Stochastic approximation for uncapacitated assortment optimization under the multinomial logit model

Yannik Peeters¹ | Arnoud V. den Boer¹,²

¹Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands
²Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Yannik Peeters, Amsterdam Business School, University of Amsterdam, Plantage Muidergracht 12, 1018 TV, Amsterdam, The Netherlands. Email: y.a.peeters@outlook.com

History
Accepted by Rene Caldentey, revenue management and marketplace design.

Abstract
We consider dynamic assortment optimization with incomplete information under the uncapacitated multinomial logit choice model. We propose an anytime stochastic approximation policy and prove that the regret—the cumulative expected revenue loss caused by offering suboptimal assortments—after $T$ time periods is bounded by $\sqrt{T}$ times a constant that is independent of the number of products. In addition, we prove a matching lower bound on the regret for any policy that is valid for arbitrary model parameters—slightly generalizing a recent regret lower bound derived for specific revenue parameters. Numerical illustrations suggest that our policy outperforms alternatives by a significant margin when $T$ and the number of products $N$ are not too small.

KEYWORDS
assortment optimization, incomplete information, multinomial logit model, stochastic approximation

1 INTRODUCTION

Assortment optimization is vital for maximizing revenue. A seller of a large number of substitute products faces the challenge of determining the most profitable subset, that is, assortment, of products to offer to consumers. In most practical situations, the seller does not know the specific demand distribution for all assortments, so that the optimal assortment optimization has to be studied in a sequential optimization framework with incomplete information. As the infrastructure of information improves to better handle incoming real-time purchase data, the necessity for computationally efficient and easily implementable data-driven algorithms arises.

Because the number of feasible assortments grows exponentially in the number of products, the problem is often studied under a particular choice model that describes how demand or choice probabilities depend on the offered assortment. Perhaps the most widely used choice model is the multinomial logit (MNL) choice model. When the seller can offer assortments of any size, the optimal assortment under the MNL model is of the form “offer the $k$ most profitable products”—see, for example, Talluri and van Ryzin (2004, proposition 6). This structure greatly simplifies the problem as the seller merely has to learn the optimal value of the one-dimensional quantity $k$ instead of high-dimensional model parameters, suggesting that a conceptually simple stochastic-approximation type of policy might work well.

In this article, we construct such a stochastic-approximation policy for the dynamic assortment optimization problem without capacity constraint and with demand characterized by the MNL model with unknown parameters, and show that it is both asymptotically optimal and has excellent numerical performance.

1.1 Literature review

Within the management science and operations research community, the problem of dynamic assortment planning has recently received much attention (see, e.g., Rusmevichientong et al., 2010; Farias et al., 2013; Sauré & Zeevi, 2013;
Agrawal et al., 2017; Cheung & Simchi-Levi, 2017; Chen & Wang, 2018; Ou et al., 2018; Agrawal et al., 2019; Kallus & Udell, 2020; Chen et al., 2021). All these papers study assortment optimization under the MNL model in a sequential decision framework. An important recent contribution closely related to our work is Chen et al. (2021). The authors construct a Trisection policy that exploits the structure of the optimal assortment under the uncapacitated MNL assortment optimization problem (“offer the k most profitable products”), and prove that its regret—the total expected revenue loss caused by offering suboptimal assortments—after $T$ time periods is bounded by $\sqrt{T}$ times a constant that does not depend on the number of products. This improves earlier regret rates that do depend on the number of products—see, for example, the bounds derived in Rusmevichientong et al. (2010, theorem 3.4), Sauré and Zeevi (2013, theorem 4), Agrawal et al. (2017, theorem 1) and Agrawal et al. (2019, theorems 1, 3, and 4), applied to a maximum capacity of $K = N$ products. In addition, Chen et al. (2021) construct an instance where the theorem 1) and Agrawal et al. (2019, theorems 1, 3, and 4), applied to a maximum capacity of $K = N$ products. In addition, Chen et al. (2021) construct an instance where the worst-case regret over a set of possible model parameters that any policy must endure is bounded from below by a constant from below by a constant $\mathcal{O}(\sqrt{T})$ regret are asymptotically optimal for any product revenue parameters. We conduct numerical experiments that demonstrate that our policy has a robust performance in different instances, and can outperform alternative algorithms by a significant margin when $T$ and the number of products are not too small.

We emphasize that our policy is not the first to have been shown to be asymptotically optimal (that is achieved by Chen et al. (2021)); the value of our policy lies in the fact that it is easy to understand and implement, and has superior numerical performance when $N$ and $T$ are not too small.

The rest of this article is organized as follows. We introduce the model in Section 2. In Section 3, we present our policy and the upper bound on its regret, as well as the lower bound result of the regret of any policy. The numerical experiments are contained in Section 4. Mathematical proofs are collected in the Appendix.

## 2 | MODEL

We consider a seller who has $N \in \mathbb{N}$ different products for sale during $T \in \mathbb{N}$ time periods, and who has to determine at the beginning of each time period which subset of products is available for purchase. We abbreviate the set of products $\{1, \ldots, N\}$ as $[N]$ and the set of time instances $\{1, \ldots, T\}$ as $[T]$. Each product $i \in [N]$ yields a known marginal revenue of $w_i > 0$. We assume that $w_i \leq 1$ for all $i \in [N]$, and that the products are ordered in strictly increasing order with respect to their marginal revenue: $w_1 < \cdots < w_N$. (Note that, in the uncapacitated MNL model, products that yield the same marginal profit to the seller can be considered as the same product.)

Each product $i \in [N]$ is associated with a preference parameter $v_i \geq 0$, unknown to the seller. At the beginning of each time period $t \in [T]$ the seller selects an assortment $S_t \subseteq [N]$ based on the purchase information available up to and including time $t - 1$. Thereafter, the seller observes a purchase $Y_t \in S_t \cup \{0\}$, where product 0 corresponds to a no-purchase. Clearly, $w_0 = 0$ and we set $v_0$—the preference parameter for product 0—equal to 1. The purchase probabilities under the MNL model are given by

$$
\mathbb{P}(Y_t = i | S_t = S) = \frac{v_i}{1 + \sum_{j \in S} v_j}, \quad \text{for all } i \in S \cup \{0\}.
$$

The expected revenue earned by the seller from an assortment $S \subseteq [N]$ is denoted by

$$
r(S, v) := \frac{\sum_{i \in S} v_i w_i}{1 + \sum_{i \in S} v_i}.
$$

The assortment decisions of the seller are described by his/her policy: a collection of probability distributions $\pi = (\pi(\cdot|h) : h \in \mathcal{H})$.
\( h \in H \) on \( 2^{[N]} \), where
\[
H := \bigcup_{n \in [T]} \{(S, Y) : Y \in S \cup \{0\}, S \subseteq [N]\} \}
\]
is the set of possible histories, and where, conditionally on \( h = (S_1, Y_1, \ldots, S_{t-1}, Y_{t-1}) \), assortment \( S_t \) has distribution \( \pi(\cdot| h) \), for all \( h \in H \) and all \( t \in [T] \). Let \( \mathbb{P}_\pi^T \) denote the probability measure of \( (S_t, Y_t : t \in \mathbb{N}) \) under policy \( \pi \) and preference vector \( v \), and let \( \mathbb{E}_\pi^T \) be the corresponding expectation operator. The objective for the seller is to find a policy \( \pi \) that maximizes the total accumulated revenue or, equivalently, minimizes the accumulated regret:
\[
\Delta_\pi(T, v) := T \cdot \max_{\pi \in \mathcal{P}} r(S, v) - \sum_{t=1}^T \mathbb{E}_\pi^T [r(S_t, v)].
\]
In addition, we consider the worst-case regret over a class \( \mathcal{V} \) of preference vectors:
\[
\Delta_\pi(T) := \sup_{v \in \mathcal{V}} \Delta_\pi(T, v).
\]
The class of preference vectors \( \mathcal{V} \) under consideration consists of all vectors \( v \) that satisfy the following assumption:

**Assumption 1** \( \mathcal{V} \) is the set of all components-wise non-negative vectors \( (v_1, \ldots, v_N) \) such that
\[
\frac{1}{1 + \sum_{i=1}^N v_i} \geq p_0,
\]
for some \( p_0 \in (0, 1) \) known to the seller.

This assumption is arguably mild as it simply puts a lower bound \( p_0 \) on the no-purchase probability that should be strictly positive but can otherwise be arbitrarily small. Note that Assumption 1 does not incorporate a “separability” condition (as in, e.g., Rusmevichientong et al., 2010, assumption 2.1) between the true optimal assortment and sub-optimal alternatives. The use of Assumption 1 is to ensure that the policy parameters guarantee convergence. For the choice of \( p_0 \), see further Remarks 1 and 2.

3 | ASSORTMENT OPTIMIZATION

In this section, we discuss the structure of the optimal assortment, propose a policy for incomplete information, and show that its regret is bounded by \( \sqrt{T} \) times a constant independent of \( N \). Then we show that, for arbitrary revenue parameters, the regret of any policy is bounded from below by a constant times \( \sqrt{T} \), implying that our policy is asymptotically optimal in a general setting. The mathematical proofs are contained in the Appendix.

For uncapacitated assortment optimization under the MNL model, it is known that the optimal assortment is of the form ‘offer the \( k \) most profitable products’ for some integer \( k \in [N] \), see, for example, Talluri and van Ryzin (2004, proposition 6). To facilitate our analysis, we define
\[
S_\rho := \{ i \in [N] : w_i \geq \rho \}, \quad \rho \in [0, 1].
\]
Then it is also known that the sequence \( \mathcal{P} := (r(S_1, v), \ldots, r(S_N, v)) \) is unimodal. Consider the function \( \rho \mapsto r(S_\rho, v) \). This function is piecewise constant and attains every value in \( \mathcal{P} \). Moreover, the assortment \( S^* = S_\rho^* \) with
\[
\rho^* := \max_{\mathcal{V} \subseteq [N]} r(S,v)
\]
is optimal under preference vector \( v \), see, for example, Chen et al. (2021, section 4). This structure has compelling computational implications, as we only need to approximate \( \rho^* \) instead of the more straightforward approach of establishing estimates of \( r(S_i, v) \) for all \( i \in [N] \).

3.1 | A policy for incomplete information

In this section, we propose a policy to iteratively establish a sequence of assortments that converges to the optimal assortment. The policy follows the intuitive approach of offering products with marginal revenue above a threshold value \( \phi_t \in [0, 1] \) at each time \( t \in [T] \). Based on the observed (no-)purchases, each next threshold value \( \phi_{t+1} \) is selected. The policy is parameterized by \( \alpha \geq 1 \) and \( \beta \geq \alpha - 1 \).

**Stochastic approximation policy SAP(\( \alpha, \beta \))**

1. Initialization. Let \( \phi_1 \in [0, 1] \). For all \( t \in \mathbb{N} \) let \( \phi_t := \phi/(t + \beta) \), for some \( \alpha \geq 1 \) and \( \beta \geq \alpha - 1 \). Put \( t := 1 \). Go to 2.
2. Assortment selection. Let \( S_t := S_{\phi_t} \) and
\[
\phi_{t+1} := \phi_t + \alpha_t \left( w_Y - \phi_t \right)
\]
Put \( t := t + 1 \). If \( t \leq T \), then go to 2, else to 3.

3. Terminate.

The policy \( \text{SAP}(\alpha, \beta) \) is a classical stochastic approximation algorithm (Robbins & Monro 1951; Kushner & Yin 1997) that relies on the observation that \( \rho^* \) is the unique solution to the fixed point equation
\[
\rho = r(S_\rho, v).
\]
This is easily verified (see, e.g., Lemma 2 from Chen et al. (2021)). Note that we do not directly observe the value of \( r(S_\rho, v) \). However, we do have the unbiased, noisy observation \( w_Y \) given offered assortment \( S_\rho \) at our disposal. As a result, the sign of \( w_Y - \phi \) approximately indicates the direction in which \( \rho^* \) is situated in relation to \( \phi \). The step sizes \( \alpha_t \) decays approximately as \( 1/t \), ensuring the correct convergence rate of \( \phi_t \) to \( \rho^* \), as \( \phi_t \) does not “keep jumping over” \( \rho^* \), and \( \phi_t \) does not converge “too slow.”
3.2 Regret upper bound

We proceed by showing that the worst-case regret of SAP $(\alpha, \beta)$ is bounded from above by $\sqrt{T}$ times a constant independent of $N$.

**Theorem 1** Let $\pi$ correspond to SAP $(\alpha, \beta)$ with $\alpha \geq 1/p_0$ and $\beta \geq \alpha - 1$. Then there exists a $\bar{C} > 0$ such that, for all $T \in \mathbb{N}$,

$$\Delta_\pi(T) \leq \bar{C}\sqrt{T}.$$  

The proof relies on two main steps. First, we show that the regret is bounded from above by the following expression:

$$\Delta_\pi(T) \leq c_1 \left( T E^\pi_\pi [\|\phi^* - \phi_{T+1}\|] + \sum_{t=1}^T E^\pi_\pi [\|\phi^* - \phi_t\|] \right),$$  

for all $\pi \in \mathcal{V}$ and some instance-independent constant $c_1$. Second, we show a recursive relationship regarding the mean squared error of $\phi_t$ with respect to $\phi^*$. Then the convergence rate of $\phi_t$ to $\phi^*$ is a consequence of the recursive relation. This second step is summarized in the following lemma.

**Lemma 1** Let $\pi \in \mathcal{V}$ and let $\pi$ correspond to SAP $(\alpha, \beta)$ with $\alpha \geq 1/p_0$ and $\beta \geq \alpha - 1$. For all $t \in \{1, \ldots, T\}$ it holds that

$$E^\pi_\pi [(\phi^* - \phi_{t+1})^2] \leq E^\pi_\pi [(\phi^* - \phi_t)^2] (1 - 2p_0 a_i) + a_i^2.$$  

As a result, there exists an instance-independent constant $c_2$ such that, for all $t = 1, \ldots, T + 1$,

$$E^\pi_\pi [(\phi^* - \phi_t)^2] \leq \frac{c_2}{t + \beta}.$$  

We then apply Jensen’s inequality to the concave function $x \mapsto \sqrt{x}$ to conclude that the mean absolute error of $\phi_t$ with respect to $\phi^*$ converges as $1/\sqrt{t}$, that is,

$$E^\pi_\pi [\|\phi^* - \phi_t\|] \leq \frac{\sqrt{c_2}}{\sqrt{t + \beta}},$$

for all $t = 1, \ldots, T + 1$. With the convergence rate of the mean absolute error in combination with (1), the proof of Theorem 1 follows.

**Remark 1** The requirement of $\alpha \geq 1/p_0$ in Theorem 1 is caused by the constant $p_0$ in (2), and ensures an $O(\sqrt{T})$ regret for all preference vectors $\pi \in \mathcal{V}$, even with small corresponding no-purchase probability. However, as the performance of SAP depends on its parameters, a large value of $\alpha$ and $\beta$ when $p_0$ is small might not be necessary in practice. The inequality (2) is also valid when $p_0$ is replaced by the instance-dependent constant $\gamma \in (p_0, 1]$.

where

$$\gamma := \inf_{\phi^* \in \mathcal{X}} \frac{r(S_\pi, \pi) - \phi}{\phi^* - \phi}.$$  

and where $\phi^*$ denotes the optimal revenue as function of the preference vector $\pi$. We still obtain an $O(\sqrt{T})$ regret bound as long as $\alpha \geq 1/\gamma$, which is a weaker requirement on $\alpha$ since $\gamma$ might be substantially larger than $p_0$.

To illustrate this possibility, consider the case where $N = 2$, $r_1 = 0.4$, and $r_2 = 1.0$. In addition, fix $p_0 \in (0, 1/2)$ and let $\mathcal{V}$ be as in Assumption 1. For all $\pi = (v_1, 1)$ such that $1/(1 + v_1 + 1) \geq p_0$ the optimal assortment is $[2]$ with $\phi^*_1 = 0.5$. In this example, it holds that the infimum in (4) is attained at $\phi = r_1 = 0.4$ and as a result

$$\gamma = r((1, 2), v) - 0.4 \approx \frac{2}{2 + v_1}.$$  

If $v_1 = 0$, then $\gamma_1 = 1 > p_0$, and if $v_1 = 2/p_0 - 2$, then $\gamma_1 = 2p_0$. This example shows that there exists instances within $\mathcal{V}$ where $\gamma$ is of the order $p_0$ as well as instances where $\gamma_1 > p_0$. Hence, the requirement that $\alpha$ is of the order $1/p_0$, although necessary for bounding the worst-case regret as done by Theorem 1, can for some preference vectors be mitigated while maintaining good case-specific performance in practice. For a numerical illustration we refer to Section 4.

**Remark 2** As long as the assumed value of $p_0$ is lower than $\gamma$ defined in (4), with $\nu$ the true preference vector, the policy performs as intended and our convergence results are valid. The speed at which $\phi_t$ converges to $\phi^*$ is negatively affected, however, if $p_0$ is substantially smaller than $\gamma$. On the other hand, if the assumed value of $p_0$ is higher than $\gamma$, then there is no guarantee of convergence. More specifically, the policy could offer sub-optimal assortments “too often,” as $\phi_t$ could get “stuck” when $\phi_t \ll \phi^*$. This corresponds to the parameter $\alpha$ being set too low. In a sense, the choice of $p_0$ captures the trade-off between generality and performance: decreasing $p_0$ enlarges the set of preference vectors for which convergence is guaranteed, but simultaneously decreases convergence speed (although not the rate as function of time). In practice, $p_0$ can be selected in a data-driven fashion, before the start of the algorithm, by offering the full assortment a number of times $n$, constructing a confidence interval $(\hat{\beta} - b_n, \hat{\beta} + b_n)$ around the observed fraction of no-purchases $\hat{\beta}$ (for example with $b_n = \sqrt{2\log(1/\delta)/n}$ for some $\delta \in (0, 1)$), and setting $p_0$ equal to the lower end of the confidence interval (unless $\hat{\beta} - b_n \leq 0$, in which case the experiment needs to be prolonged). Then, $p_0$
will with high probability be a lower bound to the true but unknown no-purchase probability.

3.3 Regret lower bound

Now that we have provided an upper bound on the regret of our SAP policy, we proceed by showing that this bound is asymptotically tight—up to a multiplicative constant—as $T$ grows large. This implies that our policy performs asymptotically optimal. We prove our regret lower bound for values $p_0$ in Assumption 1 such that

$$p_0 \leq \max_{1 \leq k < e \leq N} \frac{2w_e - 2w_k}{5w_e + w_k}.$$ (5)

Observe that this condition can always be ensured to hold by choosing $p_0$ sufficiently small.

The regret lower bound is presented below.

**Theorem 2** There exists a $C > 0$ such that, for all policies $\pi$ and all $T \in \mathbb{N}$,

$$\Delta_\pi(T) \geq C \sqrt{T}.$$ (6)

The proof of Theorem 2 is established in three steps. First, we construct two preference vectors $v^0$ and $v^1$ which are statistically “difficult to distinguish.” Second, we show that any estimator $\psi$ that has the observed purchases $Y_1, \ldots, Y_T$ as inputs and outputs either 0 or 1 must satisfy

$$\max_{j=0,1} \mathbb{P}_v[\psi \neq f] \geq \frac{1}{4}.$$ (7)

Third, we define a specific estimator $\psi$ and show that under the assumption that $\Delta_\pi(T) < C \sqrt{T}$ it follows that

$$\mathbb{P}_v[\psi \neq f] < \frac{1}{4},$$

for both $j = 0,1$. Having found a contradiction with (6), we thus conclude that the statement in Theorem 2 must hold.

The novelty of this proof is concentrated in the first step, which allows us to prove Theorem 2 for arbitrary revenue parameters. The second step and third step are conceptually the same as in Chen et al. (2021). We include the last two steps because of the slight deviation in our set-up as well as for the sake of completeness.

Starting with the first step, we define two quantities. Let $k, \ell \in [N]$ such that $k < \ell$ and

$$p_0 \leq \frac{2w_\ell - 2w_k}{5w_\ell + w_k}.$$ (8)

Note that such $k, \ell$ exist by Equation (5). Furthermore, define

$$u_k := \frac{w_\ell}{w_\ell - w_k} \quad \text{and} \quad u_\ell := \frac{w_k}{w_\ell - w_k}.$$ (9)

The idea underlying this set-up is as follows. Let $u$ denote the preference vector that has $u_k$ as its $k$th component, $u_\ell$ as its $\ell$th component and is zero everywhere else. Then it holds that

$$r([k \ell], u) = r([ \ell ], u) = \max_{S \subseteq N} r(S, u).$$ (10)

Then by perturbing $u$ in two ways we can construct two preference vectors $v^0$ and $v^1$ which are very close to $u$, but result in different solutions to the assortment optimization problem. Specifically, we define $v^0$ and $v^1$ as follows. Let $\varepsilon \in (0, 1/2]$ denote a small constant and let $v^0 = (v^0_1, \ldots, v^0_N)$ and $v^1 = (v^1_1, \ldots, v^1_N)$ denote the vectors of preference parameters such that, for $i \in [N]$,

$$v^0_i = \begin{cases} u_k, & \text{if } i = k, \\ (1 - \varepsilon)u_\ell, & \text{if } i = \ell, \\ 0, & \text{otherwise}, \end{cases}$$ (11)

$$v^1_i = \begin{cases} u_k, & \text{if } i = k, \\ (1 + \varepsilon)u_\ell, & \text{if } i = \ell, \\ 0, & \text{otherwise}. \end{cases}$$ (12)

From this set-up it follows that $\{k, \ell\}$ is the optimal assortment under $v^0$ and $\{\ell\}$ is the optimal assortment under $v^1$ (up to inclusion of products with zero preference). That is,

$$\max_{S \subseteq [N]} r(S, v^0) < w_k \quad \text{and} \quad \{k, \ell\} \in \arg \max_{S \subseteq [N]} r(S, v^0),$$

and

$$\max_{S \subseteq [N]} r(S, v^1) > w_k \quad \text{and} \quad \{\ell\} \in \arg \max_{S \subseteq [N]} r(S, v^1).$$

In what follows we abbreviate the expectation value and probability $\mathbb{E}_v[\cdot]$ and $\mathbb{P}_v[\cdot]$ as $\mathbb{E}_\psi[\cdot]$ and $\mathbb{P}_\psi[\cdot]$ for $j = 0,1$. We suppress the notation that these two notions depend on policy $\pi$, as Theorem 2 holds for any policy. Now, for the second step we use the following lemma, where we first bound the Kullback–Leibler (KL) divergence.

**Lemma 2** Let $S \subseteq [N]$. Then there is constant $c_3 > 0$ such that

$$\text{KL}(\mathbb{P}_0(\cdot|S)||\mathbb{P}_1(\cdot|S)) \leq c_3 \varepsilon^2.$$ (13)

In addition, let $\psi \in \{0,1\}$ denote an arbitrary estimator which has random purchases $Y_1, \ldots, Y_T$ as inputs. Then

$$\max_{j=0,1} \mathbb{P}_j[\psi \neq f] \geq \frac{1}{2} \left( 1 - \sqrt{2c_3 \varepsilon \sqrt{T}} \right).$$ (14)

The lemma above contains a similar statement as lemmas EC.1 and 12 from Chen et al. (2021). However, we look at different preference vectors than considered by Chen et al. (2021) because we allow arbitrary revenue parameters. The proof of the lemma above makes use of Pinsker’s inequality as well as Le Cam’s method. From Lemma 2 and by setting $\varepsilon$ equal to

$$\varepsilon := \min \left\{ \frac{1}{2}, \left( 2\sqrt{c_3 T} \right)^{-1} \right\},$$

it follows that (6) holds.

As first part of the third step, we provide a lower bound of the regret under $v^0$ and $v^1$ in terms of $T, \varepsilon, \phi_0$, and $\phi_2$, where
\( \varphi_0 \) denotes how often \( k \) and \( \ell \) are both contained in \( S_t \) for \( t \in [T] \) and \( \varphi_1 \) denotes how often \( k \) is excluded while \( \ell \) is contained in \( S_t \) for \( t \in [T] \). That is,

\[
\varphi_0 := \sum_{t=1}^{T} I \{ k, \ell \in S_t \} \quad \text{and} \quad \varphi_1 := \sum_{t=1}^{T} I \{ k \notin S_t, \ell \in S_t \}.
\]

This lower bound is stated in the lemma below.

**Lemma 3** There is a constant \( c_4 > 0 \) such that, for any policy \( \pi \),

\[
\Delta_x(T, v^0) \geq c_4 \mathbb{E}_0 \left[ \varepsilon \varphi_1 + T - \varphi_0 - \varphi_1 \right], \quad (8)
\]

and

\[
\Delta_x(T, v^1) \geq c_4 \mathbb{E}_1 \left[ \varepsilon \varphi_0 + T - \varphi_0 - \varphi_1 \right]. \quad (9)
\]

The lemma above covers similar statements shown by Chen et al. (2021) in the proof of lemma 13. Here as well, however, we look at different preference vectors than considered by Chen et al. (2021), because we allow arbitrary revenue parameters.

The remainder of the third step is established by a contradiction. To this end, assume that

\[
\Delta_x(T) < C\sqrt{T}, \quad (10)
\]

where

\[
C := \frac{c_4}{16\sqrt{2}c_3}.
\]

In addition, in view of (8) and (9) we define \( L_0 \) and \( L_1 \) as

\[
L_0 := c_4 (\varepsilon \varphi_1 + T - \varphi_0 - \varphi_1) \quad \text{and} \quad L_1 := c_4 (\varepsilon \varphi_0 + T - \varphi_0 - \varphi_1).
\]

As a consequence of the assumption in (10), we conclude by Markov’s inequality and Lemma 3 that

\[
\mathbb{P}_0 \left( L_0 > 4C\sqrt{T} \right) \leq \frac{\mathbb{E}_0 L_0}{4C\sqrt{T}} \leq \frac{\Delta_x(T, v^0)}{4C\sqrt{T}} < \frac{1}{4},
\]

and likewise

\[
\mathbb{P}_1 \left( L_1 > 4C\sqrt{T} \right) \leq \frac{\mathbb{E}_1 L_1}{4C\sqrt{T}} \leq \frac{\Delta_x(T, v^1)}{4C\sqrt{T}} < \frac{1}{4}.
\]

Next, we define the estimator \( \psi \) as

\[
\psi := \begin{cases} 
0, & \text{if } \varphi_0 > T/2, \\
1, & \text{if } \varphi_0 \leq T/2.
\end{cases}
\]

The remainder of the proof is to show that \( \psi = 1 \) implies \( L_0 > 4C\sqrt{T} \) and that \( \psi = 0 \) implies \( L_1 > 4C\sqrt{T} \). From this we conclude that

\[
\mathbb{P}_0(\psi = 1) < \frac{1}{4} \quad \text{as well as} \quad \mathbb{P}_1(\psi = 0) < \frac{1}{4}.
\]

This is a contradiction with the statement in (6), which holds by Lemma 2. Therefore, statement (10) cannot be true and we have thus proven Theorem 2.

## 4 | NUMERICAL EXPERIMENTS

In this section, we compare the performance of our proposed policy with alternatives for dynamic assortment optimization under the MNL model. In particular, we compare our policy SAP with the Thompson sampling (TS) based policy from Agrawal et al. (2017) and the Trisection (TR) policy from Chen et al. (2021) using simulated data. It is worth observing that only for SAP and TR an \( \mathcal{O}(\sqrt{T}) \) regret bound has been proven. We simulate purchase data in two scenarios.

### Scenario 1

For different values of \( N \) and \( T \) we draw \( N \) values uniformly at random from \([0.4,0.5]\), order the randomly drawn values in increasing order, and set the ordered values as the revenue parameters \( w_1, \ldots, w_N \). We draw the preference parameters \( v_1, \ldots, v_N \) uniformly at random from \([10/N, 20/N]\).

### Scenario 2

As in Scenario 1, but now we draw the unsorted revenue parameters uniformly at random from \([0,1]\), order the randomly drawn values in increasing order, and set the ordered values as the revenue parameters \( w_1, \ldots, w_N \); in addition, we draw the preference parameters \( v_1, \ldots, v_N \) uniformly at random from \([1/N, 100/N]\).

The experimental setup described by Scenario 1 follows the set-up from Chen et al. (2021). This setup ensures that finding the optimal solution is non-trivial as the optimal assortment is equal to \( \{i : w_i \geq x\} \) for some \( x \) that lies approximately within \([0.4, 0.5]\). We include the simulated results from Scenario 2 to investigate how the policies perform for a broader range of parameters: the range of possible no-purchase probabilities when offering the entire set of products is \([1/101, 1/2]\) for Scenario 2, whereas for Scenario 1 the range is \([1/21, 1/11]\).

In both simulation scenarios we set \( N \) to values ranging from 25 to 5000, roughly doubling each step, and we simulate the regret of the policies a total of 1000 times. For each \( N \), we draw the preference and revenue parameters and, for the Trisection policy, let the policy run for \( T = 500, 1000, 2500, 5000 \). We let our policy SAP and the Thompson Sampling policy run for \( T = 5000 \) and record the intermediate regret at \( t = 500, 1000, 2500 \).

For the Trisection policy, we use the version with adaptive confidence levels, see, for example, Chen et al. (2021, section 6). In addition, in line with the simulations of Chen
et al. (2021), we replace their $p_{21}$ and $p_{20}$ with $\frac{1}{p_{0}}$ and $\frac{1}{p_{0}}$ for $N_{1} = 1/21$ and $N_{0} = 1/101$ for Scenarios 1 and 2, respectively. For our policy SAP, we notice that the preference parameters satisfy Assumption 1 for $p_{0} = 1/21$ and $p_{0} = 1/101$ for Scenarios 1 and 2, respectively. Theorem 1 indicates that we should set the parameter of SAP as $\alpha = 21$ and $\beta = 20$, and $\alpha = 101$ and $\beta = 100$. Following Remark 1, we can obtain a better performance with smaller parameters. Hence, we include the performance of SAP with $\alpha = 1$ and $\beta = 0$. This policy is referred to as Adjusted SAP (ASAP). In addition, in all instances we start SAP with $\rho_{1} = w_{1}$.

### 4.1 | Results

Table 1 reports the average regret of the policies in Scenarios 1 and 2 over 1000 simulation runs. All the standard errors are within 3%. From Table 1 we see that in both scenarios, the regret of SAP and ASAP does not appear to grow large when $N$ increases, which is in line with our theoretical findings. In addition, we find that, in Scenario 1, SAP performs on par with TR for $T = 500$, 1000, and substantially outperforms TR for larger values of $T$. In Scenario 2, SAP outperforms TR significantly for all values of $N$ and $T$. We also find that the adjusted policy ASAP performs improves performance even further by several magnitudes. When we compare SAP to TS, we see that TS performs well for small values of $N$ whereas SAP is better for large values of $N$. The fact that the regret of TS appears to be growing substantially in $N$ suggests that TS is not an asymptotically optimal policy (i.e., with regret bounded by a factor that is independent of $N$).

Overall, we find that SAP and ASAP perform well and robustly for increasing values for $N$. In addition, we see that
adequate parameter adjustment for SAP can yield a significant benefit over established alternatives, especially for large values of \( N \). This suggests that designing an algorithm where \( \alpha, \beta \) are adaptively tuned may be an interesting direction for future research.

For illustrative purposes we include visual graphs in Figure 1 for the average regret for the policies for \( N \) fixed at 5000.

**CONFLICTS OF INTEREST**

The authors have no conflicts of interest to disclose.

**DATA AVAILABILITY STATEMENT**

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

**ORCID**

Yannik Peeters https://orcid.org/0000-0003-0335-8836

**REFERENCES**

Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2017). Thompson sampling for the MNL-bandit. Proceedings of the Conference on Learning Theory (COLT) (p. 76–78). Amsterdam, The Netherlands.

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

Chen, X., & Wang, Y. (2018). A note on a tight lower bound for MNL-bandit assortment selection models. *Operations Research Letters*, 46, 534–537.

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

Cheung, W. C., & Simchi-Levi, D. (2017). Assortment optimization under unknown multinomial logit choice models. *ArXiv e-print*. https://arxiv.org/abs/1704.00108

Farias, V. F., Jagabathula, S., & Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science*, 59, 305–322.

Kallus, N., & Udell, M. (2020). Dynamic assortment personalization in high dimensions. *Operations Research*, 68, 1020–1037.

Kushner, H. J., & Yin, G. G. (1997). Stochastic approximation and recursive algorithms and applications. Springer-Verlag.

Ou, M., Li, N., Zhu, S., & Jin, R. (2018). Multinomial logit bandit with linear utility functions. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 2602–2608) Stockholm, Sweden.

Peeters, Y., den Boer, A. V., & Mandjes, M. (2021). Continuous assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58, 1666–1680.

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

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

---

**How to cite this article:** Peeters, Y., & den Boer, A. V. (2022). Stochastic approximation for uncapacitated assortment optimization under the multinomial logit model. *Naval Research Logistics (NRL)*, 69(7), 927–938. https://doi.org/10.1002/nav.22068

**APPENDIX A: PROOFS OF THE RESULTS IN SECTION 3.2**

Throughout this section we abbreviate the expectation \( \mathbb{E}_v^\pi[\cdot] \) depending on \( v \in \mathcal{V} \) and \( \pi = \text{SAP}(\alpha, \beta) \) as \( \mathbb{E}[\cdot] \). In addition, recall that \( S_\phi = \{ i \in [N] : w_i \geq \phi \} \) with \( \phi \in [0, 1] \).

**Proof of Lemma 1** First, observe that

\[
\begin{align*}
\mathbb{E} \left[ (\phi^* - \phi_{i+1})^2 | \phi_i \right] &= \mathbb{E} \left[ (\phi^* - \phi_i - a_i (w_{Y_i} - \phi_i))^2 | \phi_i \right] \\
&= \mathbb{E} \left[ (\phi^* - \phi_i)^2 - 2a_i (\phi^* - \phi_i) (w_{Y_i} - \phi_i) + a_i^2 (w_{Y_i} - \phi_i)^2 | \phi_i \right] \\
&\leq (\phi^* - \phi_i)^2 - 2a_i (\phi^* - \phi_i) (h(\phi_i) - \phi_i) + a_i^2. \\
\end{align*}
\]

(A1)

For bounding the cross term in (A1), first suppose that \( \phi_i < \phi^* \). Then, we define \( I(\phi) \) as

\[
I(\phi) = \sum_{i \in S_\phi} v_i (w_i - \phi), \quad \phi \in [0, 1].
\]

Note that \( \phi^* = I(\phi^*) \) since \( \phi^* = r (S_{\phi^*}, v) \) and note that \( S_{\phi^*} \subseteq S_\phi \). Therefore,

\[
I(\phi) = \sum_{S_{\phi^*}} v_i (w_i - \phi) + \sum_{S_{\phi^*} \setminus S_{\phi^*}} v_i (w_i - \phi)
\]
implies

By induction, we show that the inequality (2)

Taking the expectation on both sides yields (2).

Next, assume that

since

\[ \sum_{i \in S_y} v_i \leq (\delta^* - \delta) \cdot \frac{1}{1 + \sum_{i \in S_y} v_i} \]

\[ \geq p_0 (\delta^* - \delta) \cdot \frac{1}{1 + \sum_{i \in S_y} v_i} \]

Therefore,

\[ h(\delta) - \delta = \frac{\sum_{i \in S_y} v_i w_i}{1 + \sum_{i \in S_y} v_i} = \frac{I(\delta) - \delta}{1 + \sum_{i \in S_y} v_i} \]

\[ \geq \frac{1}{1 + \sum_{i \in S_y} v_i} (\delta^* - \delta) \]

\[ \geq \frac{1}{1 + \sum_{i \in S_y} v_i} (\delta^* - \delta) \]

\[ \geq p_0 (\delta^* - \delta) . \]

Next, consider the case that \( \delta_i = \delta^* . \) Then, it holds that\[ h(\delta) \leq h(\delta^*) = \delta^* \text{ and since } p_0 \leq 1 \]

\[ h(\delta) - \delta = \delta^* - \delta \leq p_0 (\delta^* - \delta) . \]

Either way, the cross term in (A1) is bounded from above as

\[ -2a_t (\delta^* - \delta) (h(\delta) - \delta) \leq -2p_0 a_t (\delta^* - \delta)^2 . \]

As a result, we conclude that

\[ \mathbb{E} \left[ (\delta^* - \delta)^2 \mid \delta \right] \leq (\delta^* - \delta)^2 (1 - 2p_0 a_t + a_t^2) . \]

Taking the expectation on both sides yields (2).

For \( t = 1, \ldots, T + 1, \) denote the expected squared difference of \( \delta^* \) and \( \delta, \) as

\[ \delta_i = \mathbb{E} \left[ (\delta^* - \delta_i)^2 \right] . \]

By induction, we show that the inequality (2) implies \[ \delta_i \leq c_2/(t + \beta), \] for all \( t = 1, \ldots, T + 1 \) and some \( c_2 > 0. \) To this end, let \( c_2 \) be the constant

\[ c_2 = \max \left\{ 1 + \beta, 2a^2 \right\} . \]

For \( t = 1, \) we observe that

\[ \delta_1 \leq 1 \leq \frac{c_2}{1 + \beta} . \]

Next, assume that \( \delta_i \leq c_2/(t + \beta) \) for \( t \leq t_0 \) for some \( t_0. \) Then, for \( t > t_0, \) it follows that

\[ \frac{t + \beta}{t + \beta + 1} - 2ap_0 < 1 - 2ap_0 \leq -ap_0 < 1, \]

since \( a \geq 1/p_0 \) and so, by our choice of \( c_2, \)

\[ c_2 \left( \frac{t + \beta}{t + \beta + 1} - 2ap_0 \right) + a^2 < -c_2 + a^2 \leq 0. \]

It follows that

\[ c_2 \left( (t + \beta) - 2ap_0 - \frac{(t + \beta)^2}{t + \beta + 1} \right) + a^2 \leq 0, \]

and therefore

\[ \frac{c_2}{t + \beta} \left( 1 - 2 \frac{ap_0}{t + \beta} \right) + \frac{a^2}{(t + \beta)^2} \leq \frac{c_2}{t + \beta + 1} . \]

This, by (2) as well as by the induction hypothesis, yields

\[ \delta_{t+1} \leq \frac{c_2}{t + \beta + 1} , \]

when \( p_0 a_t < 1/2. \) If \( p_0 a_t(1/2, 1), \) it follows from (2) that

\[ \delta_{t+1} \leq \frac{a_t^2}{t + \beta + 1} , \]

where the last inequality holds because \( 2(t + \beta)^2 \geq t + \beta + 1 as t + \beta \geq 1 and c_2 \geq 2a^2, \) so that we have shown the lemma. \( \square \)

**Proof of Theorem 1** First, define \( h(\delta) := r(S, \nu) \) and

\[ \delta^* = \max_{\delta \in [0, 1]} r(S, \nu) . \]

Recall that \( \delta^* = h(\delta^*) ; \) see, for example, Chen et al. (2021, section 4). Next, note that

\[ T = \sum_{t=1}^{T} \frac{1}{a_t} (1 - (1 - a_t)) . \]  
\[ \text{(A2)} \]

Also, note that, for \( t \in [T], w_Y \), can be written in terms of \( a_t, \delta_t, \) and \( \delta_{t+1} \) as

\[ w_{Y_t} = \frac{1}{a_t} (\delta_{t+1} - (1 - a_t) \delta_t) \]  
\[ \text{(A3)} \]

As a consequence of (A2) and (A3), it holds that

\[ T \delta^* - \sum_{t=1}^{T} w_{Y_t} = \sum_{t=1}^{T} \frac{1}{a_t} \left( (\delta^* - \delta_{t+1}) - (1 - a_t) (\delta^* - \delta_t) \right) \]

\[ = \sum_{t=1}^{T} \frac{1}{a_t} (\delta^* - \delta_{t+1}) - \sum_{t=1}^{T} \frac{1}{a_t} (\delta^* - \delta_t) \]

\[ = \frac{1}{a_T} (\delta^* - \delta_{T+1}) + \sum_{t=1}^{T-1} \frac{1}{a_t} (\delta^* - \delta_{t+1}) \]

\[ - \sum_{t=2}^{T} \frac{1}{a_t} (\delta^* - \delta_t) \]

\[ = \frac{1}{a_T} (\delta^* - \delta_{T+1}) + \sum_{t=2}^{T} \frac{1}{a_t} (\delta^* - \delta_t) \]

\[ - \sum_{t=2}^{T} \frac{1}{a_t} (\delta^* - \delta_t) \]

\[ = \frac{1}{a_T} (\delta^* - \delta_{T+1}) + \sum_{t=2}^{T} \left( \frac{1}{a_{t-1}} - \frac{1}{a_t} \right) (\delta^* - \delta_t) \]

\[ \times (\delta^* - \delta_t) \leq \frac{1}{a_T} (\delta^* - \delta_{T+1}) \]

Given that \( |\delta^* - \delta_t| \leq 1, \) it follows that

\[ T \delta^* - \sum_{t=1}^{T} w_{Y_t} \leq \frac{T + \beta}{\alpha} (\delta^* - \delta_{T+1}) \]
\[ + \frac{\alpha - 1}{\alpha} \sum_{t=2}^{T} (\rho^* - \rho_t) + \frac{\beta + 1 - \alpha}{\alpha} \]

From Lemma 1 and Jensen’s inequality for the concave function \( x \mapsto \sqrt{x} \) it follows that
\[
\mathbb{E} [p^* - q_t] \geq \mathbb{E} \left[ (p^* - \rho_t)^2 \right]^{1/2} \leq \frac{\sqrt{c_2}}{t + \beta + 1}.
\]

From this, we conclude that the regret is bounded from above as
\[
\Delta_x(T) \leq \frac{T + \beta}{\alpha} \mathbb{E} [p^* - \rho_{T+1}] + \frac{\alpha - 1}{\alpha} \sum_{t=2}^{T} \mathbb{E} [p^* - \rho_t] \leq \frac{\sqrt{c_2}}{\alpha} \sqrt{T + \beta + 1} + \frac{\alpha - 1}{\alpha} \sum_{t=2}^{T} \frac{1}{\sqrt{t + \beta + 1}} + \frac{\beta + 1 - \alpha}{\alpha} \leq \frac{\sqrt{c_2}}{\alpha} \sqrt{T + \beta + 1} + \frac{2\alpha - 2}{\alpha} \sqrt{T + \beta + 1} + \frac{\beta + 1 - \alpha}{\alpha} \leq \sqrt{c_2} \frac{\sqrt{\beta + 1}}{\alpha} \sqrt{T + \sqrt{T + \frac{2\alpha - 2}{\alpha} \sqrt{T + \beta + 1}} + \frac{\beta + 1 - \alpha}{\alpha} \leq C \sqrt{T},
\]

where
\[ C = \frac{\sqrt{c_2}}{\alpha} \left( \frac{\sqrt{\beta + 1} + 2\alpha - 2}{\alpha} + \frac{\beta + 1 - \alpha}{\alpha} \right). \]

APPENDIX B: PROOFS OF THE RESULTS
IN SECTION 3.3

In this section, we abbreviate the expectation value and probability \( \mathbb{E}_x^\varepsilon [\cdot] \) and \( \mathbb{P}_x^\varepsilon (\cdot) \) as \( \mathbb{E}_j [\cdot] \) and \( \mathbb{P}_j (\cdot) \) for \( j = 0, 1 \) with \( v^0 \) and \( v^1 \) as in (7).

**Proof of Lemma 2** First, note that if \( \ell' \notin S \), then \( \mathbb{P}_0 (Y = i | S) = \mathbb{P}_1 (Y = i | S) \). Hence, we may assume that \( \ell' \in S \). In addition, note that \( \mathbb{P}_j (\cdot | S) = \mathbb{P}_j (\cdot | S \cap \{ k \ell' \}) \) for \( j = 0, 1 \). Therefore, it suffices to check \( S = \{ k \ell' \} \) and \( S = \{ \ell' \} \). To this end, define
\[
p_i = \mathbb{P}_0 (Y = i | S) \quad \text{and} \quad q_i = \mathbb{P}_1 (Y = i | S),
\]
for \( i = 0, k, \ell' \),

where \( Y \) denotes a random purchase from \( S \). First, consider the case that \( S = \{ k \ell' \} \). Then, for \( i = 0, k, \ell' \) it holds that
\[
q_i = \frac{v_i^1}{1 + u_k + (1 + \varepsilon) u_{\ell'}} \geq \min \{ 1, u_{\ell'} \},
\]

since \( v_i^1 > u_{\ell'} \) and \( \varepsilon < 1 \) and \( u_{\ell'} < u_k \) and \( 1 < u_k \).

Next, note that
\[
(1 + u_k + (1 - \varepsilon) u_{\ell'})(1 + u_k + (1 + \varepsilon) u_{\ell'}) \geq 2(1 + u_k) u_{\ell'},
\]

and as a result
\[
|p_0 - q_0| \leq \frac{2u_{\ell'} \varepsilon}{2(1 + u_k) u_{\ell'}} = \frac{\varepsilon}{1 + u_k} < \varepsilon,
\]

and
\[
|p_k - q_k| = u_k |p_0 - q_0| \leq \frac{u_k \varepsilon}{1 + u_k} < \varepsilon,
\]

and
\[
|p_{\ell'} - q_{\ell'}| = \frac{2(1 + u_k) u_{\ell'}}{(1 + u_k + (1 - \varepsilon) u_{\ell'})(1 + u_k + (1 + \varepsilon) u_{\ell'})} \leq \varepsilon.
\]

Consequently,
\[
\text{KL} (\mathbb{P}_0 (\cdot | S) || \mathbb{P}_1 (\cdot | S)) \leq \sum_{i=0, \ell'} \frac{(p_i - q_i)^2}{q_i} \leq \frac{12u_k}{\min \{ 1, u_{\ell'} \}} \varepsilon^2 = c_3 \varepsilon^2,
\]

where the first inequality is easily verified (see, e.g., Lemma 3 from Chen and Wang (2018)) and
\[
c_3 = \frac{12u_k}{\min \{ 1, u_{\ell'} \}}.
\]

Now, consider the case that \( S = \{ \ell' \} \). Similarly, we derive for \( i = 0, \ell' \) that
\[
q_i = \frac{v_i^1}{1 + (1 + \varepsilon) u_{\ell'}} > \frac{\min \{ 1, u_{\ell'} \}}{1 + 2u_k} > \frac{\min \{ 1, u_{\ell'} \}}{3u_k},
\]

since \( v_i^1 > u_{\ell'} \) and \( \varepsilon < 1 \) and \( u_{\ell'} < u_k \) and \( 1 < u_k \).

Next, note that
\[
(1 + (1 - \varepsilon) u_{\ell'})(1 + (1 + \varepsilon) u_{\ell'}) \geq 2u_{\ell'},
\]

and as a result
\[
|p_0 - q_0| \leq \frac{2u_{\ell'} \varepsilon}{2u_{\ell'}} = \varepsilon,
\]

and
\[
|p_{\ell'} - q_{\ell'}| = \frac{2u_{\ell'}}{(1 + (1 - \varepsilon) u_{\ell'})(1 + (1 + \varepsilon) u_{\ell'})} \leq \varepsilon.
\]

Consequently,
\[
\text{KL} (\mathbb{P}_0 (\cdot | S) || \mathbb{P}_1 (\cdot | S)) \leq \sum_{i=0, \ell'} \frac{(p_i - q_i)^2}{q_i} \leq \frac{6u_k}{\min \{ 1, u_{\ell'} \}} \varepsilon^2 \leq c_3 \varepsilon^2,
\]

where the first inequality is again easily verified (see, e.g., Lemma 3 from Chen and Wang (2018)). The final statement of the lemma follows by applying Pinsker’s inequality and Le Cam’s method as follows. First, consider the entire probability measures \( \mathbb{P}_0 \) and \( \mathbb{P}_1 \). By the chain rule of the KL divergence it follows that
\[
\text{KL} (\mathbb{P}_0 || \mathbb{P}_1) \leq c_3 \varepsilon^2 T.
\]

From Pinsker’s inequality it follows that the total variation (TV) norm between \( \mathbb{P}_0 \) and \( \mathbb{P}_1 \) is bounded from above as
\[
\| \mathbb{P}_0 - \mathbb{P}_1 \|_{TV} = \sup_A |\mathbb{P}_0 (A) - \mathbb{P}_1 (A)|
\]

Next, Le Cam’s method entails to consider $B := \{ \psi = 0 \}$. Then, it follows that
\[
\max_{j=0,1} \mathbb{P}_j(\psi \neq j) \geq \frac{1}{2} (\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)) \\
\geq \frac{1}{2} (1 - (\mathbb{P}_0(B) - \mathbb{P}_1(B))) \\
\geq \frac{1}{2} (1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{TV}) \geq \frac{1}{2} \left(1 - \sqrt{2c_3} \varepsilon \sqrt{T}\right),
\]
which concludes our proof.

Proof of Lemma 3 First note that, for $j = 0, 1$, the expected profit under $v^j$ is independent of the inclusion of products $i \notin \{k, \ell\}$. That is, for all $S \subseteq [N],$
\[
\mathbb{E}(S, v^j) = \mathbb{E}(S \cap \{k, \ell\}, v^j).
\]
We start by bounding the regret under $v^0$ from below. By (B1) we only need to consider $S_i = \{k, \{\ell\}, \emptyset\}$. Then, as $u_{w_k} = u_{w_{\ell}}$ and $u_k = u_{\ell} + 1$ and $\varepsilon \in (0, 1/2]$, it follows that
\[
\mathbb{E}(S_i, v^0) = u_{w_{\ell}} \frac{2(1 - \varepsilon)}{(2 + (2 - \varepsilon)u_{\ell})(2 + u_{\ell})} \\
\geq u_{w_{\ell}} \frac{2 - \varepsilon}{(2 + 3u_{\ell})^2},
\]
as well as,
\[
\mathbb{E}(S_i, v^0) = u_{w_{\ell}} \frac{2 - \varepsilon}{(2 + 3u_{\ell})^2}.
\]
and, in addition,
\[
\mathbb{E}(S_i, v^0) = u_{w_{\ell}} \frac{2 - \varepsilon}{(2 + 3u_{\ell})^2}.
\]
In general, we know that
\[
\mathbb{E}(S_i, v^0) = u_{w_{\ell}} \frac{2 - \varepsilon}{(2 + 3u_{\ell})^2}.
\]
and by taking the expectation with respect to $v^0$ and summing over $t \in [T]$ we have shown (8) with
\[
c_4 = \frac{u_{w_{\ell}}}{(2 + 3u_{\ell})^2}.
\]
Next, we bound the regret under $v^1$ from below. Again by (B1) we only need to consider $S_i = \{k, \{\ell\}, \emptyset\}. Then, as $u_{w_k} = u_{w_{\ell}}$ and $u_k = u_{\ell} + 1$ and $\varepsilon \in (0, 1/2]$, it follows that
\[
r(\{k, \ell\}, v^1) - r(\{k\}, v^1) = u_{w_{\ell}} \frac{1 + \varepsilon}{(2 + 3u_{\ell})^2},
\]
as well as,
\[
r(\{k, \ell\}, v^1) - r(\{k\}, v^1) = u_{w_{\ell}} \frac{1 + \varepsilon}{(2 + 3u_{\ell})^2}.
\]
and, in addition,
\[
r(\{k, \ell\}, v^1) - r(\{k\}, v^1) = u_{w_{\ell}} \frac{1 + \varepsilon}{(2 + 3u_{\ell})^2}.
\]
Proof of Theorem 2 As discussed in Section 3.3, we have established the first two steps: constructing two preference vectors $v_0$ and $v_1$ and showing that—as a consequence of Lemma 2—for any estimator $\psi$ that has as input the observed purchases $Y_1, \ldots, Y_T$ and outputs either 0 or 1 it holds that
\[
\max_{j=0,1} \mathbb{E}_a(\psi \neq j) \geq \frac{1}{4}.
\]
It remains to show that $v^0, v^1 \in \mathcal{V}$ and finish the third step by establishing a contradiction for a specific estimator $\psi$. First, note that for both $j = 0, 1$ we have that
\[
\sum_{i=1}^{N} v_i^j \leq \frac{5}{2} u_k - \frac{3}{2},
\]
as $u_{\ell} = u_k - 1$ and $\varepsilon \leq 1/2$. Consequently, $v^0$ and $v^1$ lie in the class $\mathcal{V}$ since
\[
p_0 \leq \frac{2u_{\ell} - w_k}{5u_{\ell} + w_k} = \frac{2}{5u_k - 1}.
\]
We continue our proof by establishing a contradiction. To this end, recall that $C = \frac{937}{}
and recall the definitions of $\wp_0$, $\wp_1$, $L_0$, and $L_1$:

$$\wp_0 := \sum_{i=1}^{T} 1 \{ k, \ell \in S_i \} \quad \text{and} \quad \wp_1 := \sum_{i=1}^{T} 1 \{ k \notin S_i, \ell \in S_i \},$$

$L_0 := c_4 (\varepsilon \wp_1 + T - \wp_0 - \wp_1)$ and
$L_1 := c_4 (\varepsilon \wp_0 + T - \wp_0 - \wp_1)$.

Now, we assume that

$$\Delta_\pi(T) < C\sqrt{T}. \quad (B3)$$

As a consequence of the assumption above, we conclude by Markov’s inequality and Lemma 3 that

$$\mathbb{P}_0 \left( L_0 > 4C\sqrt{T} \right) \leq \frac{\mathbb{E}_0 L_0}{4C\sqrt{T}} \leq \frac{\Delta_\pi(T, v^0)}{4C\sqrt{T}} < \frac{1}{4},$$

and likewise

$$\mathbb{P}_1 \left( L_1 > 4C\sqrt{T} \right) \leq \frac{\mathbb{E}_1 L_1}{4C\sqrt{T}} \leq \frac{\Delta_\pi(T, v^1)}{4C\sqrt{T}} < \frac{1}{4}.$$ 

Next, define the estimator $\psi$ as

$$\psi := \begin{cases} 
0 & \text{if } \wp_0 > T/2, \\
1 & \text{if } \wp_0 \leq T/2.
\end{cases}$$

From $\psi = 1$ it follows that

$$\varepsilon \wp_1 + T - \wp_0 - \wp_1 > \varepsilon (T - \wp_0) \geq \frac{\varepsilon T}{2},$$

as $\wp_0 + \wp_1 \leq T$ and $\varepsilon < 1$. Now note that if $\varepsilon = 1/2$, then

$$\frac{\varepsilon T}{2} = \frac{T}{4} > \frac{T}{4\sqrt{2c_3}} \geq \frac{\sqrt{T}}{4\sqrt{2c_3}},$$

since $c_3 > 1$. Also, if $\varepsilon = \left(2\sqrt{2c_3}T\right)^{-1}$, then

$$\frac{\varepsilon T}{2} = \frac{\sqrt{T}}{4\sqrt{2c_3}}.$$

From this we conclude that $\psi = 1$ implies $L_0 > 4C\sqrt{T}$ and therefore

$$\mathbb{P}_0(\psi = 1) \leq \mathbb{P}_0 \left( L_0 > 4C\sqrt{T} \right) < \frac{1}{4}. \quad (B4)$$

Now consider $\psi = 0$. This implies

$$\varepsilon \wp_0 + T - \wp_0 - \wp_1 \geq \varepsilon \wp_0 \geq \frac{\varepsilon T}{2} \geq \frac{\sqrt{T}}{4\sqrt{2c_3}},$$

as $\wp_0 + \wp_1 \leq T$ and where the last inequality is established as before. From this we conclude that $\psi = 0$ implies $L_1 > 4C\sqrt{T}$ and therefore

$$\mathbb{P}_1(\psi = 0) \leq \mathbb{P}_1 \left( L_1 > 4C\sqrt{T} \right) < \frac{1}{4}. \quad (B5)$$

We conclude that (B4) and (B5) contradict (B2). Hence, the assumption in (B3) cannot be true. □