Learning Mixed Multinomial Logits with Provable Guarantees

Anonymous Author(s)
Affiliation
Address
email

Abstract

A mixture of multinomial logits (MMNL) generalizes the single logit model, which is commonly used in predicting the probabilities of different outcomes. While extensive algorithms have been developed in the literature to learn MMNL models, theoretical results are limited. Built on the Frank-Wolfe (FW) method, we propose a new algorithm that learns both mixture weights and component-specific logit parameters with provable convergence guarantees for an arbitrary number of mixtures. Our algorithm utilizes historical choice data to generate a set of candidate choice probability vectors, each being close to the ground truth with a high probability. We further provide a sample complexity analysis to show that only a polynomial number of samples is required to secure the performance guarantee of our algorithm. Finally, we conduct simulation studies to evaluate the performance and demonstrate how to apply our algorithm to real-world applications.

1 Introduction

Multinomial logit models (MNL) are widely used in a variety of settings to predict the probabilities of different outcomes. It is termed as multinomial logistic regression in statistics and its formula for choosing the j-th option, $\frac{\exp(v_j)}{\sum_i \exp(v_i)}$, is well known as the `softmax` function for multi-class classification in machine learning. In choice modeling, it is a type of discrete choice model used to analyze and predict people's decisions.

Mixed multinomial logit (MMNL) is a mixture model consisted of multiple individual MNL components and can be adopted when the target data is heterogeneous. For instance, Yang et al. [2018] demonstrates that using a single `softmax` does not have enough capacity to model natural language and claims this as a bottleneck. To alleviate this issue, they proposed a new method called *Mixture of Softmaxes* (MoS), which has higher expressiveness and can better incorporate the contextual information. Later, new techniques are proposed on top of this architecture to further boost the efficiency, such as the Mixtape from Yang et al. [2019]. On the other hand, McFadden and Train [2000] shows that MMNL models can lift the restrictive "independence of irrelevant alternatives" (IIA) assumption present in single MNL models as well as approximate any random utility-based discrete choice models to arbitrary degree of accuracy under appropriate assumptions.

Despite its importance, the study on the learning theory of MMNL is limited. A common approach in the literature is to assume a parametric family of the mixing distribution (cf.Train [2009] and Berry et al. [1995]). But this approach suffers from the model misspecification issue, i.e., there could be systematical errors in the estimation if the assumed parametric family is different from the true one. Expectation-Maximization (EM) algorithm is a widely used non-parametric method for mixture models. But it does not guarantee the estimators' convergence to the true value of the parameters. Recent inspiring work by Jagabathula et al. [2020] creatively applies the Frank-Wolfe

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

(FW) algorithms to estimate the MMNL model and establish the sub-linear convergence of the proposed method. However, the major issue with their approach is that the convergence result only applies to the aggregate predictors, and cannot recover each individual mixture. In this paper, we propose a new *Stochastic Subregion Frank-Wolfe* (`SSRFW`) algorithm, which can generate each individual mixture within an ϵ-ball to one of the ground truth vectors with a high probability.

We summarize the main contributions of this paper as follows. First, we provide a learning algorithm for MMNL models that comes with provable convergence guarantees of the estimators. In particular, each generated mixture from our algorithm is shown to be within an ϵ-ball to one of the ground truth with a high probability. To the best of our knowledge, this is the first paper to prove these results for MMNL models with an arbitrary number of mixture types and minimal assumptions on the model parameters. Second, we further quantify the sample complexity of the proposed algorithm. We show that only a polynomial number of samples is required to achieve a high probability convergence guarantee.

1.1 Preliminaries

Consider a setting where the decision makers need to choose among a set of $[M] = \{1, \ldots, M\}$ items, where each alternative has a feature vector $\boldsymbol{z}_j \in \mathbb{R}^d$. Furthermore, we assume the population consists of K different components (types), each associated with a unique "preference vector" $\boldsymbol{\beta}_k \in \mathbb{R}^d$ and a mixture weight α_k, where $\sum_{k=1}^K \alpha_k = 1$. In other words, $\boldsymbol{\beta}_k$'s are not drawn from a parametric mixing distribution and we consider MMNL in this paper as a type of latent class models, which is why we claim our method as *non-parametric*. This is consistent with the definitions used in other literature such as Train [2009] and Jagabathula et al. [2020].

$\boldsymbol{\beta}_k$'s can also be used as user embeddings with high interpretability for decision makers who belong to the same type, which represent the corresponding type's taste on different aspects of the alternatives.

Under linear utility assumption, the probability of choosing j-th item for a particular mixture k is

$$q_j(\boldsymbol{\beta}_k) = \frac{\exp(\boldsymbol{\beta}_k^\top \boldsymbol{z}_j)}{\sum_{i=1}^M \exp(\boldsymbol{\beta}_k^\top \boldsymbol{z}_i)}$$

and the aggregated counterpart for the entire population is

$$\mathrm{g}_j = \sum_{k=1}^K \alpha_k q_j(\boldsymbol{\beta}_k), j \in [M]$$

To simplify notation, we use $\boldsymbol{q}_k$ to denote the logit vector $\boldsymbol{q}(\boldsymbol{\beta}_k) = [q_1(\boldsymbol{\beta}_k), \cdots, q_M(\boldsymbol{\beta}_k)]^\top \in \mathbb{R}^M$, and q_{kj} for $\boldsymbol{q}_j(\boldsymbol{\beta}_k)$, $j \in [M]$, $\forall k$. We will also refer to these logit vectors as *choice probability vectors*.

1.2 Related Work

Learning MMNL models remains an open-research problem. One common approach in the literature is to presume a parametric family of distributions on the parameters and apply parametric estimation methods such as the maximum likelihood estimation (MLE) method (c.f. Train [2009]) or the least square regression model to compute the parameters. A well-known work in this regime is by Berry et al. [1995]. They assume the component-specific parameters are normal distributed and proposed a two-step estimation method to learn the parameters using an aggregated market share data. A potential issue for this stream of methods is model misspecification, which leads to inaccurate predictions. In particular, if the assumed parametric family is different from the true one, there could be systematical errors in the estimation.

Non-parametric estimation is also widely recognized when learning MMNL models, with the most commonly known and used method being the Expectation-Maximization (EM) algorithm (c.f. Dempster et al. [1977]). Train [2008] studied three different types of EM algorithms using historical choice data from each individual decision maker. The two main restrictions for the EM algorithm are 1) the algorithm can get stuck in local optimum (hence no guarantee on the convergence of the estimators) and 2) the number of mixture types needs to be pre-specified (additional heuristics are needed if such information is not known in advance, which may also introduce undesired noises).

Jagabathula et al. [2020] recently developed an estimation method for mixture models also from a non-parametric perspective based on the Frank-Wolfe (FW) algorithm (cf. [Frank and Wolfe, 1956], Jaggi [2013], [Jaggi and Sulovskỳ, 2010], [Bach, 2013], Harchaoui et al. [2015],Lacoste-Julien and Jaggi [2015]), which is an iterative method originally designed to solve constrained quadratic optimization. They manage to establish a sublinear convergence rate of the proposed estimation method for MMNL models. But the convergence result only applies to the aggregate choice probabilities. In contrast, we provide the convergence guarantee for each individual choice probability (more details about their approach and the comparisons will be discussed in Section 2).

There is also another stream of work related to MMNL models from a theoretical learning perspective. The key idea is to solve a system of equations in terms of all the parameters and prove that there is one and only one solution (identifiability). Due to the entanglement of the individual logit parameters and mixture weights, as well as the high non-linearity in the system of equations, only 2-MNL models (MMNL with two mixture types) have been studied so far. Chierichetti et al. [2018] considers the setting where the mixture weights are equal (i.e., each mixture represents 50% of the population) and Tang [2020] recently studied 2-MNL with unknown weights. This paper enriches the theoretical learning literature for MMNL models by generalizing the analysis to an arbitrary number of mixture types and with minimal assumptions on the model parameters.

1.3 Main Results

The main contribution of this paper is to provide a learning algorithm for MMNL models with provable convergence guarantees of the estimators. Specifically we will show that each of the generated choice probability vectors from our algorithm is within an ϵ-ball to one of the ground truth vectors with a high probability. Subsequently, we provide the sample complexity analysis of the algorithm. We name our algorithm the Stochastic Subregion Frank-Wolfe (`SSRFW`) algorithm, while is inspired by the Frank-Wolfe framework by Jagabathula et al. [2020] but has a carefully constructed stochastic feasible region that is learned from the data.

The `SSRFW` framework assumes to work with repeated choice data, such as consumer panel or survey data. This is a reasonable assumption in many real-world scenarios (cf. Revelt and Train [1998] and Brownstone et al. [2000]). We assume the data consists of historical choice decisions from the target population of size N. Specifically, for each time period $t = 1, \ldots, T$, we assume each decision maker i's choice $X_i^{(t)}$ are i.i.d categorical random variables with support $x \in [M]$ and probability mass function (PMF) specified by $\boldsymbol{q}_k$ if he/she is of type k.

For the MMNL to be learnable, we need to assume that none of the individual MNL is identical to the rest. Let F and G be the cumulative distribution functions (CDF) for two different choice probability distribution $\boldsymbol{q}_k$ and $\boldsymbol{q}_{k'}$ respectively. Define the Kolmogorov-Smirnov (KS) distance between the two distributions as $D_{\text{KS}}(\boldsymbol{q}_k, \boldsymbol{q}_{k'}) = \sup_x |F(x) - G(x)|$. Definition 1 below is adapted from a similar assumption introduced in Kalai et al. [2010] to impose some regularity in the MMNL model.

Definition 1. *We call a mixed multinomial logit model (MMNL)* $\mathbf{g}$ *ϵ-standard if*

1. $\min_k \alpha_k \geq \epsilon$
2. $\beta_{kd} \leq \frac{1}{\epsilon}, \forall k, \forall d$
3. $D_{KS}(q_k, q_{k'}) \geq \epsilon, \forall k \neq k'$

Definition 1 imposes the following properties on the MMNL model: 1) none of the mixtures has zero weight; 2) all parameters in the model are bounded; and 3) each individual logit model has some level of separation from others. We consider this as a very mild assumption in our setting, as later we will see that we will not predetermine the number of mixtures (K) and if a mixture/component is not present in the observed data (since it has zero weight), the algorithm will simply not learn it. This is also one of the advantages of our algorithm over the traditional approaches, such as EM.

The main result is as follows:

Theorem 1. *Let* $\mathbf{g} = \sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$ *be a mixed multinomial logit (MMNL) model over a set of* M *items. Assume* $M \geq K$*. For any* $\epsilon > 0$*,* $0 < \delta < 1$*, Algorithm 1 outputs an MMNL* $\hat{\mathbf{g}} = \sum_{k=1}^{K'} \hat{\alpha}_k \hat{\boldsymbol{q}}_k$ *where* $K' \geq K$ *such that, with probability* $\geq 1 - \delta$*, there exists a many-to-one mapping* $\pi : k' \mapsto$

$k, k' \in [K'], k \in [K]$ *such that*

$$\left\|\hat{\boldsymbol{q}}_{k'} - \boldsymbol{q}_{\pi(k')}\right\| < \epsilon, \forall\, k',$$

and

$$\left|\sum_{k':\pi(k')=k} \hat{\alpha}_{k'} - \alpha_k\right| < \epsilon, \forall\, k.$$

where $\|\cdot\|$ *is the Euclidean norm. The number of samples required by Algorithm 1 is* $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$.

The outline of the paper is as follows: Section 2 and Section 3 will introduce the main algorithms, namely the Stochastic Subregion Frank-Wolfe (`SSRFW`) and the $\mathcal{Q}$-construction algorithm, where the output of the latter serves as an input to the former. In Section 4, we will discuss the theoretical properties of the `SSRFW` algorithm. We conduct simulation studies in Section 5. All the detailed proofs and parts of simulation studies are included in the appendix.

2 Stochastic Subregion Frank-Wolfe Algorithm

Recall that we observe each decision maker i's choice at time t: $x_i^{(t)}$ (as realization of $X_i^{(t)}$). Let $Y_i^t = [0, \cdots, 1, \cdots, 0]^\top \in \mathbb{R}^M$, where $Y_{ij}^t = 1$ if $x_i^{(t)} = j$. The observed aggregated choice probability vector for period t can be expressed as

$$\mathbf{y}^t = \frac{1}{N}\sum_{i=1}^{N} Y_i^t$$

where $\mathbf{y}^t \in \mathbb{R}^M$. Define $\mathcal{P} = \{\boldsymbol{q}(\beta) \in \mathbb{R}^M | \boldsymbol{\beta} \in \mathbb{R}^d\}$, the set of all valid logit vectors for the given choice set and their features. Denote $\overline{\mathcal{P}}$ as its closure. Since $\mathbf{g} = \sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$, by definition we have $\mathbf{g} \in \text{Conv}(\overline{\mathcal{P}})$.

Our main learning objective is to minimize the difference between the observed (from data) and theoretical aggregated choice probability, as follows

$$\min_{\mathbf{g}\in\text{Conv}(\overline{\mathcal{P}})} \mathcal{L}(\mathbf{g}) \equiv \min_{\mathbf{g}\in\text{Conv}(\overline{\mathcal{P}})} \frac{1}{2}\sum_{t=1}^{T}\left\|\mathbf{g} - \mathbf{y}^t\right\|^2 \tag{1}$$

Note that problem (1) is a non-convex problem in α_k and $\boldsymbol{\beta}_k$ and is in general hard to solve. We adopt the conditional gradient approach proposed by Jagabathula et al. [2020], which we will refer to as the *original FW* algorithm. Each iteration k is consisted of two steps: 1) the *supporting finding step* which searches for a new direction represented by a choice probability vector $\boldsymbol{q}^{(k)}$ (as a new latent class k) via solving a linear optimization subproblem in the search space $\text{Conv}(\overline{\mathcal{P}})$, and 2) the *proportion update step* which updates the mixture proportion $\alpha_i, i = 0, \ldots, k$, assigned to all choice probability vectors obtained so far. The final learning outcome $\hat{\mathbf{g}}$ is $\sum_{i=0}^{k} \alpha_k \boldsymbol{q}^{(k)}$ where $\boldsymbol{q}^{(0)} = \mathbf{g}^{(0)} \in \overline{\mathcal{P}}$ is a random initialization. This setup corresponds to a fully-corrective variant of the generic Frank-Wolfe algorithm, and is guaranteed to find the optimal solution to the optimization problem [Jaggi, 2013] for the decision variable $\mathbf{g}$.

The problem with the original FW approach is that the feasible region $\text{Conv}(\overline{\mathcal{P}})$ contains the complete set of all logit vectors and their limiting points. The former corresponds to *non-boundary types*, whose choice model can be characterized by a standard MNL choice model, and the latter corresponds to *boundary types*, whose standard MNL parameters become unbounded, resulting infinite utility for some options and zero for the rest. Such broad search space $\text{Conv}(\overline{\mathcal{P}})$ in FW generates mixture compositions with a considerable number of the boundary types. More specifically, this problem is rooted to the linear subproblem solved at each *support finding step*, whose optimal solution always lies on the extreme point of the feasible region which contains many of these limiting choice probability vectors $\in \overline{\mathcal{P}} \setminus \mathcal{P}$ as its extreme points. Often, each limiting choice probability vector corresponds to an unbounded $\boldsymbol{\beta}_k$ and cause the learning outcome unusable for downstream applications. Though significant effort is made to justify the legitimacy of these boundary types in the paper, the convergence result only applies to the population's aggregate choice probability, i.e. $\hat{\mathbf{g}} \to \sum_k \alpha_k \boldsymbol{q}_k$, which is

an Frank-Wolfe property. In other words, their method fails to recover each of the individual logit models ($\boldsymbol{\beta}_k$) and the corresponding mixture weights (α_k) accurately.

To remedy this issue and seek convergence guarantees for the general MMNL estimation problem, we designed the Stochastic Subregion Frank-Wolfe (SSRFW) algorithm, as shown in Algorithm 1.

Algorithm 1: Stochastic Subregion Frank-Wolfe

Input: data $\mathbf{y}$, $\mathcal{Q}$ from Algorithm 2

Initialization: $k = 0$; $\boldsymbol{\alpha}^{(0)} = [1]$, a random $\mathbf{g}^{(0)} = \boldsymbol{q}^{(0)}$ chosen from $\mathcal{Q}$

while *stopping condition not met* **do**

  $k \leftarrow k + 1$

  Compute $\boldsymbol{q}^{(k)} = \underset{\boldsymbol{v} \in \text{Conv}(\mathcal{Q})}{\arg\min} \langle \nabla \mathcal{L}\left(\mathbf{g}^{(k-1)}\right), \boldsymbol{v} - \mathbf{g}^{(k-1)} \rangle$

  Compute $\boldsymbol{\alpha}^{(k)} = \underset{\boldsymbol{\alpha} \in \Delta_k}{\arg\min} \mathcal{L}\left(\alpha_0^{(k)} \mathbf{g}^{(0)} + \sum_{s=1}^{k} \alpha_s^{(k)} \boldsymbol{q}^{(s)}\right)$ [1]

  Update $\mathbf{g}^{(k)} := \alpha_0^{(k)} \mathbf{g}^{(0)} + \sum_{s=1}^{k} \alpha_s^{(k)} \boldsymbol{q}^{(s)}$

end

Output : choice prob. $\boldsymbol{q}^{(0)}, \ldots, \boldsymbol{q}^{(k)}$

mixture weights. $\boldsymbol{\alpha}^{(k)} \in \Delta_k \subset \mathbb{R}^{k+1}$

Note in the above algorithm, after k iterations, we have generated k choice probability vectors $\boldsymbol{q}^{(s)}$, $s = 1, \ldots, k$, all of which will be the output of the learning algorithm. The mixture weight estimator is slightly different. $\boldsymbol{\alpha}^{(k)} \in \mathbb{R}^{k+1}$ represents the mixture weight vector for the $k+1$ choice probability vectors we have generated so far, where $\alpha_s^{(k)}$ refers to the s-th element in the vector, corresponding to the mixture associated with $\boldsymbol{q}^{(s)}$. In other words, the vector length of $\boldsymbol{\alpha}$ increases by 1 for each iteration and only the final vector $\boldsymbol{\alpha}^{(k)}$ is outputted from the algorithm.

The key different in SSRFW is that it takes in an additional input $\mathcal{Q}$ from the $\mathcal{Q}$ *Construction Algorithm* (discussed in detail in Section 3), which is the key to eliminate the possibility of producing *boundary types*. In particular, the new feasible region Conv($\mathcal{Q}$) is a subset of Conv($\overline{\mathcal{P}}$), where each element $\boldsymbol{q} \in \mathcal{Q}$ will be learned from data and is guaranteed to be within an ϵ-ball of the true choice probability vector for some mixture type with high probability. Subsequently, this also ensures the SSRFW algorithm to recover the mixture weight for each latent class. Note that the extreme points of Conv($\mathcal{Q}$) by definition is a subset of $\mathcal{Q}$. Since we only care about the extreme points of the feasible region, we can safely ignore any points $\in$ Conv($\mathcal{Q}$) $\setminus$ $\mathcal{Q}$ and replace Conv($\mathcal{Q}$) with $\mathcal{Q}$ in Algorithm 1.

Since Frank-Wolfe will converge to the optimal solution [Jaggi, 2013], and as we will see in Section 4 that the optimal solution to problem (1) is an interior point of the feasible region Conv($\mathcal{Q}$), the stopping condition can simply be set as $\left\|\mathbf{g}^{(k)} - \mathbf{y}\right\| \leq \epsilon$.

3 $\mathcal{Q}$ Construction Algorithm

Our primary goal in this section is to learn the input set $\mathcal{Q}$ for SSRFW (Algorithm 1). In particular for this set $\mathcal{Q} = \{\hat{\boldsymbol{q}}_\ell\}_{l=1,\ldots,L}$, we require that $\forall\ \hat{\boldsymbol{q}}_\ell$, there exists some mapping $\pi : [L] \to [K]$ such that $\left\|\hat{\boldsymbol{q}}_\ell - \boldsymbol{q}_{\pi(\ell)}\right\| \leq \epsilon$, where $\boldsymbol{q}_{\pi(\ell)}$ is one of the ground truth logit vectors. Once this property holds, each of the SSRFW 's $\boldsymbol{q}$ outputs will also be ϵ-close to some true choice probability vector as the algorithm always selects an extreme point of Conv($\mathcal{Q}$) in each iteration.

The two major components for the $\mathcal{Q}$ Construction Algorithm is first computing a distance score matrix (Section 3.1) and then creating subsamples that contain decision makers from only one mixture type through random seeding and non-uniform sampling (Section 3.2). Section 3.3 presents the algorithm and discusses the properties of the learning outcomes.

[1] $\Delta_k = \{(\alpha_1, \ldots, \alpha_{k+1}) \in \mathbb{R}^{k+1} \big| \sum_{i=1}^{k+1} \alpha_i = 1, \alpha_i \geq 0\}$

3.1 Distance Score Matrix

We first define a distance score matrix S, whose element s_{ij} measures the dissimilarity between any two decision makers i and j in term of their choice decisions. Recall that each $x_i^{(t)}$ are realizations of i.i.d categorical random variables with PMF $\boldsymbol{q}_k$ if i is from mixture k. Denote its associated empirical cumulative distribution function (CDF) as $F_T(x;i)$ when a total of T decisions are observed. In particular,

$$F_T(x;i) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{x_i^{(t)} \leq x\}}, x \in [M]$$

Define the distance score function for each pair of decision makers as

$$s_{ij} := s(i,j) = ||F_T(x;i) - F_T(x;j)||_\infty$$

where $||\cdot||_\infty$ represents the infinity norm.

3.2 Subsample Construction

We now want to create a number of subsamples such that each subsample of decision makers contains only one mixture type with high probability. Each of these subsamples is generated following a procedure that contains a *random seeding* step, and a *subsampling* step. These steps are repeated L times to create a $\mathcal{Q}$ set of size L.

Random Seeding. This step randomly samples a decision maker $i \in [N]$ from the population and can be done with simple uniform sampling technique. The selected decision maker is called a *seed*. We will generate L seeds and later investigate how large L needs to be in order to cover all mixture types in $\mathcal{Q}$.

Subsampling Strategy. This step generates an index set $I(i)$ for a seed i such that, with high probability, decision makers whose indices belonging to this set are of the same type as i. $I(i)$ is called the *subsample originated from seed* i.

Given a selected seed i, we first calculate an *accepting probability* $p_{j|i}$ which determines the likelihood of another decision maker j being accepted to the corresponding subsample $I(i)$. It is designed to be a monotonically decreasing function with respect to the distance score we defined in Section 3.1:

$$p_{j|i} = f\left(s(i,j)\right)$$

A simple example can be $p_{j|i} = 1 - s(i,j)$. Intuitively, when $s(i,j)$ is small, seed i and decision maker j's empirical CDF is close to each other, indicating there is a higher chance that they are of the same type and sharing the same choice probability $\boldsymbol{q}$. Subsequently, $p_{j|i}$ will be larger compared to a large value of $s(i,j')$ for another person j', resulting we accept j with higher probability than j'. This is consistent with our objective that we want each subsample to be composed of decision makers of the same type as the seed.

For implementation, we first initiate an empty index set $I(i)$ for the selected seed i. Then we repeat the following two steps until we reach a desired subsample size n: 1) draw a random sample from set $[N] \setminus I(i)$, 2) accept this sample into the index set with probability $p_{j|i}$ and reject with $1 - p_{j|i}$.

3.3 Algorithm and Properties

The remaining step is to obtain the set $\mathcal{Q}$ whose elements $\hat{\boldsymbol{q}}_\ell$ are computed as the average of historical choice decisions from decision makers in the subsamples I_ℓ, i.e. $\hat{\boldsymbol{q}}_\ell = \frac{1}{nT}\sum_{i \in I_\ell}\sum_{t=1}^{T} Y_i^{(t)}$. Finally, we present the $\mathcal{Q}$ Construction Algorithm in Algorithm 2.

Discussion We next discuss a few properties of the two algorithms.

`SSRFW` returns the logit vectors $\boldsymbol{q}$ instead of the parameters $\boldsymbol{\beta}$. We can simply perform maximum likelihood estimation under the single MNL setting where, if the extreme point corresponding to

Algorithm 2: The $\mathcal{Q}$ construction algorithm

Input : score matrix S, number of subsamples L, subsample size n
Initialization: $\mathcal{Q} = \texttt{set}()$
for $\ell \leftarrow 1$ **to** L **do**
  Choose seed: $i \sim U(0, N)$
  Initiate: $I_\ell = \texttt{set}()$
  while $|I_\ell| \neq n$ **do**
    $j \leftarrow \texttt{random_sample}([N] \setminus I_\ell)$
    Generate $u \sim U(0, 1)$
    if $u < p_{j|i}$ **then**
      $I_\ell.\texttt{add}(j)$
    end
  end
  Compute $\hat{\boldsymbol{q}}_\ell = \dfrac{1}{nT} \sum_{i \in I_\ell} \sum_{t=1}^{T} Y_i^{(t)}$
  $\mathcal{Q}.\texttt{add}(\hat{\boldsymbol{q}}_\ell)$
end
Output: $\mathcal{Q}$

subsample ℓ is selected during an SSRFW iteration, the log-likelihood is

$$l_\ell(\boldsymbol{\beta}) = -\sum_{t=1}^{T} \sum_{i \in I_\ell} \sum_{j \in [M]} Y_{ij}^{(t)} \left(\log \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{z}_j)}{\sum_{m \in [M]} \exp(\boldsymbol{\beta}^\top \boldsymbol{z}_m)} \right)$$

and we have $\hat{\boldsymbol{\beta}}_\ell = \arg\max_{\boldsymbol{\beta}} \; l_\ell(\boldsymbol{\beta})$. With the assumption of linear utility in $\boldsymbol{\beta}$, we can obtain the optimal solution using MLE (c.f. McFadden [1974]).

In general, we prefer learning $\boldsymbol{\beta}_k$'s compared to $\boldsymbol{q}_k$'s. This makes the algorithm more robust when item attributes change over time which results in change in the $\boldsymbol{q}_k$'s whereas $\boldsymbol{\beta}_k$ values persist. To achieve this, we just need to add two additional steps after the index set is created:

- **MNL paramter estimation:** apply the MLE step outlined above; create $\mathcal{B} = \{\hat{\boldsymbol{\beta}}_\ell\}_{\ell=1,\ldots,L}$
- **Map $\mathcal{B}$ to $\mathcal{Q}$:** fix item attributes of current interest, calculate $\mathcal{Q} = \{\hat{\boldsymbol{q}} | \hat{\boldsymbol{q}} = \boldsymbol{q}(\hat{\boldsymbol{\beta}}_\ell), \hat{\boldsymbol{\beta}}_\ell \in \mathcal{B}\}$

Next, stochasticity in our algorithm comes from two parts, namely the realization of each decision makers' choices, as well as the randomness in the process of generating subsamples. Yet we will show that under appropriate assumptions, SSRFW can still recover all individual mixture parameters with high probability.

In addition, we do not impose any restriction on the subsamples created during $\mathcal{Q}$ construction to be mutually exclusive. As long as they only contain one single mixture type (i.e. with high purity), it creates a legitimate estimation of $\hat{\boldsymbol{q}}$ or $\hat{\boldsymbol{\beta}}$ with MLE.

Last but not least, this is a more robust strategy than clustering algorithms to segment the population. Not only do we not require the number of mixtures K as a hyper-parameter, each decision maker's data being used more than once for the estimation can be thought of as a special bootstrap mechanism that has custom weights tailored to our objective in creating homogeneous subsamples.

4 Theory of the SSRFW Algorithm

Theorem 1 is the main result of this paper. To prove it, we will break down this section into two parts, where the first part will examine the provable convergence properties of the estimators, and the second part will focus on the sample complexity. We will outline the proof sketch but all detailed proofs can be found in the Appendix.

Provable Convergence. Assume K' is the total number of iterations that the algorithm has performed before reaching the stopping criteria. To simplify notation, we denote the generated outcome from k-th iteration $\boldsymbol{q}^{(k)}$ in SSRFW as $\hat{\boldsymbol{q}}_k$ to indicate they are the estimators and the final updated mixture weight corresponding to $\boldsymbol{q}^{(k)}$, $\alpha_k^{(K')}$, as $\hat{\alpha}_k$. In other words, we have $\mathbf{g}^{\texttt{SSRFW}} = \sum_{k'=1}^{K'} \hat{\alpha}_{k'} \hat{\boldsymbol{q}}_{k'}$, We use a similar notation $\mathbf{g}^{\texttt{FW}}$ for the original FW approach.

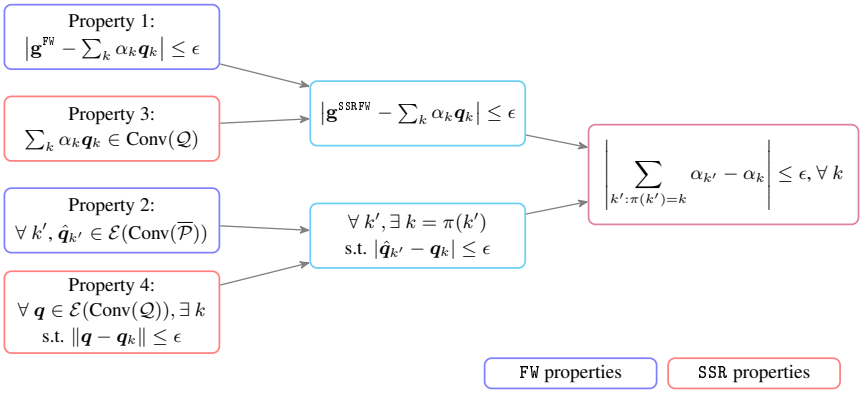


Figure 1: Proof sketch of Thoerem 1

Figure 1 describes the necessary components and steps of proving the result. In particular, Property 1 and 2 (purple boxes) are adapted from existing results of the Frank-Wolfe framework [Jaggi, 2013]. In Appendix A.1.1, we prove Property 1 in Lemma 2. Property 2 follows from the fact that each of the $\hat{\boldsymbol{q}}_{k'}$, obtained by solving a linear optimization problem, has to be an extreme point (denoted by $\mathcal{E}(\cdot)$) of the feasible region.

On the other hand, Property 3 and 4 (red boxes) hold as a consequence of the stochastic subregion we constructed in Algorithm 2. Property 3 claims that with high probability, the ground truth aggregated choice probability vector is still an interior point of the shrunken feasible region $\text{Conv}(\mathcal{Q})$ (proof in Appendix A.1.2). Together with Property 1, this tells us that $|\mathbf{g}^{\texttt{SSRFW}} - \sum_k \alpha_k \boldsymbol{q}_k| \leq \epsilon$. That is to say the aggregated choice probability vector of SSRFW also converges to the ground truth. To achieve Property 4 from the $\mathcal{Q}$ construction algorithm, we need to supply enough of data, the details of which will be discussed in the sample complexity part. Combining Property 2 and 4, we have $\forall\, k', \exists\, k = \pi(k')$ s.t. $|\hat{\boldsymbol{q}}_{k'} - \boldsymbol{q}_k| \leq \epsilon$. Finally, by doing some algebraic manipulation, we can obtain $\left|\sum_{k':\pi(k')=k} \alpha_{k'} - \alpha_k\right| \leq \epsilon$.

Sample Complexity. Recall we use T as the number of choice data records for each decision maker. Denote $T_{\min}$ as the minimum number required and $T_{\min} \geq 1$.

Define *in-types* to be the decision makers who share the same type as the seed, and *out-types* to be the ones of different types. We focus on the following three properties:

- probability of accepting the in-types,
- probability of rejecting the out-types,
- probability of covering all mixture types.

Consider a seed i, with type $k_i = k$. Assume there is a total of m_k decision makers in the population that is of type k.

Accepting In-Types. Denote the empirical choice probability for the seed i as F_n and that for another decision maker j as G_m, with n and m representing the number of independent price experiments from i and j, respectively. Recall that $s(i,j) = ||F_n(x) - G_m(x)||_\infty = \sup_x |F_n(x) - G_m(x)|$. For exposition simplicity, we use F and G to denote $F(x)$ and $G(x)$ in the subsequent presentation if no confusion incurred.

Theorem 2 (Sample Complexity I)**.** *Assume i and j are of the same type. Define $p_{j|i} = 1 - s(i,j)$. $\forall \delta > 0, \epsilon > 0$, we can achieve $p_{j|i} > 1 - \epsilon$ with probability at least $1 - \delta$ with sample size $T_{\min} = \mathcal{O}(\frac{1}{\epsilon^2} \log\left(\frac{1}{\delta}\right))$.*

Theorem 2 indicates that with sufficient amount of data, there is a high probability to accept a in-type—who shares the same type as the seed—into the subsample if it is chosen from the population after the random draw.

Rejecting Out-Types. Assume a decision maker j is of a different type from the seed i. We use F and G to denote their CDFs. Let F_n and G_m represent the corresponding empirical CDFs based on samples drawn from the two different distributions with sample size n and m, respectively. $s(i,j)$ is defined the same as before.

Theorem 3 (Sample Complexity II)**.** *Assume F and G correspond to the choice CDFs of two different types and $\sup_x |F(x) - G(x)| \geq \xi$. Let F_n and G_m be independent empirical distribution based on m and n i.i.d. samples drawn from F and G, respectively. Denote $T_{min} = min\{m, n\}$. Define $p_{j|i} = 1 - s(i,j)$. $\forall \delta > 0, \epsilon > 0$, we can achieve $p_{j|i} < 1 - \xi + \epsilon$ with probability at least $1 - \delta$ with $T_{min} = \mathcal{O}(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$.*

Theorem 3 complements Theorem 2 to rule out the out-types from being selected with a high probability. The more separable (larger ξ) their underlying distributions are, the lower accepting probabilities are.

Lemma 1. *For $\epsilon > 0$ and $0 < \delta < 1$, with minimum of choice records T required by Theorem 2 and Theorem 3, $\forall\ \boldsymbol{q}_\ell \in \mathcal{Q}$, $\left\|\boldsymbol{q}_\ell - \boldsymbol{q}_{\pi(\ell)}\right\| \leq \epsilon$ with probability at least $1 - \delta$, where $\hat{\boldsymbol{q}}_\ell = \frac{1}{nT} \sum_{i \in I_\ell} \sum_{t=1}^{T} Y_i^{(t)}$.*

Lemma 1 states that with enough data samples, the $\mathcal{Q}$-construction algorithm achieves Property 4, as shown in Figure 1, which `SSRFW`'s desired provable convergence builds upon. With Lemma 1, we conclude the proof for Theorem 1. The complete proof can be found in Appendix A.4.

As a final remark, we discuss the number of subsamples (i.e. L) needed in the $\mathcal{Q}$-construction process. We distinguish this from sample complexity as we are not requesting more data points with a larger L. Instead, we should view L as a computational complexity factor. Recall that each moving direction derived from the *support finding step* is viewed as a learned logit vector of one mixture component. Therefore, we want the $\mathcal{Q}$-construction algorithm to include at least one $\hat{\boldsymbol{q}}_\ell$ for each mixture k in order for it to be picked in `SSRFW` . This impose a requirement on L. Theorem 4 states that the expected number of the number of subsamples is controlled by the smallest mixture weight.

Theorem 4. *Assume $\alpha_1 \leq \alpha_2 \leq \cdots \alpha_K$. The expected number of subsamples L we need to construct is bounded by $\frac{1}{\alpha_1} \log \frac{1}{\alpha_1}$.*

Intuitively, this means if we have one mixture component that is very under-represented, then we need to create more subsamples to ensure that one of the seeds belongs to this mixture.

5 Conclusion

In this work, we propose the `SSRFW` algorithm, which provides an end-to-end solution for learning MMNL models using historical choice data. This novel approach utilizes a carefully designed sampling method to construct a meaningful search space. Not only does it resolve the drawback of the original Frank-Wolfe approach with boundary-type issues but also enables us to obtain provable guarantees for the model estimates and sample complexity. Another major advantages of our algorithm is that we do not assume prior knowledge of the number of mixtures K in the model. It is also more robust than traditional unsupervised approaches such as clustering method and EM.

In marketing science and operations research, MMNL provides a fundamental tool for demand prediction, which many downstream applications such as pricing and inventory position rely on. For instance, consider a pricing model that tries to maximize the revenue by solving the following optimization problem: $\max_{\boldsymbol{p}} \sum_{j=1}^{M} p_j \sum_{k=1}^{K} \alpha_k q_j(\boldsymbol{\beta}_k; \boldsymbol{p})$ where $p_j \geq 0, \forall j \in [M]$. This requires an accurate estimation of each individual MNL components. Due to page limit, we include our simulation studies and application in real-world setting in the appendix. We will also briefly discuss the requirements on the datasets in order to apply our algorithm.

References

Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

David Brownstone, David S. Bunch, and Kenneth Train. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological*, 34(5):315–338, 2000. ISSN 0191-2615.

Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Learning a mixture of two multinomial logits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 961–969. PMLR, 10–15 Jul 2018.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. A conditional gradient approach for nonparametric estimation of mixing distributions. *Management Science*, 2020.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, number CONF, pages 427–435, 2013.

Martin Jaggi and Marek Sulovskỳ. A simple algorithm for nuclear norm regularized problems. In *ICML*, 2010.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, pages 553–562, 2010.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in neural information processing systems*, pages 496–504, 2015.

D. McFadden and K. Train. Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

Daniel McFadden. *Frontiers in Econometrics*. Academic Press, New York, 1974.

David Revelt and Kenneth E. Train. Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics*, 80:647–657, 1998.

Wenpin Tang. Learning an arbitrary mixture of two multinomial logits, 2020.

Kenneth E Train. Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

James. G. Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. In *ICLR*, 2018.

Zhilin Yang, Thang Luong, Ruslan Salakhutdinov, and Quoc V. Le. Mixtape: Breaking the softmax bottleneck efficiently. In *NeurIPS*, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Appendix B.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix B.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 1, Section 4, and Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 4, and Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix B.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix B.
   (b) Did you mention the license of the assets? [Yes] See Appendix B.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

A.1 Proofs for Provable Convergence

A.1.1 Proof of Property 1

Lemma 2. *Denote* $\mathbf{g}^*$ *as the optimal solution to the Stochastic Subregion Frank-Wolfe algorithm and* $\mathbf{g}^{(k)}$ *denote the* k*-th iterate generated by Algorithm 1. Then*

$$\mathcal{L}(\mathbf{g}^{(k)}) - \mathcal{L}(\mathbf{g}^*) \leq \frac{4}{k+2}$$

for all $k \geq K$.

Proof: This Lemma follows directly from the existing results of the original Frank-Wolfe algorithm and its variants [Jaggi, 2013], which states that for an optimization problem $\min_{x \in \mathcal{D}} f(\mathbf{x})$ where f is a convex and continuously differentiable function and that the domain $\mathcal{D}$ is a compact convex set of any vector space, then for each $k \geq 1$, the iterates $\mathbf{x}^{(k)}$ of the fully-corrective Frank-Wolfe algorithm satisfy:

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq \frac{2 \cdot C_f}{k+2} \tag{2}$$

where C_f, defined as

$$C_f := \sup_{\substack{\boldsymbol{x}, \boldsymbol{s} \in \mathcal{D} \\ \gamma \in [0,1] \\ \boldsymbol{r} = \boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{x})}} \frac{2}{\gamma^2} \left(f(\boldsymbol{r}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{r} - \boldsymbol{x} \rangle \right),$$

is the *curvature constant*, which measures the "non-linearity" of function f over domain $\mathcal{D}$. The type of the Frank-Wolfe we use in Algorithm 1 is precisely the fully-corrective variant in that we optimize for α's in each iteration.

Claim 1. $\mathcal{L}(\mathbf{g}; \mathbf{y}) = ||\mathbf{g} - \mathbf{y}||^2$ *is a twice differentiable convex function. Conv*$(\mathcal{Q})$ *is a compact convex set.*

The first statement in Claim 1 is true by definition. The second statement can be shown by observing that $\mathcal{Q}$ is a finite set, hence compact, followed by the fact that convex hulls of compact set are compact.

For squared loss function $\mathcal{L}$ used in our model, Jagabathula et al. [2020] proved that $C_{\mathcal{L}} \leq 2$. The result of Lemma 2 follows by plugging $C_{\mathcal{L}}$ into Equation 2.

□

Note that with enough sample, we also have $\left\| \mathbf{y} - \sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k \right\| \leq \epsilon$ by the law of large numbers, which leads to $\mathbf{g}^{\texttt{SSRFW}}$ converges to $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$ with high probability. This shows that Frank-Wolfe can reach any tolerance level ϵ with enough number of iterations by setting appropriate stopping criteria.

A.1.2 Proof of Property 3

Lemma 3 (Wendel [1962])**.** *If* $X_1, \ldots, X_n$ *are i.i.d. random points in* $\mathbb{R}^d$ *whose distribution is symmetric with respect to the center O and assigns measure zero to every hyperplane through O, then*

$$P_n^{(d)}(O \in \textit{Conv}\{X_1, \ldots, X_n\}) = 1 - \frac{1}{2^{n-1}} \sum_{k=0}^{d-1} \binom{n-1}{k}$$

Lemma 3 is an interesting result from stochastic geometry, which states that if we randomly sample n points in a d-dimensional ball, the probability that the convex hull formed using these points contains the center point can be computed using the above formula.

Corollary 1.
$\lim_{n \to \infty} P_n^{(d)}(O \in \textit{Conv}\{X_1, \ldots, X_n\}) = 1$

Proof: When $n \geq 2d-1$, $\binom{n-1}{k}$ is a monotonically increasing function of k for $k = 0, \ldots, d-1$. We then have

$$\sum_{k=0}^{d-1} \binom{n-1}{k} \leq d\binom{n-1}{d-1} \leq d\frac{(n-1)^d}{(d-1)!}$$

Therefore, when $n \geq 2d-1$,

$$P_n^{(d)}(O \in \text{Conv}\{X_1, \ldots, X_n\}) = 1 - \frac{1}{2^{n-1}} \sum_{k=0}^{d-1} \binom{n-1}{k} \geq 1 - \frac{d}{(d-1)!} \frac{(n-1)^d}{2^{n-1}}$$

Since $\lim_{n\to\infty} \frac{(n-1)^d}{2^{n-1}} = 0$, we get

$$\lim_{n\to\infty} P_n^{(d)}(O \in \text{Conv}\{X_1, \ldots, X_n\}) = 1$$

□

Corollary 2. *With high probability, Conv($\{\boldsymbol{q}_k\}_{1,\ldots,K}$) $\subseteq$ Conv($\mathcal{Q}$)*

The proof of Corollary 2 will be included in the proof of Theorem 1 in Appendix A.4. We will also quantify what "high probability" it is referring to. Corollary 2 establishes the fact that $\mathbf{g} = \sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k \in \text{Conv}(\{\boldsymbol{q}_k\}_{k=1,\ldots,K}) \subseteq \text{Conv}(\mathcal{Q})$ with high probability. We illustrate this idea in Figure 2 for an intuitive understanding.

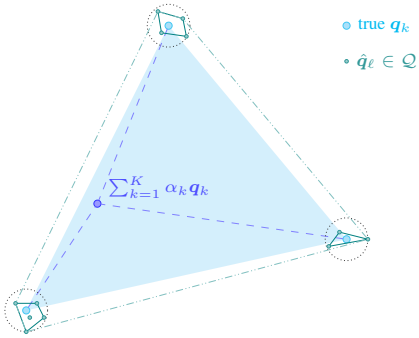


Figure 2: Constructed convex hull using logit vectors generated from subsamples

For each mixture type k, with enough data points, we have with probability $1-\delta$, $\forall \ell$ such that $\pi(\ell) = k$, $\hat{\boldsymbol{q}}_\ell$ is within an ϵ-ball of $\boldsymbol{q}_k$ given sufficient number of samples. In addition, according to Lemma 3, with high probability, such convex hull (small regions with green firm lines for each k) contains the ground truth choice probability vector $\boldsymbol{q}_k$, i.e. $\boldsymbol{q}_k \in \text{Conv}(\{\boldsymbol{q}_\ell \in \mathcal{Q} | \pi(\ell) = k\})$. Since the true aggregated choice probability $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$ is a convex combination of $\boldsymbol{q}_k$'s (so it is in the blue shaded region) and $\bigcup_k \text{Conv}(\{\boldsymbol{q}_\ell \in \mathcal{Q} | \pi(\ell) = k\}) \subset \text{Conv}(\mathcal{Q})$, we have Conv($\mathcal{Q}$) encloses the blue region and $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$ is an interior point of Conv($\mathcal{Q}$).

A.2 Proofs for Sample Complexity

We first extend the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality with the following Lemma, where both CDFs in the inequality are empirical.

Lemma 4. *Let F_n and G_m be independent empirical distribution based on m and n i.i.d. samples drown from a common cumulative distribution $F(\cdot)$. Denote $min\{m, n\}$ as T_{min}. We have*

$$\mathbb{P}\left(\sup_x |F_n(x) - G_m(x)| > \epsilon\right) \leq 4\exp\left(-\frac{1}{2}T_{\min}\epsilon^2\right)$$

Proof: Lemma 4 differs from the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality in that it investigates the tail probability of the maximum difference between *two empirical* distributions. By DKW,

we know that

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right)$$

$$\mathbb{P}\left(\sup_x |G_m(x) - F(x)| > \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right)$$

if F_n and G_m are empirical distributions of samples drawn from their true distribution function F and G respectively. We can show

$$\begin{aligned} & \mathbb{P}\left(\sup_x |F_n - F| + \sup_x |G_m - F| > \epsilon\right) \\ \leq \quad & 1 - \mathbb{P}\left(\sup_x |F_n - F| \leq \frac{\epsilon}{2} \cap \sup_x |G_m - F| \leq \frac{\epsilon}{2}\right) \\ \leq \quad & 1 - \left(1 - 2\exp\left(-\frac{1}{2}n\epsilon^2\right)\right)\left(1 - 2\exp\left(-\frac{1}{2}m\epsilon^2\right)\right) \\ \leq \quad & 4\exp\left(-\frac{1}{2}T_{\min}\epsilon^2\right) \end{aligned} \tag{3}$$

where the first inequality makes use of the fact that $\sup_x |F_n - F| + \sup_x |G_m - F| > \epsilon$ implies that either $\sup_x |F_n - F| > \frac{\epsilon}{2}$ or $\sup_x |G_m - F| > \frac{\epsilon}{2}$. The second inequality comes from the independent assumption between F_n and G_m. On the other hand, we also have

$$\begin{aligned} & \sup |F_n - F| + \sup |F - G_m| \\ \geq \quad & \sup |F_n - F| + |F - G_m| \\ \geq \quad & \sup |F_n - F + F - G_m| \\ = \quad & \sup |F_n - G_m| \end{aligned} \tag{4}$$

Combining (3) and (4), we can obtain

$$\begin{aligned} & \mathbb{P}(\sup_x |F_n(x) - G_m(x)| > \epsilon) \\ \leq \quad & \mathbb{P}(\sup_x |F_n(x) - F(x)| + \sup_x |G_m(x) - F(x)| > \epsilon) \\ \leq \quad & 4\exp\left(-\frac{1}{2}T_{\min}\epsilon^2\right). \end{aligned}$$

□

A.2.1 Proof of Theorem 2

Proof: For simpler notation, denote $s(i,j) = s$. Since i and j are of the same consumer type k, F_n and G_m are empirical distributions based on n and m samples drawn from the same distribution $\boldsymbol{q}_k$. According to Lemma 4,

$$\mathbb{P}(p_{j|i} > 1 - \epsilon) = \mathbb{P}(s < \epsilon) \geq 1 - 4\exp(-\frac{1}{2}T_{\min}\epsilon^2)$$

Let $\delta = 4\exp(-\frac{1}{2}T_{\min}\epsilon^2)$, we have $T_{\min} = \mathcal{O}(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$. □

A.2.2 Proof of Theorem 3

Proof: By definition,

$$\begin{aligned} \xi \leq \quad & \sup_x |F - G| \\ \leq \quad & \sup_x |F - F_n + F_n - G_m + G_m - G| \\ \leq \quad & \sup_x |F - F_n| + \sup_x |F_n - G_m| + \sup_x |G_m - G| \end{aligned} \tag{5}$$

Therefore,

$$
\begin{aligned}
& \mathbb{P}(\sup_x |F_n - G_m| \leq \epsilon) \qquad (6) \\
\leq\ & \mathbb{P}(\sup_x |F - F_n| + \sup_x |G_m - G| > \xi - \epsilon) \\
\leq\ & \mathbb{P}(\sup_x |F - F_n| > \frac{\xi - \epsilon}{2} \cup \sup_x |G - G_m| > \frac{\xi - \epsilon}{2}) \\
=\ & 1 - \mathbb{P}(\sup_x |F - F_n| \leq \frac{\xi - \epsilon}{2})\mathbb{P}(\sup_x |G - G_m| \leq \frac{\xi - \epsilon}{2}) \\
\leq\ & 4 \exp\left(-2T_{min}\left(\frac{\xi - \epsilon}{2}\right)^2\right)
\end{aligned}
$$

where the first inequality is based on (5) and the second inequality makes use of the fact that $\sup_x |F_n - F| + \sup_x |G_m - F| > \xi - \epsilon$ implies that either $\sup_x |F_n - F| > \frac{\xi-\epsilon}{2}$ or $\sup_x |G_m - F| > \frac{\xi-\epsilon}{2}$. The equality comes from the independent assumption between F_n and G_m. The last inequality is from the DWK inequality.

We want to restrain the sampling probability such that $p_{j|i} = 1 - s$ is within ϵ-radius of the smallest possible sampling probability, which is $1 - \xi$, i.e.

$$
\begin{aligned}
\mathbb{P}(1 - s < 1 - \xi + \epsilon) &= \mathbb{P}(s > \xi - \epsilon) \\
&= 1 - \mathbb{P}(s \leq \xi - \epsilon) \\
&\geq 1 - 4 \exp\left(-2T_{min}\left(\frac{\epsilon}{2}\right)^2\right)
\end{aligned}
$$

as a result of Equation 6.

Let $\delta = 4\exp(-\frac{1}{2}T_{min}\epsilon^2)$. Then we have $T_{min} = \mathcal{O}(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$. □

Additional discussion. We further discuss the impact of M on the sample complexity. As discussed above, we mainly relate the number of time periods we need to collect the data for, i.e. T, and ϵ and δ is through the DKW inequality. To put things simple, consider the case where we want to bound the probability that the supremum of the discrepancy between the empirical CDF and the true CDF larger than ϵ to be no bigger than δ, as follows:

$$
\mathbb{P}(\sup_{x \in [M]} |F_T(x) - F(x)| > \epsilon) \leq 2e^{-2T\epsilon^2} \overset{\text{set to}}{=} \delta
$$

This gives $T = \frac{1}{2\epsilon^2}\log\frac{2}{\delta}$. Define a random variable $U = \sup_{x\in[M]} |F_T(x) - F(x)|$. For a larger M', similarly define $V = \sup_{x\in[M']} |F_T(x) - F(x)|$. Since the set that the supremum is taken is larger for V, we have $V \geq U$. Subsequently, we have $\mathbb{P}(V > \epsilon) \geq \mathbb{P}(U > \epsilon)$.

Assume $\mathbb{P}(V > \epsilon) = \delta + \xi$ for some $\xi \geq 0$. The relationship between T and ϵ and δ is thus $T = \frac{1}{2\epsilon^2}\log\frac{2}{\delta+\xi}$ for the M' case. Now if we want to bound $\mathbb{P}(V > \epsilon)$ by δ, instead of $\delta + \xi$, we need to further increase the value of T, hence increase the sample complexity.

This analysis can be generalized directly to the two-sample DKW inequality for both in-type and out-type bounds and suggests an impact of M on the sample complexity. This agrees with the intuition that for a larger choice set, we need to collect more data since otherwise the empirical choice distribution over the set will be very sparse, making the score function computed based on the empirical CDF less credible. However, we think such impact, compared to ϵ and δ, is rather implicit as it is enclosed in the $\sup$ function.

A.3 Proof of Theorem 4

Proof: As in Algorithm 1, we use L as the number of subsamples we need to construct. Let L_k be the number of subsamples needed to hit the k-th mixture type after $k-1$ types of seeds have been selected. We have $L = L_1 + \cdots + L_K$.

We first construct a simple and fake scenario where we have K' mixture types with each mixture weight equal to α_1, i.e. $K' = \frac{1}{\alpha_1}$. Similarly, we can define L' and L'_k as above and also have

$L' = L'_1 + \cdots + L'_{K'}$. Think of L' and $L'_k, k = 1, \ldots, K'$ as random variables and we know the probability of selecting a seed from a new type k is $p_k = \frac{K'-k+1}{K'}$ since in the fake scenario, each type has the same probability α_1 of being chosen. This tells us that L'_k has a geometric distribution with expectation $\frac{1}{p_i} = \frac{K'}{K'-k+1}$.

By the linearity of expectations we have

$$\begin{aligned}
\mathbb{E}[L'] &= \mathbb{E}[L'_1 + L'_2 + \cdots + L'_{K'}] \\
&= \mathbb{E}[L'_1] + \mathbb{E}[L'_2] + \cdots + \mathbb{E}[L'_{K'}] \\
&= \frac{K'}{K'} + \frac{K'}{K'-1} + \cdots + \frac{K'}{1} \\
&= K' \cdot \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{K'}\right) \\
&= K' \cdot H_{K'}
\end{aligned}$$

where H'_K is the K'-th harmonic number. Using the asymptotics of the harmonic numbers, we get

$$\mathbb{E}[L'] \approx K' \log(K') = \frac{1}{\alpha_1} \log(\frac{1}{\alpha_1})$$

Since $\alpha_1 \leq \cdots \leq \alpha_K$, we know $K' \geq K$. On the other hand, we know the $\mathbb{E}[L_k] \leq \mathbb{E}[L'_k]$ since there is a higher probability of choosing any mixture type $k \geq 2$, due to the same reason, i.e. $\alpha_k \geq \alpha_1$. Therefore, we have $\mathbb{E}[L] \leq \frac{1}{\alpha_1} \log(\frac{1}{\alpha_1})$

We can further characterize the probability of event $\mathcal{H}$, which describes the event that all mixture types are included in the constructed set $\mathcal{Q}$ by creating L subsamples.

Claim 2. *For any $\delta > 0$, we have $\mathbb{P}(\mathcal{H}_L) \geq 1 - \delta$ with L chosen according to the criteria described below.*

Proof: Denote Z_k^L as the event that k-th mixture type is not being chosen as seed in the L trials. Similarly, we can define $Z_k^{'L}$ for the fake scenario as described above. We then have

$$\mathbb{P}(Z_k^L) = (1 - \alpha_k)^L \leq \left(1 - \frac{1}{K'}\right)^L = \mathbb{P}(Z_k^{'L}) \leq e^{-\frac{L}{K'}}$$

Denote W_k as the event that the convex hull formed by the set $\{\hat{\boldsymbol{q}}_\ell = \frac{1}{nT} \sum_{i \in I_\ell} \sum_{t=1}^T Y_i^{(t)} | \pi(i) = k\}$ for mixture type k covers the true choice probability vector. Note that each $\frac{1}{T} \sum_{t=1}^T Y_i^{(t)}$ can be viewed as a sample mean of $\boldsymbol{q}_k$ and by central limit theorem, it is symmetric with respect to $\boldsymbol{q}_k$, hence so are the $\hat{\boldsymbol{q}}_\ell$'s. According to Lemma 3 and Corollary 1,

$$\begin{aligned}
\mathbb{P}(W_k) &= 1 - \frac{1}{2^{L_k - 1}} \sum_{i=0}^{d-1} \binom{L_k - 1}{i} \\
&\geq 1 - \frac{d}{(d-1)!} \frac{(L-1)^d}{2^{L-1}} \qquad \text{when } L \geq 2d - 1
\end{aligned}$$

Putting everything together, we have

$$\mathbb{P}(\mathcal{H}) = \left(1 - \mathbb{P}(\cup_{k=1}^K Z_k^L)\right) \mathbb{P}(\cap_{k=1}^K W_k) \tag{7}$$

$$\geq \left(1 - \mathbb{P}(\cup_{k=1}^{K'} Z_k^L)\right) \left(1 - \frac{1}{2^{L-1}} \sum_{i=0}^{d-1} \binom{L-1}{i}\right)^K \tag{8}$$

$$\geq \left(1 - \frac{1}{\alpha_1} e^{-L\alpha_1}\right) \left(1 - \frac{d}{(d-1)!} \frac{(L-1)^d}{2^{L-1}}\right)^K \tag{9}$$

For any $\delta > 0$, we can then choose L such that $1 - \delta \leq$ RHS of Equation (11) and $L \geq 2d - 1$.

□

A.4 Proof of Theorem 1

We first discuss Corollary 2 and what "high probability" refers to.

Denote W_k as the event that the convex hull formed by the subsamples for a mixture type k covers the true choice probability vector, as in Appendix A.3. If we have subsampled all mixture types and for each type k, event W_k occurs, we can obtain $\boldsymbol{q}_k \in \text{Conv}(\mathcal{Q})$. Subsequently, we have $\text{Conv}(\{\boldsymbol{q}_k\}_{1,\ldots,K}) \subseteq \text{Conv}(\mathcal{Q})$.

On the other hand, we have already analyzed the probability for event that *subsampled all mixture types and Y_k occurs* $\forall k$ to occur, which is precisely $\mathcal{H}_L$ as defined above. Specifically, it happens with probability $\geq \left(1 - \frac{1}{\alpha_1} e^{-L\alpha_1}\right)\left(1 - \frac{d}{(d-1)!}\frac{(L-1)^d}{2^{L-1}}\right)^K$. As L increases, this number quickly increases to 1. This completes the proof of Corollary 2.

Finally, we combine all the results above and prove the provable convergence part in Theorem 1.

Proof: As illustrated in Figure 1, we want to show with probability $\geq 1-\delta$ we have

S.1 $|\mathbf{g}^{\texttt{SSRFW}} - \sum_k \alpha_k \boldsymbol{q}_k| \leq \epsilon$

S.2 $\forall\, k', \exists\, k = \pi(k')$ s.t. $|\hat{\boldsymbol{q}}_{k'} - \boldsymbol{q}_k| \leq \epsilon$

S.3 $\left|\sum_{k':\pi(k')=k} \alpha_{k'} - \alpha_k\right| \leq \epsilon$

First, **S.1** is proved by Lemma 2 (Property 1: $|\mathbf{g}^{\texttt{FW}} - \sum_k \alpha_k \boldsymbol{q}_k| \leq \epsilon$) and Corollary 2 (Property 3: $\sum_k \alpha_k \boldsymbol{q}_k \in \text{Conv}(\mathcal{Q})$), together with the fact that $\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \mathbf{y}^t - \sum_k \alpha_k \boldsymbol{q}_k\right| \leq \epsilon\right) > 1-\delta$ by central limit theorem. Since $\sqrt{T}(\frac{1}{T}\sum_{t=1}^T \mathbf{y}^t - \sum_k \alpha_k \boldsymbol{q}_k) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, number of samples required is also in the order of $\frac{1}{\epsilon^2}\log(\frac{1}{\delta})$.

Second, **S.2** is also a combined result by Frank-Wolfe's solving a linear program as an intermediate step (Property 2: $\forall\, k'$, $\hat{\boldsymbol{q}}_{k'} \in \mathcal{E}(\text{Conv}(\overline{\mathcal{P}}))$, $\mathcal{E}(\cdot)$ denoting the extreme point set of the input region) and by construction using Algorithm 2 (Property 4: $\forall\, \boldsymbol{q} \in \mathcal{E}(\text{Conv}(\mathcal{Q})), \exists\, k$ s.t. $\|\boldsymbol{q} - \boldsymbol{q}_k\| \leq \epsilon$).

Subsequently, we can show **S.3**. Denote K' as the number of mixtures output by the `SSRFW` algorithm. Using **S.1**, we first write

$$\left|\sum_{k'=1}^{K'} \hat{\alpha}_{k'} \hat{\boldsymbol{q}}_{k'} - \sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k\right| \leq \epsilon' \tag{10}$$

According to **S.2**, $\exists\, \pi$ such that $\pi(k') = k$ and we can write $\hat{\boldsymbol{q}}_{k'} = \boldsymbol{q}_{\pi(k')} + \boldsymbol{\epsilon}'$ where $|\boldsymbol{\epsilon}'| \leq \epsilon'$. Rearranging Eqn. (10) gives

$$\left|\sum_{k=1}^{K} \boldsymbol{q}_k \left(\sum_{k':\pi(k')=k} \hat{\alpha}_{k'} - \alpha_k\right) + \sum_{k'=1}^{K'} \hat{\alpha}_{k'} \boldsymbol{\epsilon}'\right| \leq \epsilon' \tag{11}$$

By triangle inequality, we get

$$\left|\sum_{k=1}^{K} \boldsymbol{q}_k \left(\sum_{k':\pi(k')=k} \hat{\alpha}_{k'} - \alpha_k\right)\right| - \epsilon' \leq \epsilon'$$

Since $\boldsymbol{q}_k$ is some arbitrary non-zero vector, we must have $\left(\sum_{k':\pi(k')=k} \hat{\alpha}_{k'} - \alpha_k\right) \leq 2\epsilon', \forall\, k$, which completes **S.3** by letting $\epsilon = 2\epsilon'$.

Note the above result holds assuming $K' \geq K$. To see why this is always the case, consider the linear system $Q\boldsymbol{x} = [\boldsymbol{q}_1 \boldsymbol{q}_2 \ldots \boldsymbol{q}_K][x_1, x_2, \ldots, x_K]^\top = \mathbf{g}$, where $Q \in \mathbb{R}^{M\times K}$ and $\mathbf{g} = \sum_{k=1}^K \alpha_k \boldsymbol{q}_k$. According to Definition 1, $D_{KS}(\boldsymbol{q}_k, \boldsymbol{q}_{k'}) \geq \epsilon$, we know that all $\boldsymbol{q}_k$'s are linearly independent. Since $M \geq K$, $\text{rank}(Q) = \text{rank}(Q|\mathbf{g}) = K$. The linear system has a unique solution that $\boldsymbol{x} = \boldsymbol{\alpha}$, where all x_k's are non-zero. On the other hand, we have $\mathbf{g}^{\texttt{SSRFW}} = \sum_{k'=1}^{K'} \hat{\alpha_{k'}} \hat{\boldsymbol{q}}_{k'}$ and $\|\mathbf{g}^{\texttt{SSRFW}} - \mathbf{g}\| \leq \epsilon$. Assume $K' < K$, then upto a difference of ϵ, the linear system $\hat{Q}\boldsymbol{x} = \mathbf{g}$, where $\hat{Q} = [\hat{\boldsymbol{q}}_1 \hat{\boldsymbol{q}}_2 \ldots \hat{\boldsymbol{q}}_{K'}] \in$

$\mathbb{R}^{M \times K'}$, is inconsistent. In other words, we will not be able to obtain a $\mathbf{g}^{\texttt{SSRFW}}$ that is ϵ-close to $\mathbf{g}$, making it impossible to reach the stopping condition in the SSRFW algorithm. Therefore, the algorithm will keep going for more iterations, until we have at least $K' = K$.

Finally, according to Theorem 2, Theorem 3, the sample complexity is $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$.

□

Additional discussion. In the proof, we showed that number of mixtures returned by SSRFW , K', is at least the ground truth number of mixtures, K. A natural question to ask is that how the misaligned number of mixtures affect the learning result, if $K' \neq K$. In many situations, this would not be a problem.

Consider the case that $\exists k_1, k_2$, such that $\pi(k_1) = \pi(k_2) = k$ while the rest are all one-to-one mapping. According to Theorem 1, we have $\mathbb{P}(|\boldsymbol{q}_{k_i} - \boldsymbol{q}_k| < \epsilon) \geq 1 - \delta$, for $i = 1, 2$ and $\mathbb{P}(|\alpha_{k_1} + \alpha_{k_2} - \alpha_k| < \epsilon) \geq 1 - \delta$. We can view the ground-truth MMNL model as an $(M+1)$-MNL model, where the original k-th mixture is now divided into two MNL components which share the identical logit parameters, where one of them has mixture weight α_{k_1} and the other one $\alpha_k - \alpha_{k_1}$. It is not hard to see that for the first component, we learned the correct mixture weight with an ϵ-close logit vector $\hat{\boldsymbol{q}}_{k_1}$ while for the second, the mixture weight is off by at most ϵ with an ϵ-close logit vector $\hat{\boldsymbol{q}}_{k_2}$.

Finally, we give some comment on the mapping function π. Note that we do not need this information other than using it as a tool in the proofs, though we can design heuristics to learn the mapping. In real world applications, we do not know the ground truth parameters, so we cannot derive such mapping anyways. On the other hand, if we run the algorithm multiple times, we will get different results due to the stochasticity embedded in the algorithm. In general, the lexicographic order of the mixtures is not important and can be reordered arbitrarily.

B Numerical Experiments

In this section, we demonstrate the performance of our algorithm using numerical simulation and on the Nielsen Consumer Panel Data (`https://www.chicagobooth.edu/research/kilts/datasets/nielsenIQ-nielsen`).

B.1 Simulation Study

Consider in a setting where there are N consumers (i.e. decision makers) who make purchase decisions among a set of M products for T periods. Working with synthetic data allows us to measure the model performance regarding consumer type recovery since we have the knowledge of the underlying ground truth.

Data Generation As an example experiment, we use following parameters: $N = 2000, K = 5, M = 10, d = 10$. We vary the total number of experiment epochs from $T = 5$ to 300. Mixture types are indexed by capital letters, i.e., "A", "B", "C", "D", "E", and products are indexed by numbers from 1 to 10. In the experiment, we also include an offset option, allowing the consumer to choose to not buy anything.

Features of different options and preference vector β_k's are randomly generated in the interval $[-1, 1]$. α_k's are randomly generated such that $\sum_k \alpha_k = 1$ and the minimum mixture proportion $\min_k \alpha_k \geq \frac{1}{K+3}$ to ensure that not a particular type is under-represented. We set $L = 75$.

Subsample Purity We first examine the purity of each subsample obtained. The reason we are interested in this quantity is that if a constructed subsample has high purity, it means that the empirical $\hat{\boldsymbol{q}}$ we obtained from the subsample contains only one mixture type and thus can provide an estimate for the logit vector for that type with high accuracy. Figure 3 shows the average subsample purity with respect to the total number of experiment epochs. We can see that the constructed subsamples can achieve 90% purity with as few as 30 experiment epochs, and quickly reaches 99% around $T = 150$.

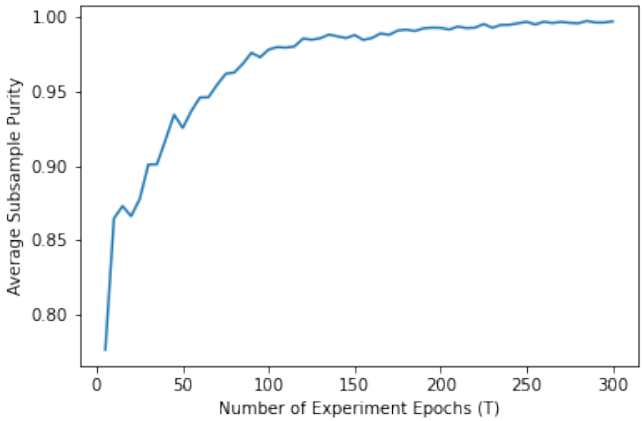
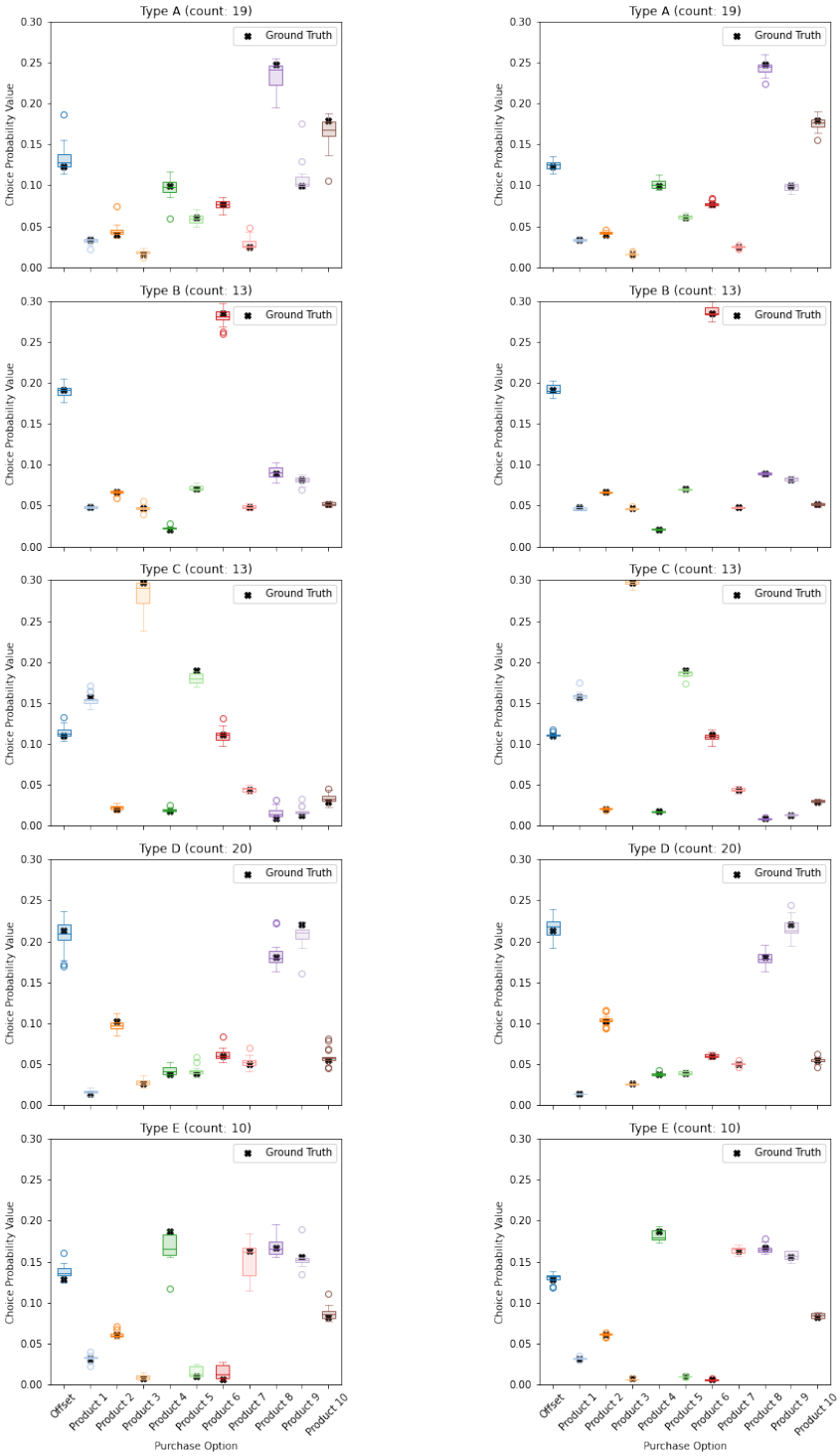


Figure 3: Subsample Purity Distribution

Quality of $\mathcal{Q}$ Next we evaluate the quality of the constructed set $\mathcal{Q}$ after applying Algorithm 2 to the simulated data. According to `SSRFW`, the estimated choice probability outcome is essentially a subset of these candidate vectors in $\mathcal{Q}$—the ones emitted by `SSRFW` at each iteration. Therefore, the higher the quality of $\mathcal{Q}$—in the sense it is concentrated near the ground truth—the better mixture estimation we can obtain using the learning algorithm `SSRFW`.

We claim each subsample is a representation of a particular mixture type in the ground truth if the majority of the selected consumers in that subsample are of that type. For instance, if a subsample contains 47 of type A, 2 of type B, and 1 of type E, we will evaluate the estimates $\hat{\boldsymbol{q}}$ from this subsample against $\hat{\boldsymbol{q}}_A$.

In Figure 4, we plot the distribution of choice probability values with respect to each product in the option set for all mixture types, with Figure 4a showing the result with 50 experiment epochs and Figure 4b with 300 experiment epochs.

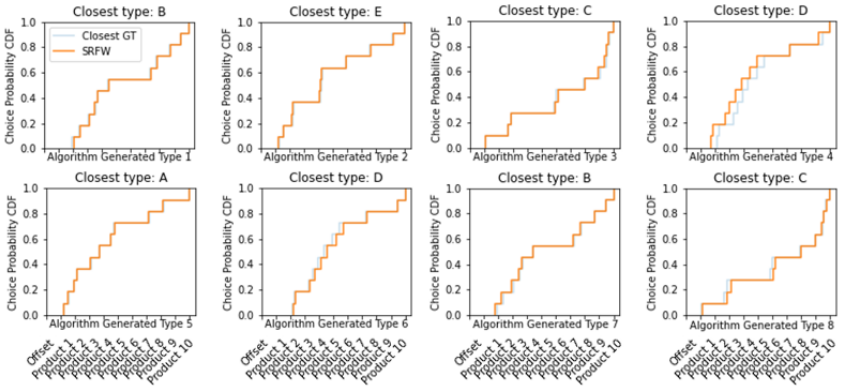


(a) $T = 50$ (b) $T = 300$

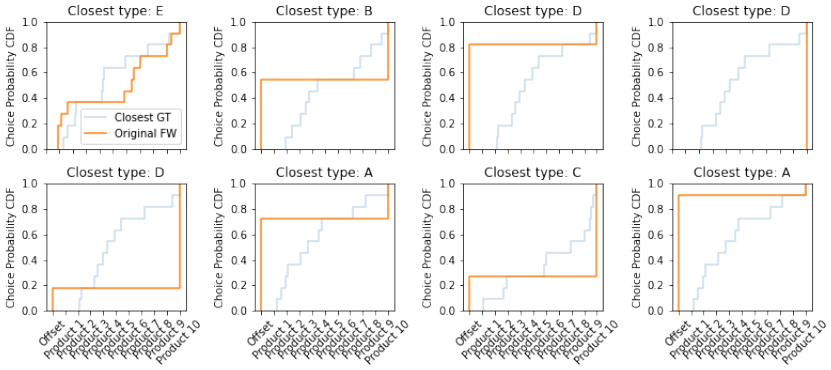
Figure 4: Choice Probability Estimation Result for All Consumer Types

These figures show that the estimated choice probability values are very concentrated near the ground truth when $T = 300$ and are also reasonably good even with a small $T = 50$, which we will proceed with for the following evaluation.

Mixture Recovery After feeding the constructed $\mathcal{Q}$ to Algorithm 1, `SSRFW` generates 8 mixture types in terms of logit vectors (as expected, larger than the ground truth number of mixtures). We now want to compare these generated logit vectors against the ground truth. We refer to the mapping $\pi(j) = i$ as the "closest type" between a generated type j and an ground truth (GT) type i again using the majority rule mentioned above.



(a) `SSRFW`



(b) Original `FW`

Figure 5: Comparison of cumulative distribution of choice probability vectors

Figure 5 compares j and $\pi(j)$ using the cumulative distribution (CDF) of the choice probability vectors, with orange lines representing learning outcome and blue lines the ground truth. We can see that the results from `SSRFW` are very close to the true mixture CDFs (Figure 5a). In contrast, the original `FW` algorithm [Jagabathula et al., 2020] does not recover the true choice probabilities (Figure 5b), which are mostly *boundary-types* that correspond to the extreme points of the feasible region to the individual LP subproblem in each FW iteration.

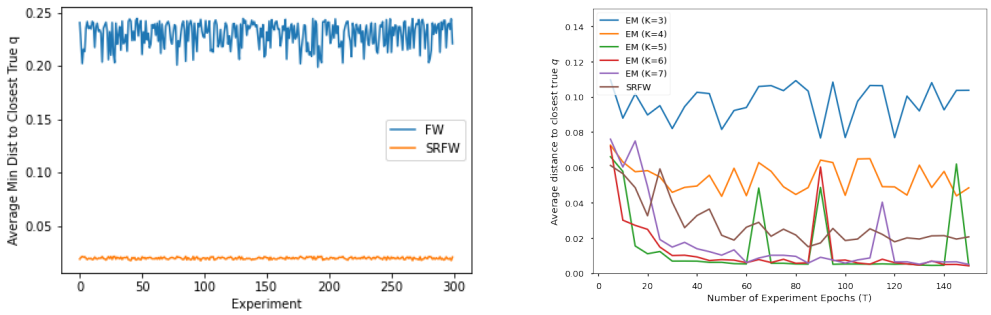
We further compare the mixture weight estimates α's in Table 1. From the table, we can see that if we have a one-to-one mapping (such as A and E), the α estimates from the algorithm are close to the true values. If we have many-to-one mappings, i.e., the algorithm outputs multiple mixtures for the same latent class, the sum of the estimated mixture weights is also close to the true values of each mixture.

A Further Comparison with Original FW and EM We define another performance measure as the weighted average distance between the predicted choice probabilities and its closest ground truth, i.e. $\| \sum_k \alpha_k \boldsymbol{q}_k - \sum_k \hat{\alpha}_k \sum_{\pi(i)=k} \hat{\boldsymbol{q}}_i^{\text{Alg}} \|$, where α_k and $\boldsymbol{q}_k$ are ground truth and Alg is one of `FW`, `SSRFW`, and EM. We vary the number of experiment epochs T to examine how additional data can help improve the performance of MMNL learning algorithms.

Table 1: Estimation of Mixture Proportion α

Type	True $\boldsymbol{\alpha}$	$\hat{\boldsymbol{\alpha}}^{\texttt{SSRFW}}$	Type-wise Sum
A	0.2000	$\hat{\alpha}_5$: 0.1904	0.1904
B	0.2364	$\hat{\alpha}_1$: 0.0713 $\hat{\alpha}_7$: 0.1607	0.2320
C	0.1636	$\hat{\alpha}_3$: 0.0741 $\hat{\alpha}_8$: 0.0757	0.1498
D	0.2182	$\hat{\alpha}_4$: 0.0387 $\hat{\alpha}_6$: 0.2143	0.2530
E	0.1818	$\hat{\alpha}_2$: 0.1748	0.1748

In Figure 6a, we can see that the total discrepancy is consistently smaller using the SSRFW compared to the original FW, which further justifies the good performance of our algorithm in predicting individual mixtures. In Figure 6b, we can see that both SSRFW and EM benefit from having more repetitive choice data, especially when $T \leq 30$, which agrees with the observations we learned from Sample Purity and Quality of $\mathcal{Q}$ part.



(a) Comparison with the original FW

(b) Comparison with EM

Figure 6: Performance comparison with other algorithms

Since EM requests the number of mixtures as an additional hyperparameter, we compare the performance of SSRFW to EM with various different hyperparameter $\tilde{K}$ values, as shown in Figure 6b. We can see from the plot that when $\tilde{K} < K$, SSRFW usually outperforms the EM algorithm. On the other hand, when $\tilde{K} >= K$, EM achieves a marginally better performance. However, in general we do not have this information, therefore we adopted the conventional strategy (c.f. Train [2008]) to use AIC/BIC to determine the best $\tilde{K}$, which corresponds to the lowest of these two criteria respectively. Both AIC/BIC measure the relative quality of statistical models for a given dataset, by balancing the trade-off between goodness of fit and the simplicity of the model and are commonly used in model selection. In our experiment, $K = 3$ gives the lowest value for both AIC and BIC. This indicates we should choose $K = 3$, yet it will result in worse estimation quality as shown in Figure 6b.

Last but not least, we can see that the major issue of EM is that it suffers from its instability. In reality, there is no way to tell whether the EM method is stuck at local optimum vs global optimum. In contrast, our algorithm is designed to give theoretical guarantees on the estimators. In summary, the combination of SSRFW's theoretical guarantees on the learning result, sample complexity analysis, as well as the capability to recover individual level of mixture parameters makes the unique contribution of this work.

One limitation in our framework is related to data collection. For each decision maker, we require a series of historical choice data over a fixed set of options in order to get a reliable empirical cumulative distribution of their choice probability vector. However, this problem is not unique to our framework and EM also requires the same type of data. Thanks to the high digitalization level of many industries,

we can see that such data are widely available in a variety of settings, ranging from training NLP models with mixture of topics of text corpus to learning consumer preferences via data that record their purchases over some product sets, which we will discuss next.

B.2 Experiments on Nielsen Consumer Panel Data

In this section, we demonstrate how we have applied `SSRFW` to the Nielsen Consumer Panel data. This comprehensive dataset is provided by the Kilts Center for Marketing at the University of Chicago Booth School of Business, NielsenIQ, and Nielsen. It contains panelists (i.e., households) purchase decisions on grocery items included in the NielsenIQ food and nonfood departments (roughly 1.4M UPC codes) dated back to 2004 with regular annual updates The panel size varies from 40K to 60K and the characteristics include product description, brand, multipack, size, etc. This panel data is widely used for longitudinal studies in marketing science.

B.2.1 Data Curation

We consider applying the algorithm to a substitute set of products under a particular category. This is a realistic setting as consumers usually choose one item from the substitution set. We curated data for six different categories, including yogurt, cereal, snack, candy, soft drinks and pet food and provide some summary statistics in Table 2.

Table 2: Nielsen case study: categories and data information

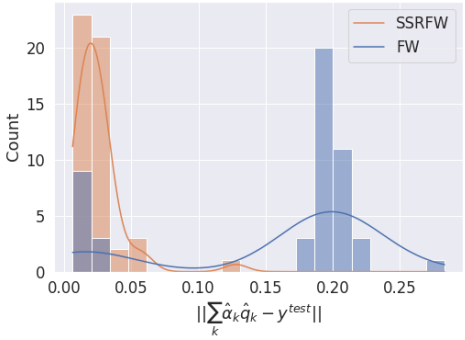
Category	Panel size	Number of features	Average # purchases
yogurt	1443	9	178
pet-food	2451	8	403
candy	1499	14	127
cereal	1085	13	96
snack	665	16	61
soft-drinks	412	12	209

B.2.2 Experiment Setup

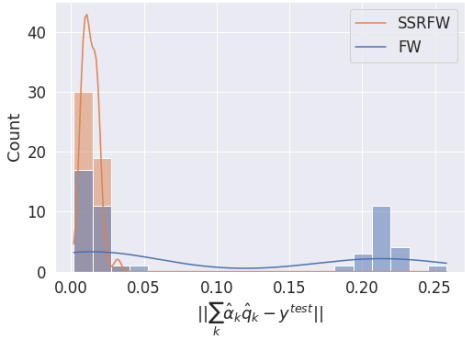
We cannot evaluate the model performance the same as in simulation studies since we no longer have the ground truth knowledge. Instead, we split the data into a training and test set, with the former used for learning the model parameters and the latter for evaluation. Specifically, we apply our algorithm to the training set, and use the learned parameters to compute the theoretical aggregated market share $\sum_k \hat{\alpha}_k \hat{\boldsymbol{q}}_k$. Then we compute its distance to the aggregated market share of the test set, i.e. $\| \sum_k \hat{\alpha}_k \hat{\boldsymbol{q}}_k - y^{\text{test}} \|$. The assumption is, if we have similar mixture composition in the training and test set, then the estimated parameter values from the training set should yield aggregated choice probability values close to that of the test set. To avoid randomness in the data split, we used a 10-fold cross validation, with the entire process repeated for five times.

B.2.3 Results

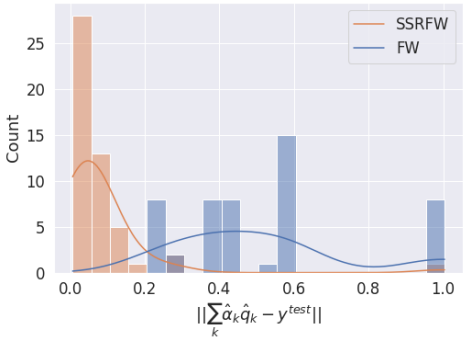
Figure 7 plots the distribution of $\| \sum_k \hat{\alpha}_k \hat{\boldsymbol{q}}_k - y^{\text{test}} \|$ from the repeated runs of both algorithms. We can see that `SSRFW` in general outperforms the original FW algorithm in that the discrepancy is distributed close 0.



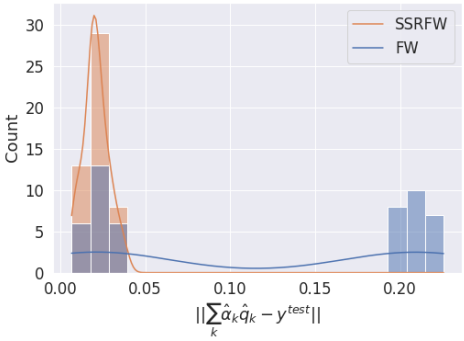
(a) Category: yogurt



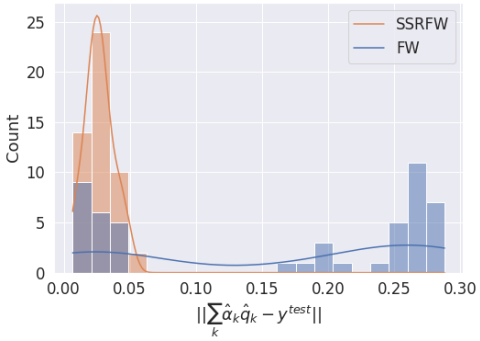
(b) Category: pet food



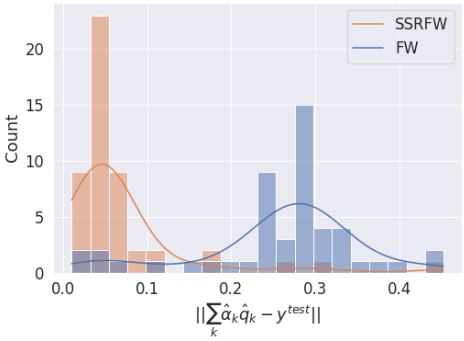
(c) Category: candy



(d) Category: cereal



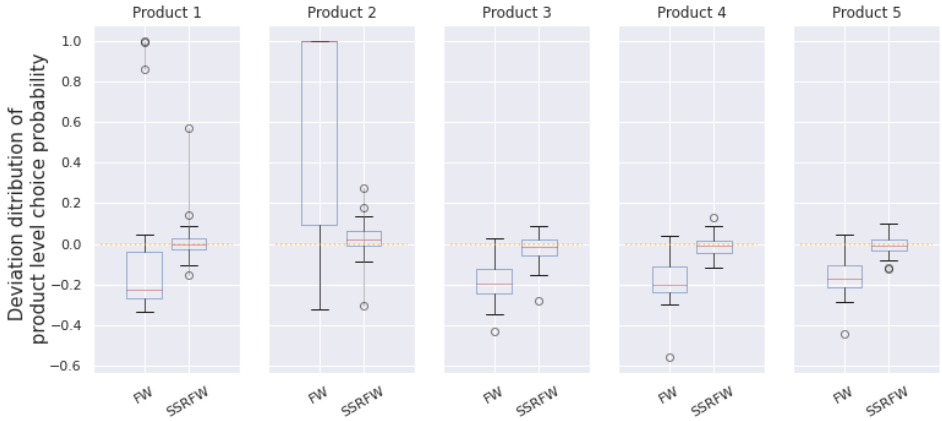
(e) Category: snack



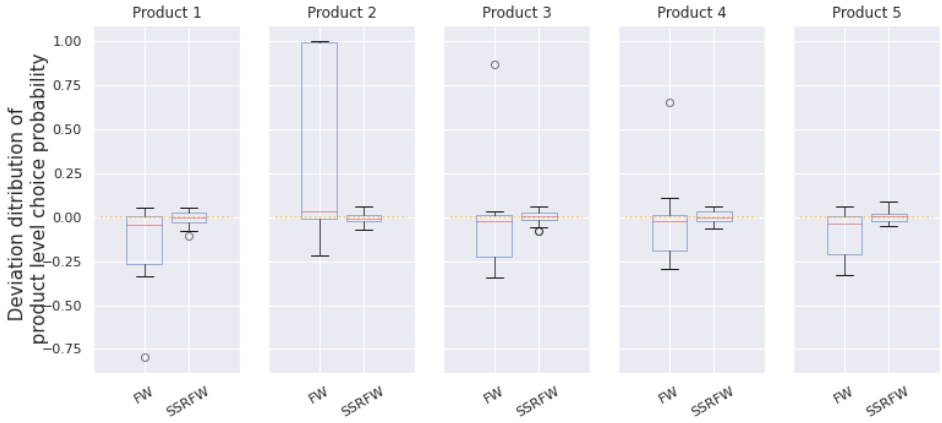
(f) Category: soft drinks

Figure 7: $\| \sum_k \hat{\alpha}_k \hat{\boldsymbol{q}}_k - y^{\text{test}} \|$ for the six product categories

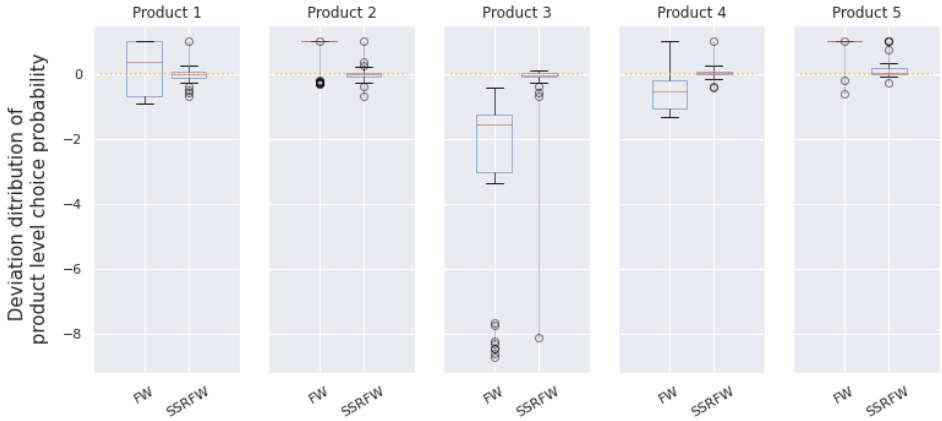
Figure 8 plots the deviation of product-level choice probability values from the test set, $\frac{|\sum_k \hat{\alpha}_k \hat{q}_{kj} - y_j^{\text{test}}|}{y_j^{\text{test}}}$, for $j = 1, 2, 3, 4, 5$. The light orange horizontal line indicates a zero deviation and we observe that the predicted aggregated choice probability per product level from `SSRFW` is more concentrated around zero than the original FW. In addition, there is a smaller variance with respect to different runs.



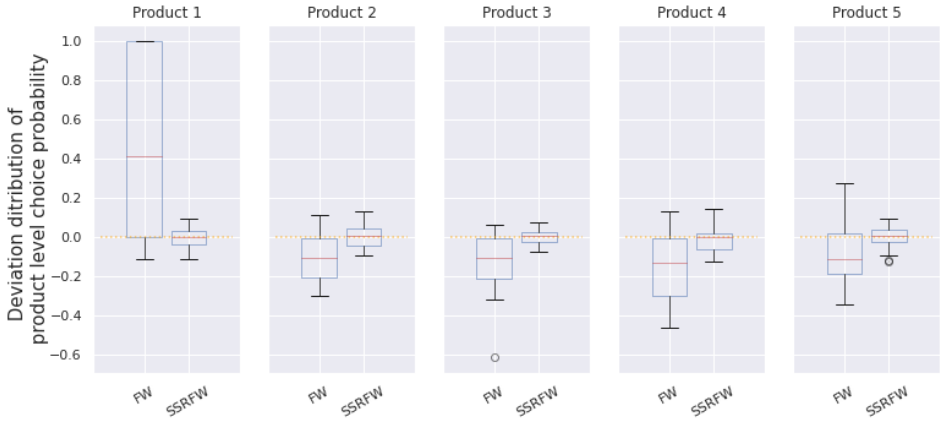
(a) Category: yogurt



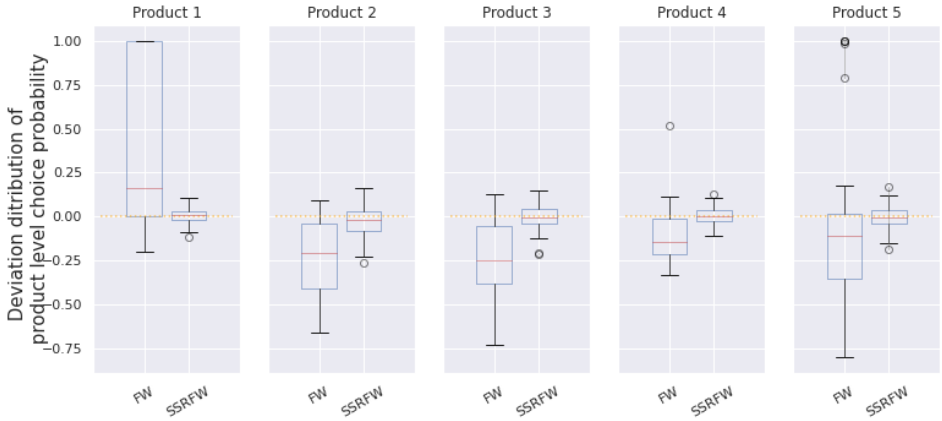
(b) Category: pet food



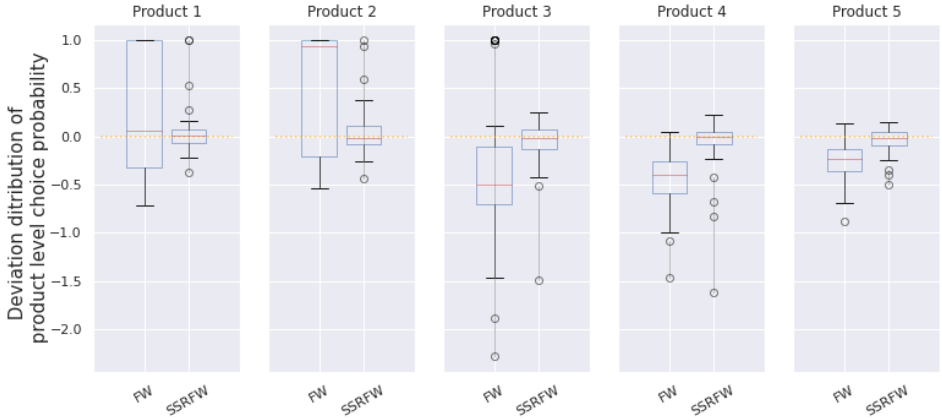
(c) Category: candy



(d) Category: cereal



(e) Category: snack



(f) Category: soft drinks

Figure 8: Deviation-from-test distribution of product-level choice probability values

Next, we examine the number of iterations required until convergence. We also report the percentage of active directions for the given number of iterations. We think these two metrics measures the effectiveness of each direction being chosen during the learning process.

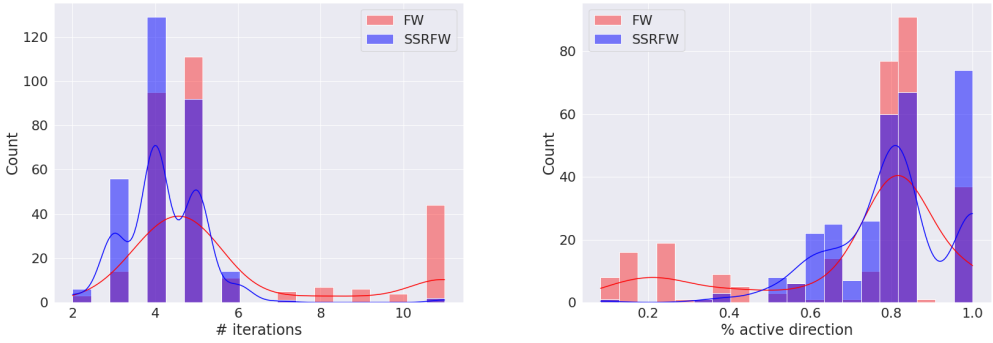
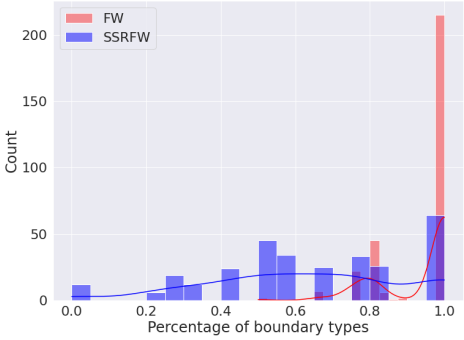


Figure 9: # iterations and % active directions

From Figure 9, we can see the FW algorithm usually requires more iterations, yet the percentage of active directions is lower than `SSRFW`. This suggests that more computation is wasted during the learning process.

The last metric we look at is the percentage of *boundary types* in the learning result. Figure 10 shows that the original FW still exhibits the same problem of generating boundary-type logit vectors while `SSRFW` is much less likely to suffer from this problem.

Figure 10: Percentage of boundary types



Finally, as requested by Kilts Center for Marketing at the University of Chicago School of Business, we make the following disclaimers:

- Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.
- The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.