusion. We remark that θ^* could be either finite-dimensional or infinite-dimensional. Given an offline dataset $\mathcal{D} = \{S_i, A_i, R_i\}_{i=1}^n$, where n is the sample size, one can estimate the model parameter θ^* via maximum likelihood estimator (MLE). Specifically, define the likelihood-based loss function $\widehat{L}_n(\theta)$ as

$$\widehat{L}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log \pi_A(A_i|S_i;\theta).$$

Then the MLE of the unknown parameter θ^* is $\widehat{\theta}_{\mathrm{ML},n} \in \arg\min_{\theta\in\Theta} \widehat{L}_n(\theta)$, where Θ is a pre-specified parameter space. Let

$$\mathcal{V}(s;\widehat{\theta}_{\mathrm{ML},n}) \doteq \sum_{i\in s} \pi_A(i|s;\widehat{\theta}_{\mathrm{ML},n})r_i.$$

Here, we define $\mathcal{V}(s;\theta^*)$ as the *value function* of s with the customer choice model for π_A depending on the parameter θ^*. The plug-in estimator of the optimal assortment based on (3) is

$$\widehat{s}_{\mathrm{ML},n} \in \arg\max_{s\in\mathbb{S}} \left\{\mathcal{V}(s;\widehat{\theta}_{\mathrm{ML},n})\right\}.$$

The MLE-based approach first plugs in the MLE of θ^*, and then directly optimizes the corresponding estimated value function.

As discussed before, a disadvantage of the above estimate-then-optimize approach is that the estimation error of $\widehat{\theta}_{\mathrm{ML},n}$ caused by insufficient data coverage may result in the overestimation of $\mathcal{V}(s;\theta^*)$, which will propagate to downstream optimization. Alternatively, we can quantify the estimation uncertainty by considering the following likelihood-ratio-test-based confidence region (Owen, 1990):

$$\Omega_n(\alpha_n) \doteq \{\theta\in\Theta : \widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) \leqslant \alpha_n\},$$

where $\alpha_n > 0$ is pre-specified. Later we analyze the MNL model as a special case (Please see Section 6). With α_n chosen as $\mathcal{O}(d/n)$, we establish in Theorem 6.1 that $\theta^* \in \Omega_n(\alpha_n)$ with high probability. Such a guarantee does not require any data coverage assumption on assortments.

For now on, for simplicity, we drop α_n and write Ω_n for $\Omega_n(\alpha_n)$ when there is no ambiguity.

In order to robustify assortment optimization against plug-in estimation errors, we consider a pessimistic version of (3) by taking the estimation uncertainty from Ω_n into account. Specifically, we propose the **Pessimistic ASsortment opTimizAtion (PASTA)** by solving

$$\widehat{s}_{\mathrm{PASTA},n} \in \arg\max_{s\in\mathbb{S}} \min_{\theta\in\Omega_n} \mathcal{V}(s;\theta). \tag{4}$$

Here, for a fixed assortment $s \in \mathbb{S}$, the inner layer of minimization computes the worst-case value among all possible model parameters θ within the confidence set Ω_n. In particular, if the estimated value $\mathcal{V}(s;\widehat{\theta}_{\mathrm{ML},n})$ for s is highly uncertain due to insufficient data coverage, the worst-case value $\min_{\theta\in\Omega_n}\mathcal{V}(s;\theta)$ is likely much smaller than $\mathcal{V}(s;\theta^*)$. In that case, the outer layer of (4) may prefer another assortment with a relatively higher worst-case value. In this way, the inner layer of (4) rules out those assortments with less frequency in the offline data. Hence, one essential advantage of such a strategy is that it avoids an overestimation of the value function. In other words, by the plug-in approach, with a non-negligible chance, the estimated value $\mathcal{V}(s;\widehat{\theta}_{\mathrm{ML},n})$ can be much larger than the truth $\mathcal{V}(s;\theta^*)$, which further leads to a possibly sub-optimal assortment but optimized by the MLE-based approach. In contrast, PASTA is aware of insufficient data coverage, and hence more pessimistic about those highly uncertain value estimates. In the next section, we theoretically analyze the advantage of the PASTA approach.

5. Theoretical Results

In this section, we show that the PASTA method (4) enjoys a generic regret guarantee under a weak assumption of *positivity at optimum* that is $\pi_S(s^*) > 0$. Specifically, given $\widehat{s}_{\mathrm{PASTA},n}$ in (4), we adopt the following regret as the performance metric to evaluate the PASTA's performance

$$\mathcal{R}(\widehat{s}_{\mathrm{PASTA},n}) = \mathcal{V}(s^*;\theta^*) - \mathcal{V}(\widehat{s}_{\mathrm{PASTA},n};\theta^*).$$

We aim to derive a regret bound for $\mathcal{R}(\widehat{s}_{\mathrm{PASTA},n})$ under generic conditions. Denote $L(\theta) = \mathbb{E}[-\log\pi_A(A|S;\theta)]$ as the population loss function. All detailed proofs can be found in the Appendix 9.

We first show that whenever $\theta^* \in \Omega_n$ that the confidence region covers the true parameter, the regret of the PASTA method can be calibrated by the worst-case estimation error among $\theta \in \Omega_n$ of the value function at the optimal assortment s^*.

Lemma 5.1. *Let $\widehat{s}_{\mathrm{PASTA},n}$ be the solution by the PASTA* *method defined in* (4)*. If* $\theta^* \in \Omega_n$*, then*

$$\mathcal{R}(\widehat{s}_{\mathrm{PASTA},n}) \leqslant \max_{\theta \in \Omega_n} \big\{ \mathcal{V}(s^*; \theta^*) - \mathcal{V}(s^*; \theta) \big\}.$$

Proof of Lemma 5.1.

$$\begin{aligned} &\mathcal{V}(s^*; \theta^*) - \mathcal{V}(\widehat{s}_{\mathrm{PASTA},n}; \theta^*) \\ &\leqslant \mathcal{V}(s^*; \theta^*) - \min_{\theta \in \Omega_n} \mathcal{V}(\widehat{s}_{\mathrm{PASTA},n}; \theta) \qquad (\text{by } \theta^* \in \Omega_n) \\ &\leqslant \mathcal{V}(s^*; \theta^*) - \min_{\theta \in \Omega_n} \mathcal{V}(s^*; \theta) \qquad (\text{by } \widehat{s}_{\mathrm{PASTA},n} \text{ solves (4)}) \\ &= \max_{\theta \in \Omega_n} \big\{ \mathcal{V}(s^*; \theta^*) - \mathcal{V}(s^*; \theta) \big\}. \qquad \square \end{aligned}$$

Lemma 5.1 highlights the benefit of pessimistic optimization. In particular, guarantees for the regret of estimate-then-optimize approaches require the control of estimation error uniformly over all $s \in \mathbb{S}$, that is, $\sup_{s \in \mathbb{S}} \big| \mathcal{V}(s; \widehat{\theta}_{\mathrm{ML},n}) - \mathcal{V}(s; \theta^*) \big|$, in order to control error caused by the greedy policy, i.e., $\big| \mathcal{V}(\widehat{s}_{\mathrm{ML},n}, \widehat{\theta}_{\mathrm{ML},n}) - \mathcal{V}(\widehat{s}_{\mathrm{ML},n}, \theta^*) \big|$. This may typically entail that the value function is estimated uniformly well across $s \in \mathbb{S}$. It further explains the reason why an estimate-then-optimize approach would require the positivity Assumption 3.3 for all $s \in \mathbb{S}$. In contrast, our Lemma 5.1 suggests that controlling the estimation error at s^* can be enough. Therefore, it is sufficient for our PASTA method to require the positivity only at optimum.

Next, we impose the following assumptions to obtain the regret guarantee of our algorithm.

Assumption 5.1.
(I) [POSITIVITY AT OPTIMUM] The probability of observing the optimal assortment is positive, that is, $\pi_S(s^*) > 0$.
(II) [LIKELIHOOD-BASED CONCENTRATION] For any $0 < \delta < 1$, with probability at least $1 - \delta$, we have: (1) $\theta^* \in \Omega_n$, and (2)

$$\sup_{\theta \in \Omega_n} \Big| L(\theta) - L(\theta^*) - \Big(\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \Big) \Big| \leqslant \alpha_n.$$

We emphasize that *PASTA only requires the positivity at optimum*. Compared to the positivity at all assortments in Assumption 3.3, our Assumption 5.1 (I) is much weaker and hence more plausible to be satisfied. Assumption 5.1 (II) is a generic condition for likelihood-based concentration. We later justify that (II) above indeed holds under the general MNL model in Theorem 6.2. In particular, Statement (1) of Part (II) requires the validity of the likelihood-ratio-test-based confidence region Ω_n while Statement (2) of Part (II) requires the concentration of the likelihood-based localized empirical process (van der Vaart & Wellner, 1996).

The positivity at optimum is associated with a finite constant $C_{s^*} = 1/\pi_s(s^*)$ related to the learning performance. We also denote $r_{s^*} \doteq \max_{j \in s^*} r_j$ as the largest possible revenue among all items in s^*. Notice that both constants C_{s^*} and r_{s^*} depend on the optimal assortment s^* only. In the following lemma, we establish the estimation error bound at the optimal assortment s^*.

Lemma 5.2. *Under Assumption 5.1, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*, we have for any* $\theta \in \Omega_n$*,*

$$\mathcal{V}(s^*; \theta^*) - \mathcal{V}(s^*; \theta) \lesssim r_{s^*} C_{s^*} \sqrt{\alpha_n}.$$

Combining Lemmas 5.1 and 5.2, we summarize the regret bound for PASTA in the following theorem.

Theorem 5.3. *Under Assumption 5.1, for any* $0 < \delta < 1$*, with probability* $1 - \delta$*, we have*

$$\mathcal{R}(\widehat{s}_{\mathrm{PASTA},n}) \lesssim r_{s^*} C_{s^*} \sqrt{\alpha_n}.$$

6. Application: Multinomial Logit Model

In this section, we consider the Multinomial Logit Model (MNL) for customer choices $\pi_A(a|s)$. This is one of the most widely used models in assortment optimization literature (Feng et al., 2022). Under the MNL model, we will verify Assumption 5.1 (II) and establish the regret bound for PASTA in this case.

Given the item-specific features $\{x_i\}_{i \in [N]}$, MNL assumes that customer's preference for the i-th item is proportional to $\exp(x_i^\top \theta^*)$, where $\theta^* \in \Theta$ is the underlying unknown parameter. Here, we assume that the parameter space $\Theta \subseteq \mathbb{R}^d$ is compact with $\theta_{\max} \doteq \sup_{\theta \in \Theta} \|\theta\|_2 < +\infty$. Given an assortment s, the customer choice probability under MNL is given by

$$\pi_A(i|s; \theta^*) = \frac{\exp(x_i^\top \theta^*)}{1 + \sum_{j \in s} \exp(x_j^\top \theta^*)}, \quad \forall i \in s. \tag{5}$$

Moreover, the probability of no-purchase is normalized to $\pi_A(0|s; \theta^*) = 1/(1 + \sum_{j \in s} \exp(x_j^\top \theta^*))$. Based on (3) and the MNL model (5), the objective function for assortment optimization can be written as

$$\mathcal{V}(s; \theta) = \frac{\sum_{i \in s} r_i \exp(x_i^\top \theta)}{1 + \sum_{i \in s} \exp(x_i^\top \theta)}.$$

We first justify Statement (1) of Assumption 5.1 under the MNL model. To this end, given the compactness of Θ, there exists a finite constant $C_A > 0$ such that for all $\theta \in \Theta$, $s \in \mathbb{S}$ and $i \in s$, we have $1/\pi_A(i|s; \theta) \leqslant C_A$.

Lemma 6.1. *Consider the MNL model* (5) *with a compact set* Θ*. Assume that* $\theta^* \in \Theta$*. For any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*, we have*

$$\widehat{L}_n(\theta^*) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) \lesssim \frac{C_A d}{n} \cdot \log \frac{\theta_{\max}}{\delta}.$$

Lemma 6.1 suggests that, with α_n chosen as $\frac{C_A d}{n} \log \frac{\theta_{\max}}{\delta}$, we can guarantee that $\theta^* \in \Omega_n$ with high probability, which justifies Statement (1) of Assumption 5.1 (II). In particular, the order of α_n is $\mathcal{O}(d/n)$. Notice that Lemma 6.1 does not depend on the distribution of S, which implies that no data coverage assumption on the observed assortments is required. The assumption that $\theta^* \in \Theta$ for Θ a compact set requires that given any assortment, every product has a chance of being selected by the customer in the data. This is a mild requirement as θ^* is always finite.

Next, we justify Statement (2) of Assumption 5.1 (II) in the following theorem.

Lemma 6.2. *Consider the MNL model* (5)*. Suppose conditions in Lemma 6.1 hold, and $L(\theta)$ and $L_n(\theta)$ are uniformly and strongly convex. Let $\alpha_n \eqsim \frac{C_A d}{n} \log \frac{\theta_{\max}}{\delta}$. For $0 < \delta < 1$, with probability $\geqslant (1 - \delta)$, we have*

$$\sup_{\theta \in \Omega_n} \left| L(\theta) - L(\theta^*) - \left(\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \right) \right| \leqslant \alpha_n.$$

Finally, with the Assumption of positivity only at optimum (Assumption 5.1 (I)), we can apply Theorem 5.3 to establish the regret bound for PASTA in MNL.

Theorem 6.3. *Consider the MNL model (given in Equation* (5)*). Assume that the conditions in Lemma 6.2 hold and that $\pi_S(s^*) > 0$. Fix a $\delta \in (0, 1)$. Suppose $\widehat{s}_{\mathrm{PASTA},n}$ is output of PASTA with $\alpha_n \eqsim \frac{C_A d}{n} \log \frac{\theta_{\max}}{\delta}$. Then with probability at least $1 - \delta$, we have*

$$\mathcal{R}(\widehat{s}_{\mathrm{PASTA},n}) \lesssim r_{s*} C_{s*} \sqrt{\frac{C_A d}{n} \cdot \log \frac{\theta_{\max}}{\delta}}.$$

We remark that under the MNL model (given in Equation (5)), the order of regret is $\mathcal{O}(\sqrt{d/n})$. This is due to the concentration rate of MNL's empirical likelihood ratio in Lemma 6.1. Such a rate of regret bound matches those in the literature under parametric model assumptions (Qian & Murphy, 2011; Mo & Liu, 2022). However, existing literature requires the positivity $\pi_S(s) > 0$ at every $s \in \mathbb{S}$. In contrast, Theorem 6.3 only requires positivity $\pi_S(s^*) > 0$ at the optimal assortment s^*. Furthermore, we can show that $\min_{i \in N} \pi_A(i \mid S = [N]; \theta) \leqslant 1/N$ for any $\theta \in \Theta$, which implies that $C_A \geqslant N$. Therefore, our regret is of order at least $\sqrt{N}$, where N is the total number of available items. It is an interesting problem to establish the minimax lower bound of offline assortment optimization in terms of N, n, d and the cardinal number of s^*. This will investigated in a subsequent work.

7. *PASTA* Algorithm

In this section, we propose an efficient algorithm for solving the max-min problem given in Optimization Problem (4) for the MNL model. Specifically, let

$$\mathcal{V}(s; \theta) = \sum_{i \in s} \frac{r_i \exp(x_i^\top \theta)}{1 + \sum_{j \in s} \exp(x_j^\top \theta)}$$

and given the confidence set Ω_n, we wish to solve

$$\max_{s \in \mathbb{S}} \min_{\theta \in \Omega_n} \mathcal{V}(s; \theta).$$

The proposed iterative algorithm is executed for a maximum of T iterations. At the t-th iteration, given s_t and θ_t from the previous iteration, we consecutively execute the following two steps:

- Step 1: Compute the optimal assortment s_{t+1} given θ_t (see Section 7.1).
- Step 2: Compute the optimal θ_{t+1} using s_{t+1} (see Section 7.2).

The corresponding pseudo-code is presented in Algorithm 1 below.

Algorithm 1 PASTA

```
Input: offline dataset $\{(S_i, A_i, R_i)\}_{i=1}^n$; $\alpha_n$; $\{r_i\}_{i=1}^N$; $\{x_i\}_{i=1}^N$; maximum number of iterations $T$
Output: the solution to pessimistic assortment optimization $\widehat{s}$
$\widehat{L}_n(\theta) \doteq -\frac{1}{n} \sum_{i=1}^n \log \pi_A(A_i | S_i; \theta)$
$\widehat{\theta}_{\mathrm{ML},n} \leftarrow \arg\min_{\theta \in \Theta} \widehat{L}_n(\theta)$
$\Omega_n \leftarrow \{\theta \in \Theta : \widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) \leqslant \alpha_n\}$
$t \leftarrow 0; \theta_t \leftarrow \widehat{\theta}_{\mathrm{ML},n}$        /* Initialize $\theta_0$ as $\widehat{\theta}_{\mathrm{ML},n}$ */
for $t = 1$ to $T$ do
    $s_t \leftarrow$ SolveLP($\theta_{t-1}, \{r_i\}_{i=1}^N, \{x_i\}_{i=1}^N$)
                    /* Section 7.1 */
    $\theta_t \leftarrow$ SolveGD($s_t, \Omega_n, \{r_i\}_{i=1}^N, \{x_i\}_{i=1}^N$)
                    /* Section 7.2 */
end for
$\hat{s} \leftarrow s_T$
```

7.1. Optimal Assortment Computation

Given the MNL model parameter θ_t, computing the assortment s_{t+1} that maximizes the expected revenue can be formulated as a linear programming (LP) problem.

Suppose that an assortment s can be represented by an N-dimensional binary vector $\gamma \in \{0, 1\}^N$ where $\gamma_j = 1$ if and only if $j \in s$. Suppose that $s \in \mathbb{S}$ corresponds to the following feasible set for γ with M linear inequality constraints:

$$\Gamma = \left\{ \gamma \in \{0, 1\}^N : \sum_{j \in N} a_{ij} \gamma_j \leqslant b_i \text{ for } i \in [M] \right\},$$

where the matrix of constraint coefficients $[a_{ij}]_{i \in [M], j \in [N]}$ is a totally unimodular matrix (Pang, 2017). In other words, based on the one-to-one correspondence between s and γ, we have $s \in \mathbb{S}$ if and only if $\gamma \in \Gamma$.

Next, we denote $v_i = \exp(x_i^\top \theta_t)$ as the preference score for the i-th item. The customer choice probability under the MNL model (5) becomes $\pi_A(i|s) = \frac{v_i}{1+\sum_{j \in s} v_j}$. The optimization for s_{t+1} can be formulated as

$$\max_{\gamma \in \Gamma} \frac{\sum_{i \in [N]} r_i v_i \gamma_i}{1 + \sum_{i \in [N]} v_i \gamma_i}, \tag{6}$$

which is equivalent to the following linear programming problem (Davis et al., 2013):

$$\begin{aligned} \max_{w_j : j \in [N] \cup \{0\}} \quad & \sum_{j \in [N]} r_j w_j \\ \text{subject to} \quad & \sum_{j \in [N]} w_j + w_0 = 1 \\ & \sum_{j \in [N]} a_{ij} \frac{w_j}{v_j} \leqslant b_i w_0 \quad \forall i \in [M] \\ & 0 \leqslant \frac{w_j}{v_j} \leqslant w_0 \quad \forall j \in [N]. \end{aligned} \tag{7}$$

In particular, we can recover the optimal solution to Problem (6), denoted as γ^*, using the optimal solution to Problem (7), denoted by w^*, via the following formula:

$$\gamma_j^* = \frac{w_j^*}{v_j w_0^*} \quad \forall j \in [N]. \tag{8}$$

To conclude, at the t-th iteration, in order to compute an optimal assortment s_{t+1} for a given θ_t, we first solve an LP problem in (7) for w^*. Then we recover γ^* via (8). Finally, the updated assortment s_{t+1} is obtained by the correspondence $i \in s_{t+1}$ if and only if $\gamma_j^* = 1$.

7.2. Model Parameter Computation

For a given optimized assortment s_{t+1} from Section 7.1, we aim to search for the worst-case MNL parameter θ_{t+1} from the confidence set Ω_n that minimizes the expected revenue. In particular, we employ a gradient descent with line search (GDLS) method to compute θ_{t+1} by solving the following problem

$$\min_{\theta \in \Omega_n} \mathcal{V}(s_{t+1}; \theta). \tag{9}$$

Here, we remark that $\mathcal{V}(s_{t+1}; \theta) = \frac{\sum_{i \in s_{t+1}} r_i \exp(x_i^\top \theta)}{1+\sum_{i \in s_{t+1}} \exp(x_i^\top \theta)}$ is a locally Lipschitz function in θ. Given a feasible initial parameter $\theta^{(0)} \in \Omega_n$, we run at most L gradient descent steps. Suppose β_ℓ is the step size for gradient descent in the ℓ-th step. At each step $\ell = 1, 2, \cdots, L$, we do a line search to maintain the feasibility. In particular, given $\theta^{(\ell-1)} \in \Omega_n$, we first evaluate the gradient as $\xi_\ell = \nabla_\theta \mathcal{V}(s_{t+1}; \theta^{(\ell-1)})$. Then we initiate β_ℓ with a pre-specified step size $\beta_\ell = \widetilde{\beta}$, and check whether $\theta^{(\ell)} = \theta^{(\ell-1)} - \beta_\ell \xi_\ell$ is feasible, i.e. $\theta^{(\ell)} \in \Omega_n$. If not, we set $\beta_\ell \leftarrow c\beta_\ell$ for some pre-specified $c \in (0, 1)$, and recompute $\theta^{(\ell)} = \theta^{(\ell-1)} - \beta_\ell \xi_\ell$. Such a search is repeated until $\theta^{(\ell)}$ is feasible. We provide the pseudocode in Algorithm 2 for the overall process. Note that $L, \widetilde{\beta}, c$ are all hyper-parameters. In all of our numerical studies, we set $L = 2$, $\widetilde{\beta} = 0.01$ and $c = \frac{1}{2}$, which performs well empirically.

Algorithm 2 Gradient Descent with Line Search (GDLS)

```
Input: assortment $s_{t+1}$; feasible set $\Omega_n$; initial parameter $\theta^{(0)}$; initial step size $\widetilde{\beta}$; step shrinkage constant $c$; number of descent steps $L$
Output: the updated parameter $\theta_{t+1}$
$\ell \leftarrow 0$
for $\ell = 1$ to $L$ do
    $\xi_\ell = \nabla_\theta \mathcal{V}(s_{t+1}; \theta^{(\ell-1)})$    /* compute the gradient */
    $\beta_\ell \leftarrow \widetilde{\beta}$
    $\theta^{(\ell)} \leftarrow \theta^{(\ell-1)} - \beta_\ell \xi_\ell$
    while $\theta^{(\ell)} \notin \Omega_n$ do
        $\beta_\ell \leftarrow c\beta_\ell$    /* decrease the step size */
        $\theta^{(\ell)} \leftarrow \theta^{(\ell-1)} - \beta_\ell \xi_\ell$
    end while
end for
$\theta_{t+1} \leftarrow \theta^{(\ell)}$
```

8. Experiments

We compare the PASTA method with assortment optimization without pessimism (referred to as the *baseline* method in the sequel). Our method and the baseline method are evaluated on synthetic data for which the optimal assortment s^* and true parameter θ^* are known so that the true regrets can be computed. We describe the data generation process and the baseline method in details below.

8.1. Data Generation

We consider the assortment optimization scenarios described by N, K, d, n and p, where N is the total number of available products; K is the cardinality constraint of the assortments, i.e., $\mathbb{S} = \{s : |s| \leqslant K\}$; d is the dimension of θ^* and $\{x_j\}_{j=1}^N$; n is the sample size of the offline dataset; p is the probability for sampling the optimal assortment s^*. Similar to (Chen et al., 2020), we first generate the true preference vector θ^* as a uniformly random unit d-dim vector. For $i \in \{1, \ldots, N\}$, we generate r_i (the reward of product i) uniformly from the range $[0.5, 0.8]$ and generate x_i (the feature of product i) as uniformly random unit d-dim vector such that $\exp(x_i^\top \theta^*) \leqslant \exp(-0.6)$ to avoid degenerate cases, where the optimal assortments include too few items. Given such information, the true optimal assortment s^* can be computed. Then, we generate an offline dataset $\mathcal{D} = \{(S_i, A_i, R_i)\}_{i=1}^n$ with n samples. For $i \in \{1, \ldots, n\}$, we generate S_i following the distribution π_S

such that $\pi_S(s^*) = p$ and $\pi_S(s) = \frac{1-p}{|\mathbb{S}|-1}$, where $0 < p < 1$ is the probability of observing the optimal assortment s^*. After the assortment S_i is sampled, the customer choice (action) A_i is sampled according to the probability computed by MNL as in Eq. (5) with the true parameter θ^*.

8.2. Baseline

In our experiments, we use the gradient descent method to find $\widehat{\theta}_{\mathrm{ML},n}$ that minimizes the empirical negative log-likelihood function. Then given $\widehat{\theta}_{\mathrm{ML},n}$, the baseline method solves the assortment optimization problem by solving the linear programming problem in (7).

8.3. Performance Comparison

For a given (N, K, d, n, p), we repeat the data generation process in Section 8.1 to randomly generate 50 offline datasets. The solutions of PASTA and the baseline method are recorded in these experiments. For hyper-parameters, we set $\alpha_n = 2\widehat{L}_{\mathrm{ML}}$ where $\widehat{L}_{\mathrm{ML}} = \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n})$ and the maximum of iteration $T = 30$. We measure the performance with two metrics: (1) the *average regret* of the solutions which indicates how far the performance of the solutions is to that of the *optimal* performance (i.e., revenue of s^*); (2) the *assortment accuracy* of the solutions (with respect to the optimal assortment s^*). The assortment accuracy of an assortment s is defined as the ratio of the number of correctly chosen products to the number of products in s^*. The key results are summarized below.

Effect of Sample Size. We set $N = 40$, $K = 8$, $d = 16$ and $p = 0.9$. We then gradually increase the number of samples n . The result is presented in Figure 1 indicating that PASTA significantly outperforms the baseline method. While the performance of the baseline method improves with increasing number of samples, the PASTA method maintains a regret that is less than 25% of that of the baseline method. The same experiment repeated with an increased number of products ($N = 60$, $K = 15$) demonstrates that the gain of the PASTA method is stable, as presented in Figure 1.

Effect of Probability of Sampling Optimal Assortment in Offline Data. We set $N = 40$, $K = 8$, $d = 16$, $n = 150$, and let $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. We also study the effect of p in scenarios with an increased total number of products ($N = 60$, $K = 15$). As can be seen in Figure 2, the gain of pessimistic assortment optimization is consistent and robust for varying values of p.

Effect of Dimension of Features. We set $N = 20$, $K = 5$, $p = 0.9$, $n = 150$, and let $d \in \{8, 20, 32, 64, 128\}$. In order to characterize the effect of dimension d, we generate d elements of θ^* independently from Uniform$[-1, 1]$. The results are presented in Figure 3. We observe that while both the regret of the baseline method and that of the pessimistic assortment optimization increase with increasing dimensions of features, the PASTA method maintains its performance gain as the dimension d varies.

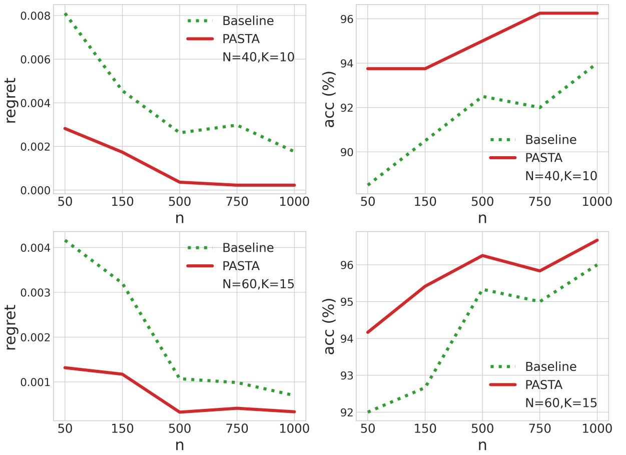

Figure 1. Performance comparison between PASTA and the baseline method with varying number of samples (n). On the left is the average regret (the lower the better) while the assortment accuracy (the higher the better) is on the right.

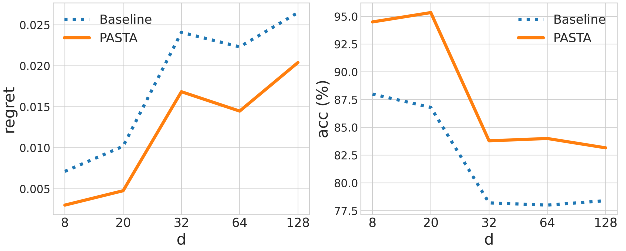

Figure 3. Comparison between PASTA and the baseline method with increasing dimensions of product features (d).

9. Conclusion

This work addresses the issue of insufficient data coverage in offline assortment optimization problems. This becomes more challenging as the number of choices grows quickly as a function of the number of items N. We presented a framework of pessimistic assortment optimization and provided theoretical justifications for our approach. We then performed an in-depth study of the Multinomial Logit Model (MNL), and derived a finite-sample regret bound of pessimistic assortment optimization for this popular model. We presented an efficient algorithm to solve the pessimistic assortment optimization problem for MNL, and demonstrated significant improvements of our approach over the baseline method by extensive numerical studies.

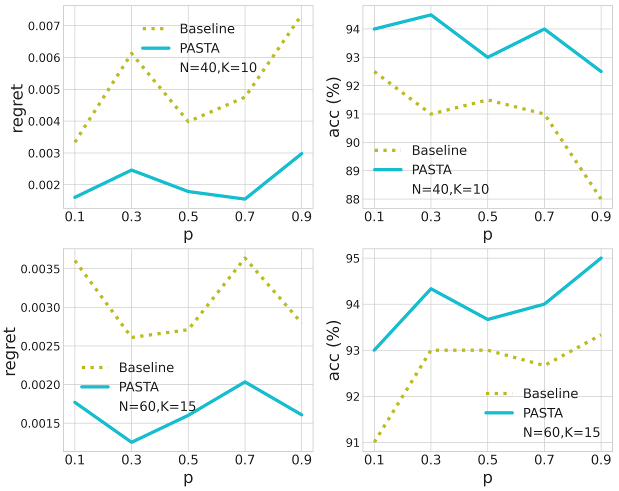

Figure 2. Comparison between PASTA and the baseline method with varying probability of the optimal assortment (p). Top row: $N = 40$; bottom row: $N = 60$.

Acknowledgements

Ethan X. Fang is partially supported by NSF DMS-2230795, NSF DMS-2230797. Cong Shi is partially supported by an Amazon Research Award.

References

Aouad, A., Farias, V., and Levi, R. Assortment optimization under consider-then-choose choice models. *Management Science*, 67(6):3368–3386, 2021.

Aouad, A., Feldman, J., and Segev, D. The exponomial choice model for assortment optimization: an alternative to the MNL model? *Management Science*, 2022.

Bertsimas, D. and Kallus, N. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

Caro, F. and Gallien, J. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment optimization with changing contextual information. *Journal of Machine Learning Research*, 2020.

Chen, X., Shi, C., Wang, Y., and Zhou, Y. Dynamic assortment planning under nested logit models. *Production and Operations Management*, 30(1):85–102, 2021a.

Chen, X., Wang, Y., and Zhou, Y. Optimal policy for dynamic assortment planning under multinomial logit models. *Mathematics of Operations Research*, 46(4): 1639–1657, 2021b.

Davis, J. M., Gallego, G., and Topaloglu, H. Assortment planning under the multinomial logit model with totally unimodular constraint structures, 2013. URL `https://people.orie.cornell.edu/jmd388/publications/MNLConstr.pdf`. Working Paper, Cornell University, Ithaca, NY.

Désir, A., Goyal, V., Segev, D., and Ye, C. Constrained assortment optimization under the Markov chain–based choice model. *Management Science*, 66(2):698–721, 2020.

Diaconis, P. and Saloff-Coste, L. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

Feldman, J. and Topaloglu, H. Bounding optimal expected revenues for assortment optimization under mixtures of multinomial logits. *Production and Operations Management*, 24(10):1598–1620, 2015.

Feng, Q., Shanthikumar, J. G., and Xue, M. Consumer choice models and estimation: A review and extension. *Production and Operations Management*, 31(2):847–867, 2022.

Flores, A., Berbeglia, G., and Van Hentenryck, P. Assortment optimization under the sequential multinomial logit model. *European Journal of Operational Research*, 273 (3):1052–1064, 2019.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. Offline reinforcement learning with instrumental variables in confounded markov decision processes, 2022. URL `https://arxiv.org/abs/2209.08666`. Working Paper, Northwestern University, Evanston, IL.

Gallego, G. and Topaloglu, H. Constrained assortment optimization for the nested logit model. *Management Science*, 60(10):2583–2601, 2014.

Harsha, P. Communication complexity, 2011. URL `https://www.tifr.res.in/~prahladh/teaching/2011-12/comm/lectures/l12.pdf`. Lecture Notes, Tata Institute of Fundamental Research (TIFR), Mumbai, India.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, Cambridge, UK, 2015.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096. PMLR, 18–24 Jul 2021.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

Li, S., Luo, Q., Huang, Z., and Shi, C. Online learning for constrained assortment optimization under Markov chain choice model, 2022. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4079753. Working Paper, University of Michigan, Ann Arbor, MI.

Liu, N., Ma, Y., and Topaloglu, H. Assortment optimization under the multinomial logit model with sequential offerings. *INFORMS Journal on Computing*, 32(3):835–853, 2020.

McFadden, D. *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, Berkeley, CA, 1973.

McFadden, D. Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272, 1981.

Mo, W. and Liu, Y. Efficient learning of optimal individualized treatment rules for heteroscedastic or misspecified treatment-free effect models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2): 440–472, 2022.

Owen, A. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

Pang, R. 1 totally unimodular matrices - stanford university, 2017. URL https://theory.stanford.edu/~jvondrak/MATH233B-2017/lec3.pdf.

Qian, M. and Murphy, S. A. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Rusmevichientong, P., Shen, M., and Shmoys, D. A PTAS for capacitated sum-of-ratios optimization. *Operations Research Letters*, 37:230–238, 07 2009.

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.

Rusmevichientong, P., Sumida, M., and Topaloglu, H. Dynamic assortment optimization for reusable products with random usage durations. *Management Science*, 66(7): 2820–2844, 2020.

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

Sen, B. A gentle introduction to empirical process theory and applications, 2018. URL http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf. Lecture Notes, Columbia University, New York, NY.

Talluri, K. and van Ryzin, G. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, New York, NY, 1996.

Wang, Y., Chen, X., and Zhou, Y. Near-optimal policies for dynamic multinomial logit assortment selection models. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

A. Proof of Theoretical Results

Throughout the proofs, we use θ^* to denote the true parameter and n to denote the number of samples. We use $\widehat{L}_n$ to denote the empirical negative log-likelihood function, i.e., $\widehat{L}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log \pi_A(A_i|S_i;\theta)$ where $\{(S_i, A_i, R_i)\}_{i=1}^n$ is the offline dataset. We use L and $\widehat{\theta}_{\mathrm{ML},n}$ to respectively denote the negative log-likelihood function , i.e., $L(\theta) = -\mathbb{E}[\log \pi_A(A|S;\theta)]$, and the MLE of θ^*, i.e., $\widehat{\theta}_{\mathrm{ML},n} \in \arg\min_{\theta\in\Theta}\left\{\widehat{L}_n(\theta)\right\}$. The confidence region Ω_n is defined as $\Omega_n \doteq \{\theta \in \Theta : \widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) \leqslant \alpha_n\}$.

For a general pair of random variables (X, Y), assume that the conditional probability density function of Y given X is parameterically modeled by $p(y|x;\theta)$ for parameter θ. For technical reasons, we will consider the following distances.

Definition A.1 (Squared Hellinger Distance)**.**

$$h^2(p(\cdot|x;\theta_1), p(\cdot|x;\theta_2)) = \frac{1}{2}\int\left(\sqrt{p_1(y|x;\theta_1)} - \sqrt{p_2(y|x;\theta_2)}\right)^2 dy. \tag{10}$$

Definition A.2 (Hellinger Distance)**.**

$$h(p(\cdot|x;\theta_1), p(\cdot|x;\theta_2)) = \sqrt{h^2(p(\cdot|x;\theta_1), p(\cdot|x;\theta_2))}. \tag{11}$$

Definition A.3 (Generalized Squared Hellinger Distance)**.**

$$H^2(\theta_1,\theta_2) = \mathbb{E}_X\Big[h^2(p(\cdot|X;\theta_1), p(\cdot|X;\theta_2))\Big]. \tag{12}$$

Definition A.4 (Generalized Hellinger Distance)**.**

$$H(\theta_1,\theta_2) = \mathbb{E}_X\Big[\sqrt{h^2(p(\cdot|X;\theta_1), p(\cdot|X;\theta_2))}\Big]. \tag{13}$$

In our theoretical results, we particularly consider $p(y|x;\theta) = \pi_A(a|s;\theta)$ as the conditional density of A given S (hereafter denoted as $A|S$).

A.1. Proof of Lemma 5.2

Under Assumption 5.1, for any $0 < \delta < 1$, with probability at least $1-\delta$, we have for any $\theta \in \Omega_n$,

$$\mathcal{V}(s^*;\theta^*) - \mathcal{V}(s^*;\theta) \lesssim\ r_{s^*} C_{s^*}\sqrt{\alpha_n}, \tag{14}$$

where $r_{s^*} \doteq \max_{j\in s^*} r_j$ is the largest possible revenue among all items in the optimal assortment.

Proof of Lemma 5.2. For any θ such that $\widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) \leqslant \alpha_n$, i.e., $\theta \in \Omega_n$, we have

$$\begin{aligned}
\mathcal{V}(s^*;\theta^*) - \mathcal{V}(s^*;\theta) &\leqslant \Big|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\theta^*)\Big| \\
&\leqslant \Big|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n}) + \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n}) - \mathcal{V}(s^*;\theta^*)\Big| \\
&\leqslant \Big|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n})\Big| + \Big|\mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n}) - \mathcal{V}(s^*;\theta^*)\Big| \\
&\leqslant 2\max_{\theta\in\Omega_n}\Big|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n})\Big| \qquad \text{(By Assumption 5.1 (II) } \theta^* \in \Omega_n\text{)}.
\end{aligned}$$

With Lemma B.2, we have that for any $\theta \in \Theta$,

$$\Big|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n})\Big| \leqslant r_{s^*} C_{s^*}\mathbb{E}_S\Big[||\pi_A(\cdot|S;\theta) - \pi_A(\cdot|S;\widehat{\theta}_{\mathrm{ML},n})||_1\Big]$$

where $||\cdot||_1$ is the ℓ_1-norm, $r_{s*} \doteq \max_{j\in s*} r_j$ is the largest possible revenue among all items and $C_{s*} = 1/\pi_S(s^*)$.

In Lemma B.3, we establish that

$$\mathbb{E}_S\left[||\pi_A(\cdot|S;\theta) - \pi_A(\cdot|S;\widehat{\theta}_{\mathrm{ML},n})||_1\right] \leqslant 2\sqrt{2}\sqrt{H^2(\theta, \widehat{\theta}_{\mathrm{ML},n})},$$

where H^2 is the generalized squared Hellinger distance defined in (12) with $p(y|x;\theta) = \pi_A(a|s;\theta)$ as the conditional density of $A|S$.

Combining the above two inequalities, we have that for any $\theta \in \Theta$,

$$\left|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML},n})\right| \lesssim r_{s*}C_{s*}\sqrt{H^2(\theta, \widehat{\theta}_{\mathrm{ML},n})}. \tag{15}$$

In the following, we use the fact that $\log x \leqslant 2(\sqrt{x}-1)$ for any $x \geqslant 0$ to show that for any $s \in \mathbb{S}$ and any θ:

$$\begin{aligned}
&-\int \pi_A(a|s;\theta^*)\log\frac{\pi_A(a|s;\theta)}{\pi_A(a|s;\theta^*)}da \\
&\geqslant -2\int \pi_A(a|s;\theta^*)\left(\sqrt{\frac{\pi_A(a|s;\theta)}{\pi_A(a|s;\theta^*)}}-1\right)da \\
&= \int \Big(\pi_A(a|s;\theta^*) + \pi_A(a|s;\theta) - 2\sqrt{\pi_A(a|s;\theta)\pi_A(a|s;\theta^*)}\Big)da \\
&= \int \Big(\sqrt{\pi_A(a|s;\theta^*)} + \sqrt{\pi_A(a|s;\theta)}\Big)^2 da \\
&\geqslant \int \Big(\sqrt{\pi_A(a|s;\theta^*)} - \sqrt{\pi_A(a|s;\theta)}\Big)^2 da,
\end{aligned}$$

which implies that

$$L(\theta) - L(\theta^*) \geqslant 2H^2(\theta;\theta^*). \tag{16}$$

By Lemma B.4, we have that for any $\theta \in \Omega_n$,

$$H^2(\theta, \widehat{\theta}_{\mathrm{ML},n}) \leqslant 2H^2(\theta^*, \theta) + 2H^2(\theta^*, \widehat{\theta}_{\mathrm{ML},n}) \leqslant L(\widehat{\theta}_{\mathrm{ML},n}) - L(\theta^*) + \{L(\theta) - L(\theta^*)\}. \tag{17}$$

From Assumption 5.1, we have that with probability at least $1-\delta$, for any $\theta \in \Omega_n$,

$$\left|L(\theta) - L(\theta^*) - \Big(\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*)\Big)\right| \leqslant \alpha_n.$$

In other words, under Assumption 5.1, with probability at least $1-\delta$ for any $\theta \in \Omega_n$,

$$L(\theta) - L(\theta^*) \leqslant \left|\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*)\right| + \alpha_n \leqslant 2\alpha_n. \tag{18}$$

Plugging Eq. (18) into Eq. (17), we have that with probability at least $1-\delta$, for any $\theta \in \Omega_n$,

$$H^2(\theta, \widehat{\theta}_{\mathrm{ML},n}) \leqslant 4\alpha_n. \tag{19}$$

Combining the above inequality and Eq. (15), we have that, with probability at least $1-\delta$, $\left|\mathcal{V}(s^*;\theta) - \mathcal{V}(s^*;\widehat{\theta}_{\mathrm{ML}}))\right| \lesssim r_{s*}C_{s*}\sqrt{\alpha_n}$ for all $\theta \in \Omega_n$. This concludes the proof. □

A.2. Proof of Lemma 6.1

Consider the MNL model (5) with a compact set Θ. Assume that $\theta^* \in \Theta$. For $0 < \delta < 1$, with probability at least $1-\delta$, we have

$$\widehat{L}_n(\theta^*) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) \lesssim \frac{C_A d}{n}\log\frac{\theta_{\max}}{\delta}. \tag{20}$$

Proof of Lemma 6.1. Fix $0 < \delta < 1$. Suppose $\alpha_n \eqsim \frac{C_A d}{n} \log \frac{2\theta_{\max}}{\delta}$. Define an oracle confidence set as

$$\widetilde{\Omega}_n \doteq \{\theta \in \Theta : L(\theta) - L(\theta^*) \leqslant \alpha_n\}.$$

In particular, $\theta^* \in \widetilde{\Omega}_n$. By Lemmas B.1 and A.2, we also have with probability at least $1 - \delta/2$,

$$L(\widehat{\theta}_{\mathrm{ML},n}) - L(\theta^*) \leqslant 2C_A H^2(\widehat{\theta}_{\mathrm{ML},n}, \theta^*) \lesssim \alpha_n,$$

that is, $\widehat{\theta}_{\mathrm{ML},n} \in \widetilde{\Omega}_n$.

Define $\mathcal{F}_n \doteq \left\{\log \frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)} : \theta \in \widetilde{\Omega}_n\right\}$. In particular, for any $f \in \mathcal{F}_n$, we have $\|f\|_\infty \leqslant 2 \log C_A$. Let $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_n} \doteq \sup_{f \in \mathcal{F}_n} |(\mathbb{P}_n - \mathbb{P})(f)|$ be the envelope. By Talagrand's inequality, with probability at least $1 - \delta/2$, we have for any $f \in \mathcal{F}_n$,

$$(\mathbb{P}_n - \mathbb{P})(f) \lesssim \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_n} + \sqrt{\frac{1}{n}\left\{(\log C_A)\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_n} + \sup_{f \in \mathcal{F}_n} \mathbb{E}(f - \mathbb{E}f)^2\right\}\log\frac{2}{\delta}} + \frac{\log C_A}{n}\log\frac{2}{\delta}. \tag{21}$$

For the variance term, we have

$$\begin{aligned}
\sigma^2_{\mathcal{F}_n} &\doteq \sup_{f \in \mathcal{F}_n} \mathbb{E}(f - \mathbb{E}f)^2 \leqslant \sup_{\theta \in \widetilde{\Omega}_n} \mathbb{E}\left(\log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}\right)^2 \\
&\leqslant \sup_{\theta \in \widetilde{\Omega}_n} \mathbb{E}\left\{2C_A h^2\big(\pi_A(\cdot|S;\theta^*), \pi_A(\cdot|S;\theta)\big)\right\} && \text{(by Lemma B.5)} \\
&= 2C_A \sup_{\theta \in \widetilde{\Omega}_n} H^2(\theta^*, \theta) \\
&\leqslant 2C_A \sup_{\theta \in \widetilde{\Omega}_n} [L(\theta) - L(\theta^*)] && \text{(by (17))} \\
&= 2C_A\alpha_n. && \text{(by definition of } \widetilde{\Omega}_n\text{)}
\end{aligned}$$

For the expected envelope, our goal below is to apply Sen (2018, Theorem 7.13) (stated in Theorem B.6). Consider the covering number $\mathcal{N}(\epsilon, \mathcal{F}_n, L^2(\mathbb{Q}))$ for any given $\epsilon > 0$ and finitely supported probability measure $\mathbb{Q}$. By Lemma A.1, based on the MNL model (5), for some $\mathsf{L} < +\infty$, $\mathcal{F}_n$ is a class of L-Lipschitz functions with respect to the index space $(\Theta, \|\cdot\|_2)$. Then in terms of the bracketing number $\mathcal{N}_{[]}$ and covering number $\mathcal{N}$, for any $\epsilon > 0$ and probability measure $\mathbb{Q}$, we have

$$\mathcal{N}(\epsilon\mathsf{L}, \mathcal{F}_n, L^2(\mathbb{Q})) \leqslant \mathcal{N}_{[]}(2\epsilon\mathsf{L}, \mathcal{F}_n, L^2(\mathbb{Q})) \leqslant \mathcal{N}(\epsilon, \Theta, \|\cdot\|_2).$$

By $\Theta \subseteq \mathbb{R}^d$ and Θ is compact, we further have $\mathcal{N}(\epsilon, \Theta, \|\cdot\|_2) \lesssim \left(\frac{1}{\epsilon}\right)^d$. Therefore,

$$\mathcal{N}(\epsilon, \mathcal{F}_n, L^2(\mathbb{Q})) \lesssim \left(\frac{\mathsf{L}}{\epsilon}\right)^d.$$

By Theorem B.6, we further have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_n} \lesssim \sqrt{\frac{d}{n}\sigma^2_{\mathcal{F}_n}\log\frac{\mathsf{L}}{\sigma_{\mathcal{F}_n}}} \vee \left\{\frac{d}{n} \times 2\log(C_A)\log\frac{\mathsf{L}}{\sigma_{\mathcal{F}_n}}\right\} \lesssim \sqrt{\frac{C_A d}{n}\alpha_n} \eqsim \frac{C_A d}{n}\sqrt{\log\frac{2\theta_{\max}}{\delta}} \lesssim \alpha_n.$$

The Talagrand's inequality (21) becomes

$$(\mathbb{P}_n - \mathbb{P})(f) \lesssim \alpha_n.$$

In particular, in the case of $\widehat{\theta}_{\mathrm{ML},n} \in \widetilde{\Omega}_n$ corresponding to $f(\widehat{\theta}_{\mathrm{ML},n}) \in \mathcal{F}_n$, we have

$$\widehat{L}_n(\theta^*) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}) = -\mathbb{P}_n f(\widehat{\theta}_{\mathrm{ML},n}) \lesssim -\mathbb{P}f(\widehat{\theta}_{\mathrm{ML},n}) + \alpha_n = \underbrace{L(\theta^*) - L(\widehat{\theta}_{\mathrm{ML},n})}_{\leqslant 0} + \alpha_n \leqslant \alpha_n.$$

This complete the proof. □

Lemma A.1. *Consider the MNL model:*

$$\pi_A(i|s;\theta) = \frac{\exp(x_i^\top\theta)}{1+\sum_{j\in s}\exp(x_j^\top\theta)}; \quad \pi_A(0|s;\theta) = \frac{1}{1+\sum_{j\in s}\exp(x_j^\top\theta)}; \quad \forall i \in s,\ s \in \mathbb{S},\ \theta \in \Theta.$$

Let $\theta_{\max} \doteq \max_{\theta\in\Theta}\|\theta\|_2$*,* $x_{\max} \doteq \max_{j\in[N]}\|x_j\|_2$*. If* Θ *is compact, that is,* $\theta_{\max} < +\infty$*, then the log-likelihood ratio* $\log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}$ *is a uniformly Lipschitz function in* $\theta \in \Theta$*.*

Proof of Lemma A.1.

$$\begin{aligned}\left\|\frac{\partial}{\partial\theta}\log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}\right\|_2 &= \frac{1}{\pi_A(A|S;\theta)}\left\|\frac{\partial\pi(A|S;\theta)}{\partial\theta}\right\|_2 \\ &= [1-\pi_A(A|S;\theta)]\left\|x_A + \sum_{j\in S}\exp(x_j^\top\theta)(x_A - x_j)\right\|_2 \\ &\leqslant x_{\max} + 2Nx_{\max}\exp(x_{\max}\theta_{\max}) \doteq \mathsf{L} < +\infty.\end{aligned} \tag{22}$$

That is, $\theta \mapsto \log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}$ is L-Lipschitz. □

Lemma A.2 (Concentration of Parametric MLE in Hellinger Distance)**.** *Consider the MNL model* (5)*. For* $0 < \delta < 1$*, with probability at least* $1-\delta$*, we have*

$$H^2(\widehat{\theta}_{\mathrm{ML},n},\theta^*) \lesssim \frac{d}{n}\log\frac{\theta_{\max}}{\delta}.$$

Proof of Lemma A.2. We follow from Fu et al. (2022, Corollary 2) as a special case, where our data are generated i.i.d. instead of being a general Markov chain. □

A.3. Proof of Lemma 6.2

Consider the MNL model (5). Suppose conditions in Lemma 6.1 hold, and $L(\theta)$ and $L_n(\theta)$ are uniformly and strongly convex. Let $\alpha_n \eqsim \frac{C_A d}{n}\log\frac{\theta_{\max}}{\delta}$. For $0 < \delta < 1$, with probability at least $1-\delta$, we have

$$\sup_{\theta\in\Omega_n}\left|L(\theta) - L(\theta^*) - \left(\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*)\right)\right| \leqslant \alpha_n.$$

Proof of Lemma 6.2. Fix $0 < \delta < 1$. By the strong convexity assumption on $L(\theta)$ and $L_n(\theta)$, there exists a constant $\mu > 0$ such that for any $\theta \in \Theta$,

$$\mu\|\theta - \theta^*\|_2^2 \leqslant L(\theta) - L(\theta^*); \quad \mu\|\theta - \widehat{\theta}_{\mathrm{ML},n}\|_2^2 \leqslant \widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n}).$$

By Lemma 6.1, with probability at least $1-\delta/2$, we have $\widehat{\theta}_{\mathrm{ML},n} \in \widetilde{\Omega}_n$. Then for any $\theta \in \Omega_n$, we have

$$\begin{aligned}&\|\theta - \theta^*\|_2 \leqslant \|\theta - \widehat{\theta}_{\mathrm{ML},n}\|_2 + \|\widehat{\theta}_{\mathrm{ML},n} - \theta^*\|_2 && \text{(by triangular inequality)} \\ &\leqslant \frac{1}{\sqrt{\mu}}\sqrt{\widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_{\mathrm{ML},n})} + \frac{1}{\sqrt{\mu}}\sqrt{L(\widehat{\theta}_{\mathrm{ML},n}) - L(\theta^*)} && \text{(by strong convexity)} \\ &\lesssim \sqrt{\alpha_n} && \text{(by } \theta \in \Omega_n \text{ and } \widehat{\theta}_{\mathrm{ML},n} \in \widetilde{\Omega}_n \text{ respectively).}\end{aligned}$$

The above implies that with probability at least $1-\delta/2$, we have $\Omega_n \subseteq \bar{\Omega}_n$, where $\bar{\Omega}_n$ is a ball centered around θ^* with radius $\sqrt{\alpha_n}$:

$$\bar{\Omega}_n \doteq \left\{\theta \in \Theta : \|\theta - \theta^*\|_2 \leqslant \sqrt{\alpha_n}\right\}.$$

Define $\bar{\mathcal{F}}_n \doteq \left\{\log \frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)} : \theta \in \bar{\Omega}_n\right\}$. Let $\|\mathbb{P}_n - \mathbb{P}\|_{\bar{\mathcal{F}}_n} \doteq \sup_{f \in \bar{\mathcal{F}}_n} |(\mathbb{P}_n - \mathbb{P})(f)|$ be the envelop. By Talagrand's inequality, with probability at least $1 - \delta/2$, we have for any $f \in \bar{\mathcal{F}}_n$,

$$|(\mathbb{P}_n - \mathbb{P})(f)| \lesssim \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\bar{\mathcal{F}}_n} + \sqrt{\frac{1}{n}\left\{(\log C_A)\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\bar{\mathcal{F}}_n} + \sup_{f \in \bar{\mathcal{F}}_n} \mathbb{E}(f - \mathbb{E}f)^2\right\}\log\frac{2}{\delta}} + \frac{\log C_A}{n}\log\frac{2}{\delta}. \tag{23}$$

For the variance term, we have

$$\begin{aligned}
\sigma^2_{\bar{\mathcal{F}}_n} \doteq \sup_{f \in \bar{\mathcal{F}}_n} \mathbb{E}(f - \mathbb{E}f)^2 &\leqslant \sup_{\theta \in \bar{\Omega}_n} \mathbb{E}\left(\log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}\right)^2 \\
&\lesssim \sup_{\theta \in \bar{\Omega}_n} \|\theta - \theta^*\|_2^2 && \text{(by Lipschitzness in Lemma A.1)} \\
&\leqslant \alpha_n && \text{(by definition of } \bar{\Omega}_n\text{)}.
\end{aligned}$$

For the expected envelope, by Theorem B.6, we further have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\bar{\mathcal{F}}_n} \lesssim \sqrt{\frac{d}{n}\sigma^2_{\bar{\mathcal{F}}_n}\log\frac{\mathsf{L}}{\sigma_{\bar{\mathcal{F}}_n}}} \vee \left\{\frac{d}{n} \times 2\log(C_A)\log\frac{\mathsf{L}}{\sigma_{\bar{\mathcal{F}}_n}}\right\} \lesssim \sqrt{\frac{d}{n}\alpha_n} \lesssim \alpha_n.$$

Therefore, the Talagrand's inequality (23) gives that for any $f \in \bar{\mathcal{F}}_n$

$$|(\mathbb{P}_n - \mathbb{P})(f)| \lesssim \alpha_n.$$

In other words, with probability at least $1 - \delta$, $\Omega_n \subseteq \bar{\Omega}_n$ corresponding to $\{f(\theta)\}_{\theta \in \Omega_n} \subseteq \bar{\mathcal{F}}_n$, we have

$$\sup_{\theta \in \Omega_n}\left|L(\theta) - L(\theta^*) - \left(\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*)\right)\right| \leqslant \sup_{f \in \bar{\mathcal{F}}_n}|(\mathbb{P}_n - \mathbb{P})(f)| \lesssim \alpha_n.$$

□

B. Technical Lemmas

Lemma B.1. *Suppose $C_A > 2$ and $C_A \geqslant 1/\pi_A(i|s;\theta)$ for all $\theta \in \Theta$, $s \in \mathbb{S}$ and $i \in s$. Then for any $\theta \in \Theta$:*

$$|L(\theta) - L(\theta^*)| \leqslant 2C_A H^2(\theta, \theta^*), \tag{24}$$

Proof of Lemma B.1. By definition,

$$|L(\theta^*) - L(\theta)| = \left|\mathbb{E}\left\{\mathbb{E}_A\left[\log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}\middle|S\right]\right\}\right|.$$

In particular, for a fixed $s \in [N]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\log\frac{\pi_A(A|S;\theta^*)}{\pi_A(A|S;\theta)}\middle|S = s\right] = \mathrm{KL}\Big(\pi_A(\cdot|s;\theta^*)\Big\|\pi_A(\cdot|s;\theta)\Big) \\
&\leqslant \frac{\log(C_A - 1)}{1 - 2/C_A}\Big\{1 - \big[1 - h^2(\pi_A(\cdot|s;\theta^*), \pi_A(\cdot|s;\theta))\big]^2\Big\} \\
&\quad \text{(by log-Sobolev inequality (Diaconis \& Saloff-Coste, 1996, Theorem A.1))} \\
&= \frac{C_A\log(C_A - 2 + 1)}{C_A - 2}\{2h^2 - (h^2)^2\} \ \Big(\text{with } h^2 = h^2\big(\pi_A(\cdot|s;\theta^*), \pi_A(\cdot|s;\theta)\big)\Big) \\
&\leqslant 2C_A h^2(\pi_A(\cdot|s;\theta^*), \pi_A(\cdot|s;\theta)).
\end{aligned}$$

Therefore,

$$|L(\theta^*) - L(\theta)| \leqslant \mathbb{E}\left\{2C_A h^2(\pi_A(\cdot|S;\theta^*), \pi_A(\cdot|S;\theta))\right\} = 2C_A H^2(\theta^*, \theta).$$

□

Lemma B.2. *Let $C_{s*} \doteq \frac{1}{\pi_S(s*)}$ and $r_{s*} \doteq \max_{j \in s*} r_j$, then the following inequality holds for any $\theta_1, \theta_2 \in \Theta$:*

$$\left|\mathcal{V}(s*;\theta_1) - \mathcal{V}(s*;\theta_2)\right| \leqslant r_{s*} C_{s*} \mathbb{E}_S\left[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1\right]$$

where $||\cdot||_1$ denotes the L^1 norm.

Proof of Lemma B.2.

$$\left|\mathcal{V}(s*;\theta_1) - \mathcal{V}(s*;\theta_2)\right| = \left|\mathbb{E}_S\left[\frac{\mathbb{I}(S=s*)}{\pi_S(S)}\sum_{i \in S} r_{i,S}\Big(\pi_A(i|S;\theta_1) - \pi_A(i|S;\theta_2)\Big)\right]\right| \tag{25}$$

$$\leqslant \mathbb{E}_S\left[\left|\frac{\mathbb{I}(S=s*)}{\pi_S(S)}\sum_{i \in S} r_{i,S}\Big(\pi_A(i|S;\theta_1) - \pi_A(i|S;\theta_2)\Big)\right|\right] \tag{26}$$

$$\leqslant \left|\left|\frac{\mathbb{I}(S=s*)}{\pi_S(S)}\right|\right|_\infty \mathbb{E}_S\left[\mathbb{I}(S=s*)\left|\sum_{i \in S} r_{i,S}\Big(\pi_A(i|S;\theta_1) - \pi_A(i|S;\theta_2)\Big)\right|\right] \tag{27}$$

$$\leqslant C_{s*}\mathbb{E}_S\left[\mathbb{I}(S=s*)\sum_{i \in S}\left|r_{i,S}\Big(\pi_A(i|S;\theta_1) - \pi_A(i|S;\theta_2)\Big)\right|\right] \tag{28}$$

$$\leqslant r_{s*}C_{s*}\mathbb{E}_S\left[\sum_{i \in S}\left|\Big(\pi_A(i|S;\theta_1) - \pi_A(i|S;\theta_2)\Big)\right|\right] \tag{29}$$

$$= r_{s*}C_{s*}\mathbb{E}_S\left[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1\right], \tag{30}$$

where Eq. (25) comes from the sample-based estimation of $\mathbb{E}[R(s)]$ (Eq. (1)), Eq. (27) comes from the Hölder's inequality, Eq. (28) comes from the fact that $\left|\left|\frac{\mathbb{I}(S=s*)}{\pi_S(S)}\right|\right|_\infty = C_{s*}$ because $\frac{\mathbb{I}(S=s*)}{\pi_S(S)}$ has the value zero everywhere except at $s = s*$. The last equality follows from the definition of L^1 norm. □

Lemma B.3. *For any $\theta_1, \theta_2 \in \Theta$,*

$$\mathbb{E}_S\left[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1\right] \leqslant 2\sqrt{2}\sqrt{H^2(\theta_1,\theta_2)}.$$

Proof of Lemma B.3. We first use the facts that (1) $L^1 = \frac{1}{2}\text{TV}$ where TV is the total variation distance and (2) TV $\leqslant \sqrt{2}h$ (Harsha, 2011) where h is the Hellinger distance to have that for any $s \in \mathbb{S}$,

$$||\pi_A(\cdot|s;\theta_1) - \pi_A(\cdot|s;\theta_2)||_1 \leqslant 2\sqrt{2}h\big(\pi_A(\cdot|s;\theta_1), \pi_A(\cdot|s;\theta_2)\big). \tag{31}$$

From (31), we have

$$||\pi_A(\cdot|s;\theta_1) - \pi_A(\cdot|s;\theta_2)||_1 \leqslant 2\sqrt{2}h\big(\pi_A(\cdot|s;\theta_1), \pi_A(\cdot|s;\theta_2)\big),$$
$$||\pi_A(\cdot|s;\theta_1) - \pi_A(\cdot|s;\theta_2)||_1^2 \leqslant 8h^2\big(\pi_A(\cdot|s;\theta_1), \pi_A(\cdot|s;\theta_2)\big).$$

Taking expectation with respect to S on both sides, we have

$$\mathbb{E}_S\left[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1^2\right] \leqslant 8\mathbb{E}_S\left[h^2(\pi_A(\cdot|S;\theta_1), \pi_A(\cdot|S;\theta_2))\right],$$
$$\mathbb{E}_S\left[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1^2\right] \leqslant 8H^2(\theta_1,\theta_2).$$

By the Jensen's inequality, we have

$$\left[\mathbb{E}_S||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1\right]^2 \leqslant \mathbb{E}_S\left[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1^2\right].$$

This implies that

$$\mathbb{E}_S\Big[||\pi_A(\cdot|S;\theta_1) - \pi_A(\cdot|S;\theta_2)||_1\Big] \leqslant 2\sqrt{2}\sqrt{H^2(\theta_1,\theta_2)}.$$

□

Lemma B.4 (Properties of H and H^2)**.** *For any $\theta_1, \theta_2, \theta_3 \in \Theta$, the following inequalities hold:*

$$\begin{aligned} H(\theta_1,\theta_2) &\leqslant H(\theta_1,\theta_3) + H(\theta_2,p_3); \\ \big(H(\theta_1,\theta_2)\big)^2 &\leqslant H^2(\theta_1,\theta_2) \leqslant H(\theta_1,\theta_2); \\ H^2(\theta_1,\theta_2) &\leqslant 2H^2(\theta_1,\theta_3) + 2H^2(\theta_2,\theta_3). \end{aligned}$$

Proof of Lemma B.4. For ease of notation, for $i = 1,2,3$, we use p_i to denote π_A parametrized by θ_i, i.e., $p_i(a|s) = \pi_A(a|s;\theta_i)$.

(1) Notice that for any $s \in \mathbb{S}$, $\sqrt{h^2(p_1(\cdot|s), p_2(\cdot|s))}$ is just the regular Hellinger distance that satisfies the triangular inequality. Hence we have

$$\sqrt{h^2(p_1(\cdot|s), p_2(\cdot|s))} \leqslant \sqrt{h^2(p_1(\cdot|s), p_3(\cdot|s))} + \sqrt{h^2(p_2(\cdot|s), p_3(\cdot|s))}. \tag{32}$$

Take expectation of both side of Eq. (32) with respect to S, we have

$$\mathbb{E}_S\Big[\sqrt{h^2(p_1(\cdot|S), p_2(\cdot|S))}\Big] \leqslant \mathbb{E}_S\Big[\sqrt{h^2(p_1(\cdot|S), p_3(\cdot|S))}\Big] + \mathbb{E}_S\Big[\sqrt{h^2(p_2(\cdot|S), p_3(\cdot|S))}\Big].$$

By the definition of H, this means that

$$H(\theta_1,\theta_2) \leqslant H(\theta_1,\theta_3) + H(\theta_2,\theta_3).$$

(2) For the first inequality, by applying the Jensen's inequality, we have

$$\bigg(\mathbb{E}\Big[\sqrt{h^2(p_1(\cdot|S), p_2(\cdot|S))}\Big]\bigg)^2 \leqslant \mathbb{E}\bigg[\Big(\sqrt{h^2(p_1(\cdot|S), p_2(\cdot|S))}\Big)^2\bigg]. \tag{33}$$

Then the inequality follows.

For the second inequality, we have

$$\begin{aligned} H(\theta_1,\theta_2) - H^2(\theta_1,\theta_2) =& \mathbb{E}_S\Big[h(p_1(\cdot|S), p_2(\cdot|S)) - h^2(p_1(\cdot|S), p_2(\cdot|S))\Big] && (34) \\ =& \mathbb{E}_S\Big[\Big(1 - h(p_1(\cdot|S), p_2(\cdot|S))\Big)h(p_1(\cdot|S), p_2(\cdot|S))\Big]. && (35) \end{aligned}$$

Note that for any s, $\Big(1 - h(p_1(\cdot|s), p_2(\cdot|s))\Big)$ is a non-negative function because the Hellinger distance is no larger than 1, and $h(p_1(\cdot|s), p_2(\cdot|s))$ is also a positive function because it is a metric. We have Eq. (35) as the expectation of non-negative functions, and thus we have

$$\mathbb{E}_S\Big[\Big(1 - h(p_1(\cdot|S), p_2(\cdot|S))\Big)h(p_1(\cdot|S), p_2(\cdot|S))\Big] \geqslant 0.$$

Then the inequality follows.

(3) Notice that for any a, b, c, we have $(a-b)^2 \leqslant 2(a-c)^2 + 2(b-c)^2$. With this fact, we have that for any s,

$$\begin{aligned} h^2(p_1(\cdot|s), p_2(\cdot|s)) &= \frac{1}{2}\int \Big(\sqrt{p_1(a|s)} - \sqrt{p_2(a|s)}\Big)^2 da && (36) \\ &\leqslant \frac{1}{2}\int 2\Big(\sqrt{p_1(a|s)} - \sqrt{p_3(a|s)}\Big)^2 + 2\Big(\sqrt{p_2(a|s)} - \sqrt{p_3(a|s)}\Big)^2 da && (37) \\ &= 2\cdot\frac{1}{2}\int \Big(\sqrt{p_1(a|s)} - \sqrt{p_3(a|s)}\Big)^2 da + 2\cdot\frac{1}{2}\int \Big(\sqrt{p_2(a|s)} - \sqrt{p_3(a|s)}\Big)^2 da && (38) \\ &= 2h^2(p_1(\cdot|s), p_3(\cdot|s)) + 2h^2(p_2(\cdot|s), p_3(\cdot|s)). && (39) \end{aligned}$$

This implies that

$$H^2(\theta_1, \theta_2) \leqslant 2H^2(\theta_1, \theta_3) + 2H^2(\theta_2, \theta_3). \tag{40}$$

□

Lemma B.5 (Log-Density Ratio Variance Bound)**.** *Suppose $X \sim p$ is an $\mathbb{R}$-valued random variable with probability density function p, and p_1, p_2 are two other probability density functions for X such that p_1 and p_2 are uniformly bounded from below by C^{-1} on the support of p. Then we have*

$$\mathbb{E}_{X\sim p}\left(\log\frac{p_1(X)}{p_2(X)}\right)^2 \leqslant 2Ch^2(p_1, p_2),$$

where h^2 is the squared Hellinger distance in (10)*.*

Proof of Lemma B.5. By $\log(x) \leqslant 2(\sqrt{x} - 1)$ for any $x \geqslant 0$, we have

$$\begin{aligned}
&\mathbb{E}_{X\sim p}\left(\log\frac{p_1(X)}{p_2(X)}\right)^2 = \int\left(\log\frac{p_1(X)}{p_2(X)}\right)^2 p(x)\mathrm{d}x \\
&\leqslant 4\int \max\left\{\left(\sqrt{\frac{p_1(x)}{p_2(x)}} - 1\right)^2, \left(\sqrt{\frac{p_2(x)}{p_1(x)}} - 1\right)^2\right\} p(x)\mathrm{d}x \\
&= 4\int \max\left\{\frac{1}{p_2(x)}\left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2, \frac{1}{p_1(x)}\left(\sqrt{p_2(x)} - \sqrt{p_1(x)}\right)^2\right\} p(x)\mathrm{d}x \\
&\leqslant 4C\int\left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 \mathrm{d}x \\
&= 2Ch^2(p_1, p_2).
\end{aligned}$$

□

Theorem B.6 (Sen (2018, Theorem 7.13))**.** *Let $\mathcal{F}$ be a measurable function class, such that $\sup_{f\in\mathcal{F}}\|f\|_\infty \leqslant f_{\max}$ for some constant $f_{\max} < +\infty$. Assume that for $A \geqslant ef_{\max}$, $d \geqslant 2$, $0 \leqslant \epsilon \leqslant f_{\max}$, and every finitely supported probability measure $\mathbb{Q}$, we have the covering number (Sen, 2018) as:*

$$\mathcal{N}(\epsilon, \mathcal{F}, L^2(\mathbb{Q})) \lesssim \left(\frac{A}{\epsilon}\right)^d. \tag{41}$$

Let $\sigma_{\mathcal{F}}^2 \doteq \sup_{f\in\mathcal{F}}\mathbb{E}(f - \mathbb{E}f)^2$. Then we have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim \sqrt{\frac{d}{n}\sigma_{\mathcal{F}}^2\log\frac{A}{\sigma_{\mathcal{F}}}} \vee \left\{\frac{d}{n}f_{\max}\log\frac{A}{\sigma_{\mathcal{F}}}\right\}.$$

Proof of Theorem B.6. In this proof, we denote X as the underlying random variable, $\{X_i\}_{i=1}^n$ are n i.i.d. copies of X, and for any $f \in \mathcal{F}$, $\mathbb{P}_n(f) \doteq \frac{1}{n}\sum_{i=1}^n f(X_i)$, $\mathbb{P}(f) \doteq \mathbb{E}[f(X)]$. Without loss of generality, assume that $0 \in \mathcal{F}$, and for any $f \in \mathcal{F}$, $\mathbb{P}(f) = 0$. Let $\{\epsilon_i\}_{i=1}^n$ be i.i.d. Rademacher random variables that are independent of $\{X_i\}_{i=1}^n$. By symmetrization, we have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leqslant 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|. \tag{42}$$

Conditional on $\{X_i\}_{i=1}^n$, by Dudley's entropy bound, we have

$$\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i\frac{f(X_i)}{\sqrt{n}}\right| \leqslant \int_0^{\sigma_{\mathcal{F},n}}\sqrt{\log\mathcal{N}(u, \mathcal{F}, L^2(\mathbb{P}_n))}\mathrm{d}u, \tag{43}$$

where we consider the $L^2(\mathbb{P}_n)$ as the metric on $\mathcal{F}$, that is, for any $f \in \mathcal{F}$, $\|f\|_{L^2(\mathbb{P}_n)}^2 = \frac{1}{n}\sum_{i=1}^n f(X_i)^2$. We also denote $\sigma_{\mathcal{F},n}^2 \doteq \sup_{f\in\mathcal{F}} \|f\|_{L^2(\mathbb{P}_n)}^2$, and $\mathbb{E}_\epsilon$ to emphasize that the expectation is taken with respect to the Rademacher random variables $\{\epsilon_i\}_{i=1}^n$ but holding $\{X_i\}_{i=1}^n$ as fixed. By (41) with $\mathbb{Q}$ chosen as $\mathbb{P}_n$, we have

$$\begin{aligned}(43) &\lesssim \sqrt{d}\int_0^{\sigma_{\mathcal{F},n}} \sqrt{\log\frac{A}{\delta}}\mathrm{d}\delta \\ &\leqslant 2\sqrt{d}\sigma_{\mathcal{F},n}\sqrt{\log\frac{A}{\sigma_{\mathcal{F},n}}} \qquad \text{(by Lemma B.7)}.\end{aligned} \tag{44}$$

In particular, $\log(A/\sigma_{\mathcal{F},n}) \geqslant \log(A/f_{\max}) \geqslant 1$ by assumption, which satisfies the condition for Lemma B.7. Combining (42), (43) and (44), we have

$$\begin{aligned}\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} &\lesssim \sqrt{\frac{d}{n}} \times \mathbb{E}\sqrt{\sigma_{\mathcal{F},n}^2 \log\frac{A}{\sigma_{\mathcal{F},n}}} \\ &\leqslant \sqrt{\frac{d}{n}} \times \sqrt{\frac{1}{2}\mathbb{E}(\sigma_{\mathcal{F},n}^2)\log\frac{A^2}{\mathbb{E}(\sigma_{\mathcal{F},n}^2)}} \qquad \left(\text{by the concavity of } u \mapsto \sqrt{u\log\frac{A^2}{u}}\right),\end{aligned} \tag{45}$$

where $\mathbb{E}$ takes expectation with respect to $\{X_i\}_{i=1}^n$. Notice that

$$\mathbb{E}(\sigma_{\mathcal{F},n}^2) = \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{P}_n(f^2) \leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left\{|(\mathbb{P}_n - \mathbb{P})(f^2)| + \mathbb{P}(f^2)\right\} \leqslant \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}^2} + \sigma_{\mathcal{F}}^2,$$

where we define $\mathcal{F}^2 \doteq \{f^2 : f \in \mathcal{F}\}$. We aim to apply (42), (43), (44), (45) to $\mathcal{F}^2$. Notice that $\sup_{f^2\in\mathcal{F}^2}\|f^2\|_\infty \leqslant f_{\max}^2$, $\sigma_{\mathcal{F}^2,n}^2 \doteq \sup_{f^2\in\mathcal{F}^2}\|f^2\|_{L^2(\mathbb{P}_n)}^2 \leqslant f_{\max}^2 \sup_{f\in\mathcal{F}}\|f\|_{L^2(\mathbb{P}_n)}^2 = f_{\max}^2\sigma_{\mathcal{F},n}^2$. By $\|f^2 - g^2\|_{L^2(\mathbb{P}_n)} = \sqrt{\mathbb{P}_n[(f+g)^2(f-g)^2]} \leqslant 2f_{\max}\sqrt{\mathbb{P}_n(f-g)^2} = 2f_{\max}\|f-g\|_{L^2(\mathbb{P}_n)}^2$ for any $f, g \in \mathcal{F}$, we further have

$$\mathcal{N}(2f_{\max}\epsilon, \mathcal{F}^2, L^2(\mathbb{P}_n)) \leqslant \mathcal{N}(\epsilon, \mathcal{F}, L^2(\mathbb{P}_n)) \lesssim \left(\frac{A}{\epsilon}\right)^d.$$

Therefore, applying (42), (43), (44) and (45) to $\mathcal{F}^2$, we have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}^2} \lesssim \sqrt{\frac{d}{n}} \times \sqrt{\frac{1}{2}f_{\max}^2\mathbb{E}(\sigma_{\mathcal{F},n}^2)\log\frac{4A^2}{\mathbb{E}(\sigma_{\mathcal{F},n}^2)}}.$$

Define $B \doteq \sqrt{\frac{1}{2}\mathbb{E}(\sigma_{\mathcal{F},n}^2)\log\frac{A^2}{\mathbb{E}(\sigma_{\mathcal{F},n}^2)}}$. Then we have

$$\mathbb{E}(\sigma_{\mathcal{F},n}^2) - \sigma_{\mathcal{F}}^2 \lesssim \sqrt{\frac{d}{n}}f_{\max}B.$$

By $u \mapsto u\log\frac{A^2}{u}$ is non-decreasing on $u \in (0, A^2/e]$ and non-increasing on $u \in [A^2/e, +\infty)$, we have

$$B^2 = \frac{1}{2}\mathbb{E}(\sigma_{\mathcal{F},n}^2)\log\frac{A^2}{\mathbb{E}(\sigma_{\mathcal{F},n}^2)} \lesssim \frac{1}{2}\left\{\left(\sigma_{\mathcal{F}}^2 + \sqrt{\frac{d}{n}}f_{\max}B\right) \wedge \frac{A^2}{e}\right\}\log\frac{A^2}{\left(\sigma_{\mathcal{F}}^2 + \sqrt{\frac{d}{n}}f_{\max}B\right) \wedge \frac{A^2}{e}}.$$

In particular, $B \leqslant \sqrt{\frac{A^2}{2e}}$, $\sigma_{\mathcal{F}}^2 \leqslant f_{\max}^2 \leqslant A^2/e^2 < A^2/e$. Then the cap A^2/e is inactive as $d/n \to 0$ asymptotically. Therefore,

$$B^2 \lesssim \left(\sigma_{\mathcal{F}}^2 + \sqrt{\frac{d}{n}}f_{\max}B\right)\log\frac{A}{\sigma_{\mathcal{F}}}.$$

In particular, B is bounded by both roots of the corresponding quadratic equation:

$$B \lesssim \frac{1}{2}\left\{\sqrt{\frac{d}{n}}f_{\max}\log\frac{A}{\sigma_{\mathcal{F}}} + \sqrt{\frac{d}{n}f_{\max}^2\left(\log\frac{A}{\sigma_{\mathcal{F}}}\right)^2 + 4\sigma_{\mathcal{F}}^2\log\frac{A}{\sigma_{\mathcal{F}}}}\right\} \lesssim \left\{\sqrt{\frac{d}{n}}f_{\max}\log\frac{A}{\sigma_{\mathcal{F}}}\right\} \vee \sqrt{\sigma_{\mathcal{F}}^2\log\frac{A}{\sigma_{\mathcal{F}}}}.$$

Combined with (45), we further have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim \sqrt{\frac{d}{n}}B \lesssim \sqrt{\frac{d}{n}\sigma_{\mathcal{F}}^2 \log \frac{A}{\sigma_{\mathcal{F}}}} \vee \left\{\frac{d}{n} f_{\max} \log \frac{A}{\sigma_{\mathcal{F}}}\right\}.$$

□

Lemma B.7. *Suppose* $a, A > 0$ *such that* $\log(A/a) \geqslant 1$. *Then we have*

$$\int_0^a \sqrt{\log \frac{A}{u}} \mathrm{d}u \leqslant 2a\sqrt{\log \frac{A}{a}}.$$

Proof of Lemma B.7. Define

$$f(a) \doteq \begin{cases} 2a\sqrt{\log \frac{A}{a}} - \int_0^a \sqrt{\log \frac{A}{u}} \mathrm{d}u, & a > 0; \\ 0, & a = 0. \end{cases}$$

Then f is continuous at 0. Moreover, for $a > 0$, we have

$$f'(a) = \sqrt{\log \frac{A}{a}} - \frac{1}{\sqrt{\log \frac{A}{a}}},$$

which is nonnegative if $\log(A/a) \geqslant 1$. As $a \to 0^+$, we further have

$$\begin{aligned} &\frac{f(a)}{a} = 2\sqrt{\log \frac{A}{a}} - \frac{1}{a}\int_0^a \sqrt{\log \frac{A}{u}} \mathrm{d}u \\ &= \sqrt{\log \frac{A}{a}} - \frac{1}{2a}\int_0^a \frac{1}{\sqrt{\log \frac{A}{u}}} \mathrm{d}u \qquad \text{(by integration-by-part)} \\ &\geqslant \sqrt{\log \frac{A}{a}} - \frac{1}{2\sqrt{\log \frac{A}{a}}}; \\ &\liminf_{a \to 0^+} \frac{f(a)}{a} \geqslant +\infty. \end{aligned}$$

Therefore, for any $a \geqslant 0$, we have $f(a) \geqslant 0$, which concludes the proof. □