Appendix A: Proofs

In order to prove the main result Theorem1, we first break down the process by proving the 4 properties mentioned in Section 4

A.1. Proof of Property 1

The proof of property 1 is relatively simple. As in our SSFRW algorithm, each step involves solving a linear optimization problem over a convex region, which is Conv($\mathcal{Q}$). We deduce from the classical convex optimization theory that the solution must be one of the extreme points of Conv($\mathcal{Q}$). As $\hat{q}_{\hat{k}}$ is the solution we get after each iteration in SSFRW algortihm, we naturally have $\hat{q}_{\hat{k}} \in \mathcal{E}(\text{Conv}(\mathcal{Q})), \forall \hat{k}$.

A.2. Proof of Property 2

Property 3 follows from the simple convergence result known for FW algorithm. Since our SSRFW algorithm inherits the same procedure as FW algorithm, the convergence result is also valid for SSRFW algorithm. We state the convergence result below as Lemma 1.

Proof of Lemma 1:

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 $\hat{k} \geq 1$, the iterates $\mathbf{x}^{(\hat{k})}$ of the fully-corrective Frank-Wolfe algorithm satisfy:

$$f(\boldsymbol{x}^{(\hat{k})}) - f(\boldsymbol{x}^{\texttt{OPT}}) \leq \frac{2 \cdot C_f}{\hat{k}+2} \tag{6}$$

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 1 follows by plugging $C_{\mathcal{L}}$ into Equation 6.

A.3. Proof of Property 3

In order to prove property 2, we will need some results about the sample complexity of our algorithm. These results will be stated and proved below.

We first state below the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality with the following Lemma, which is very useful in our proof.

LEMMA 5. *[DKW inequality (c.f. Dvoretzky et al. (1956))] Suppose $X_1, ..., X_n$ are n independent random variables with the same cumulative distribution function $F(\cdot)$. Let F_n denote the associated empirical distribution function. Then we have :*

$$\mathbb{P}(\sup_x |F(x) - F_n(x)| > \epsilon) \leq 2\exp(-2n\epsilon^2), \forall \epsilon > 0 \tag{7}$$

This is a classical inequality named after Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz, who gave a prove in 1956. And we omit the proof here since it can be found in any classical statistic textbook. The DKW inequality will be used to prove Lemma 2.

A.3.1. Proof of Lemma 2

In order to prove Lemma 2, we need the following lemma:

LEMMA 6. *Suppose $F(x)$ and $G(x)$ are CDF of two choice probability vectors $\boldsymbol{q}$ and $\boldsymbol{p}$, where $x \in [M]$. $\forall \epsilon > 0$, if $\sup_x |F(x) - G(x)| \leq \epsilon$, then $\|\boldsymbol{q} - \boldsymbol{p}\|_2 \leq \epsilon\sqrt{1 + 4(M-1)}$.*

Proof of Lemma 6:

Since $\sup_x |F(x) - G(x)| \leq \epsilon$, we have $\forall x \in [M], |F(x) - G(x)| \leq \epsilon$. If we denote the ith element of $\boldsymbol{p}$ and $\boldsymbol{q}$ as $\boldsymbol{p}(i)$ and $\boldsymbol{q}(i)$, respectively, then $F(x) = \sum_{i=1}^{x} \boldsymbol{q}(i), \forall x \in [M]$ and $G(x) = \sum_{i=1}^{x} \boldsymbol{p}(i), \forall x \in [M]$. Thus $|F(1) - G(1)| = |\boldsymbol{q}(1) - \boldsymbol{p}(1)| \leq \epsilon$. And $|F(2) - G(2)| = |\boldsymbol{q}(1) + \boldsymbol{q}(2) - \boldsymbol{p}(1) - \boldsymbol{p}(2)| \leq \epsilon$. Using the triangle inequality, we have $|\boldsymbol{q}(2) - \boldsymbol{p}(2)| - |\boldsymbol{q}(1) - \boldsymbol{q}(2)| \leq |\boldsymbol{q}(1) + \boldsymbol{q}(2) - \boldsymbol{p}(1) - \boldsymbol{p}(2)|$, so $|\boldsymbol{q}(2) - \boldsymbol{p}(2)| \leq 2\epsilon$. Similarly, we can prove that for any $i > 1$, we have $|\boldsymbol{q}(i) - \boldsymbol{p}(i)| \leq 2\epsilon$ by applying the triangle inequality $|\boldsymbol{q}(i) - \boldsymbol{p}(i)| - |F(i-1) - G(i-1)| \leq |\boldsymbol{q}(i) + F(i-1) - \boldsymbol{p}(i) - G(i-1)| = |F(i) - G(i)|$. So $\|\boldsymbol{q} - \boldsymbol{p}\|_2^2 = \sum_{i=1}^{M} (\boldsymbol{q}(i) - \boldsymbol{p}(i))^2 \leq (\epsilon^2 + (M-1)4\epsilon^2)$. So we have $\|\boldsymbol{q} - \boldsymbol{p}\|_2 \leq \epsilon\sqrt{1 + 4(M-1)}$.

Q.E.D.

Proof :

Let us denote i_0 as the seed of subsample $\boldsymbol{I}_\ell$, and $\boldsymbol{q}^{i_0} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_{i_0}^t$ as the empirical choice vector of i_0. According to the Q construction algorithm, we have $s(i, i_0) \leq \epsilon'$, for every $i \in \boldsymbol{I}_\ell$, where ϵ' is the precision parameter in the algorithm. Now let's just set $\epsilon' = \frac{\epsilon}{\sqrt{1+4(M-1)}}$, we have $s(i, i_0) \leq \frac{\epsilon}{\sqrt{1+4(M-1)}}$, for every $i \in \boldsymbol{I}_\ell$, which implies $\sup_x |F_T(x; i) - F_T(x; i_0)| \leq \frac{\epsilon}{\sqrt{1+4(M-1)}}$ based on the definition of $s(i, i_0)$. Next, by applying lemma 6, we have $\|\boldsymbol{q}^i - \boldsymbol{q}^{i_0}\| \leq \epsilon, \forall i \in \boldsymbol{I}_\ell$. Hence, we have

$$\left\|\bar{\boldsymbol{q}}_\ell - \boldsymbol{q}^{i_0}\right\|_2 = \left\|\frac{1}{n}\sum_{i \in \boldsymbol{I}_\ell} \boldsymbol{q}^i - \boldsymbol{q}^{i_0}\right\|_2 \leq \frac{1}{n}\sum_{i \in \boldsymbol{I}_\ell} \left\|\boldsymbol{q}^i - \boldsymbol{q}^{i_0}\right\|_2 \leq \epsilon, \tag{8}$$

where $\bar{\boldsymbol{q}}_\ell = \frac{1}{n}\sum_{i \in \boldsymbol{I}_\ell} \boldsymbol{q}^i$.

On the other hand, according to the DKW inequality (7), we have $\sup_x |F_T(x; i_0) - F(x)| \leq \frac{\epsilon}{\sqrt{1+4(M-1)}}$ with a probability of at least $1 - \delta$ if $T = O(\frac{M}{\epsilon^2} log(\frac{1}{\delta}))$, where $F(x)$ is the CDF corresponding to the ground truth type of i_0. We denote this ground truth type as $\pi(\ell)$. Then by lemma 6, with a probability of at least $1 - \delta$, we have $\left\|\boldsymbol{q}^{i_0} - \boldsymbol{q}_{\pi(\ell)}\right\|_2 \leq \epsilon$. Combing the analysis above and applying the triangle inequality, we have

$$\left\|\bar{\boldsymbol{q}}_\ell - \boldsymbol{q}_{\pi(\ell)}\right\|_2 \leq \left\|\bar{\boldsymbol{q}}_\ell - \boldsymbol{q}^{i_0}\right\|_2 + \left\|\boldsymbol{q}^{i_0} - \boldsymbol{q}_{\pi(\ell)}\right\|_2 \leq 2\epsilon, \tag{9}$$

which completes the proof.

Q.E.D.

A.4. Proof of Property 4

To prove Property 4, we first need to prove Theorem 2.

Proof of Theorem 2:

Without loss of generality, we assume $\alpha_1 \le \alpha_2 \le ... \le \alpha_K$. Denote Z_k^L as the event that none of the generated L sample seeds is from the k-th mixture type. Similarly, $\tilde{Z}_k^L$ represents the same event but assuming that the true mixture type is from a special distribution that includes $\tilde{K}$ mixture types, each with a mixture weight equals to α_1, where $\tilde{K} = \frac{1}{\alpha_1}$. We have the following inequality

$$\mathbb{P}(Z_k^L) = (1-\alpha_k)^L \le (1-\alpha_1)^L = \mathbb{P}(\tilde{Z}_k^L) = (1-\frac{1}{\tilde{K}})^L \le e^{-\frac{L}{\tilde{K}}} \tag{10}$$

The first inequality in (10) is because α_1 is the smallest mixture weight, i.e., $\alpha_1 \le \alpha_k$ for any $k = 1, ..., K$. The third equality follows from $\tilde{K} = \frac{1}{\alpha_1}$, which implies $\alpha_1 = \frac{1}{\tilde{K}}$. To get the last inequality, we note that $\lim_{n\to\infty}(1-\frac{1}{n})^n = \frac{1}{e}$ and the sequence $\{(1-\frac{1}{n})^n\}_{n=1}^{\infty}$ is an increasing sequence. The monotonicity of the sequence can be verified using the Arithmetic Mean Geometric Mean Inequality (AM-GM inequality) which states $\frac{x1+x2+\cdots+x_n}{n} \ge \sqrt[n]{x_1 x_2 \cdots x_n}$ for any numbers $x_1, x_2, \cdots x_n \ge 0$. Specifically, we have $(1-\frac{1}{n})^n = 1 \cdot \underbrace{(1-\frac{1}{n}) \cdot (1-\frac{1}{n}) \cdots (1-\frac{1}{n})}_{n} \le [\frac{1+(1-\frac{1}{n})+(1-\frac{1}{n})+\cdots+(1-\frac{1}{n})}{n+1}]^{n+1} = [\frac{1+n(1-\frac{1}{n})}{n+1}]^{n+1} = (\frac{n}{n+1})^{n+1} = (1-\frac{1}{n+1})^{n+1}$, which shows $\{(1-\frac{1}{n})^n\}_{n=1}^{\infty}$ is an increasing sequence. Therefore we have $(1-\frac{1}{\tilde{K}})^{\tilde{K}} \le \frac{1}{e}$. Then $(1-\frac{1}{\tilde{K}})^L = ((1-\frac{1}{\tilde{K}})^{\tilde{K}})^{\frac{L}{\tilde{K}}} \le (\frac{1}{e})^{\frac{L}{\tilde{K}}} = e^{-\frac{L}{\tilde{K}}}$.

On the other hand, we note that $\mathcal{H}_L = \cap_{k=1}^K (Z_k^L)^C$. $(Z_k^L)^C$ is the complement of Z_k^L, which means that at least one of the generated L sample seeds is from the k-th mixture type. Hence,

$$\mathbb{P}(\mathcal{H}_L) = \mathbb{P}(\cap_{k=1}^K (Z_k^L)^C) = \mathbb{P}((\cup_{k=1}^K Z_k^L)^C) \tag{11}$$

$$= 1 - \mathbb{P}(\cup_{k=1}^K Z_k^L) \tag{12}$$

$$\ge 1 - \sum_{k=1}^K \mathbb{P}(Z_k^L) \tag{13}$$

$$\ge 1 - \sum_{k=1}^K \mathbb{P}(\tilde{Z}_k^L) \tag{14}$$

$$\ge 1 - K \cdot e^{-L\alpha_1} \tag{15}$$

$$\ge 1 - \tilde{K} \cdot e^{-L\alpha_1} \tag{16}$$

$$= 1 - \frac{1}{\alpha_1} e^{-L\alpha_1} \tag{17}$$

The first inequality follows from the fact that $\mathbb{P}(\cup_{k=1}^K Z_k^L) \le \sum_{k=1}^K \mathbb{P}(Z_k^L)$, and the second inequality follows from (10) which shows $\mathbb{P}(Z_k^L) \le \mathbb{P}(\tilde{Z}_k^L)$. The third inequality also comes from (10) along with the fact that $\tilde{K} = \frac{1}{\alpha_1}$. The fourth inequality is due to $\tilde{K} \ge K$, which can be verified by contradiction. Specifically, suppose $K > \tilde{K}$, we have $\sum_{k=1}^K \alpha_k \ge \sum_{k=1}^K \alpha_1 = K\alpha_1 > \tilde{K}\alpha_1 = 1.$, which contradicts with the fact that $\sum_{k=1}^K \alpha_k = 1$. From the above, we can see that for any $\delta > 0$, we can choose $L \ge \frac{1}{\alpha_1} log(\frac{1}{\alpha_1 \delta})$, so that $\mathbb{P}(\mathcal{H}_L) \ge 1 - \delta$. Q.E.D.

Theorem 2 shows that with sufficient large L, we have $\mathbb{P}(\mathcal{H}_L)$ is close to 1, which guarantees that all mixture types can be generated from our $\mathcal{Q}$ construction with a high probability when L is large. To analyze the probability for $\sum_{k=1}^K \alpha_k \boldsymbol{q}_k$ to locate in $\text{Conv}(\mathcal{Q})$, we first define a smaller set.

DEFINITION 4. Define Θ as the smallest convex hull formed by K selected $\bar{\boldsymbol{q}}_l$, with each $\bar{\boldsymbol{q}}_l \in \mathcal{Q}$ and its corresponding seed represents a ground truth type.

Theorem 2 shows that with sufficient large L, we have the seeds generated from our $\mathcal{Q}$ construction represents all the mixture types and Θ is formed as the smallest convex hull by a subset of $\mathcal{Q}$, with each element's seed representing each ground truth type. Therefore, it is clear that Θ is a subset of $\text{Conv}(\mathcal{Q})$.

PROPOSITION 3. $\Theta \subset Conv(\mathcal{Q})$.

According to Proposition 3, the probability that $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$ lies in $\text{Conv}(\mathcal{Q})$ is lower bounded by that having $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k$ in Θ. It motivates us to analyze the probability of $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k \in \Theta$ in order to bound the probability of $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k \in \text{Conv}(\mathcal{Q})$.

Consider a set of seeds denoted by $\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_{\hat{K}}$. The empirical choice probability of each subsample falls within an ϵ range of its seed's ground-true mixture type with a high probability $1-\delta$, for $0<\delta<1$, when $T = O(\frac{M}{\epsilon^2} log(\frac{1}{\delta}))$ according to Lemma2.

Proof of Lemma 3: Conditioned on each extreme point of Θ is within an ϵ distance of the ground truth, the probability of $\sum_{k=1}^{K} \alpha_k \boldsymbol{q}_k \in \Theta$ is lower bounded by the volume ratio between Θ and $\Theta + \epsilon\mathcal{B}$, where $\mathcal{B}$ denotes a unit ball. Here the notation "+" means Minkowski sum for point sets, where $A+B = \{a+b | a \in A, b \in B\}$. $\Theta + \epsilon\mathcal{B}$ is obtained by drawing a circle radius ϵ centered on every point in Θ. Figure 11 is a illustration of the relationship between $\Theta + \epsilon\mathcal{B}$ and $\text{Conv}(q_1, .., q_K)$. In the figure, the inner triangle represents our Θ and the outer triangle with the blue dotted line is $\text{Conv}(q_1, .., q_K)$. Note that each ground truth q_k is within a distance of ϵ from each corresponding extreme point in Θ with a high probability according to the construction of Θ and Lemma 2. Hence we have $\text{Conv}(q_1, .., q_K)$ is contained in $\Theta + \epsilon\mathcal{B}$ with a high probability. Thus $\frac{V(\Theta)}{V(\text{Conv}(q_1,\ldots,q_K))} \geq \frac{V(\Theta)}{V(\Theta+\epsilon\mathcal{B})}$, where $V(\cdot)$ denotes the volume of a set. Here we define the Lebesgue integration of constant 1 over the set as the volume of this set. This inequality can be easily seen in Figure 11. In the figure, we are using the area enclosed by black circles and an inner triangle as the area of $\Theta + \epsilon\mathcal{B}$. Note that the figure is only for illustration. In fact $\Theta + \epsilon\mathcal{B}$ should be obtained continuously by drawing ϵ ball on every point in $\mathcal{B}$, which means the real area of $\Theta + \epsilon\mathcal{B}$ is the area surrounded by the black dashed lines and circles. Blue dots in Figure 11 represent the ground truth $q_1, ..., q_K$, which are within an ϵ distance from each extreme point in Θ (black dots) with a high probability according to Lemma 2. So it is clear that $\text{Conv}(q_1, ..., q_K)$, the area formed by blue triangle, is contained inside $\Theta + \epsilon\mathcal{B}$, thus $V(\text{Conv}(q_1, ..., q_K))$ is smaller than $V(\Theta + \epsilon\mathcal{B})$. We also remind the readers that as in figure 11, Θ is always contained in blue triangle $\text{Conv}(q_1, .., q_K)$, this is because we are defining our Θ as the worst case scenario. Then it suffices to study the behavior of $V(\Theta + \epsilon\mathcal{B})$. In fact, we have the following relation:

$$V(\Theta + \epsilon\mathcal{B}) \leq V(\Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta) = V((\frac{\epsilon_0}{\epsilon_0 - \epsilon})\Theta), \tag{18}$$

where ϵ_0 denotes the radius of the largest ball that can be contained in $\text{Conv}(q_1, ...q_K)$, i.e., $\epsilon_0 = \max\{\tilde{\epsilon} : \tilde{\epsilon}\mathcal{B} \subset \text{Conv}(q_1, ...q_K)\}$. This inequality tells us that we can somehow upper bound the volume of $\Theta + \epsilon\mathcal{B}$ by the volume of $(\frac{\epsilon_0}{\epsilon_0 - \epsilon})\Theta$, which makes it easier for us to obtain the bound of ratio between $V(\Theta)$ and $V(\Theta + \epsilon\mathcal{B})$.

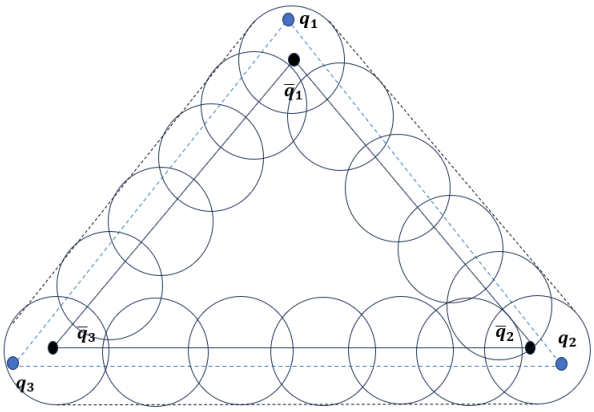

Figure 11 $\Theta + \epsilon\mathcal{B}$ **and Conv**$(q_1, ..., q_K)$

To prove relation (18), we need to first prove that $(\epsilon_0 - \epsilon)\mathcal{B} \subset \Theta$. We will illustrate it using triangles for simplicity, any higher dimension follows a similar argument. We first draw some auxiliary lines, as in Figure 12, the outer blue triangle is our $\text{Conv}(q_1, .., q_K)$, we draw three auxiliary lines in black dashed lines with each parallel to one of the edges of $\text{Conv}(q_1, ..., q_K)$, the distance between the auxiliary line and the edge is ϵ. In figure 12, the black dashed lines are auxiliary lines, while the black triangle is our Θ. Some facts can be deduced from the figure. If we denote the triangle formed by three auxiliary lines as Ω, we can see first that Ω is always contained in Θ. This is because the distance between any q and $\hat{q}$ is smaller than ϵ, thus the distances between the corresponding edges of Θ and $\text{Conv}(q_1, ..., q_K)$ should also be smaller than ϵ, making Ω contained in Θ. Secondly, triangle Ω and $\text{Conv}(q_1, ..., q_K)$ are similar to each other because their edges are parallel. So their inner circle shares the same center, and we denote it by $\mathcal{O}$. In figure 12, we draw line $\mathcal{O}b$ perpendicular to the line connecting q_2 to q_3, then this line is also perpendicular to the auxiliary line parallel to the line connecting q_2 to q_3, we denote the foot point by a. Since ϵ_0 is the radius of inner circle of $\text{Conv}(q_1, .., q_K)$, which is represented by a blue circle in Figure 12 we have $\mathcal{O}b = \epsilon_0$. Moreover, according to the construction from the auxiliary lines, the distance between auxiliary line and the corresponding edge of $\text{Conv}(q_1, ..., q_K)$, which is the length of the line connecting a to b is ϵ. Thus we can imply that the radius of the inner circle of Ω is $\epsilon_0 - \epsilon$ and since $\Omega \in \Theta$, we can conclude that $(\epsilon_0 - \epsilon)\mathcal{B} \subset \Theta$.

After noting that $(\epsilon_0 - \epsilon)\mathcal{B} \subset \Theta$, then for any $x \in \Theta + \epsilon B$, we can write x as $x = y + z$, where $y \in \Theta$ and $z \in \epsilon\mathcal{B}$. So $\frac{\epsilon_0 - \epsilon}{\epsilon} z \in (\epsilon_0 - \epsilon)\mathcal{B} \subset \Theta$, we have $z \in \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$. Thus $x = y + z \in \Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$, for any $x \in \Theta + \epsilon\mathcal{B}$, which means $\Theta + \epsilon\mathcal{B} \subset \Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$, $V(\Theta + \epsilon\mathcal{B}) \leq V(\Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta)$.

For the equality in the relation above, we need to use the convexity of Θ. For one side, assume that $\theta \in (1 + \frac{\epsilon}{\epsilon_0 - \epsilon})\Theta$. We can write θ as $\theta = \frac{\epsilon_0 - \epsilon}{\epsilon_0}\theta + \frac{\epsilon}{\epsilon_0}\theta$. And since $\theta \in (1 + \frac{\epsilon}{\epsilon_0 - \epsilon})\Theta$, we have $\frac{\epsilon_0 - \epsilon}{\epsilon_0}\theta = \frac{1}{1 + \frac{\epsilon}{\epsilon_0 - \epsilon}}\theta \in \Theta$, while $\frac{\epsilon}{\epsilon_0}\theta = \frac{\epsilon}{\epsilon_0 - \epsilon}\frac{1}{1 + \frac{\epsilon}{\epsilon_0 - \epsilon}}\theta \in \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$. Thus $\theta \in \Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$, $(1 + \frac{\epsilon}{\epsilon_0 - \epsilon})\Theta \subset \Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$.

On the other hand, suppose we have $x \in \Theta + \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$, we can write x as $x = y + z$, where $y \in \Theta$, $z \in \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$. Since $z \in \frac{\epsilon}{\epsilon_0 - \epsilon}\Theta$, we can further write $z = \frac{\epsilon}{\epsilon_0 - \epsilon}z'$, for some $z' \in \Theta$. Now we can see that $\frac{1}{1 + \frac{\epsilon}{\epsilon_0 - \epsilon}}x = \frac{\epsilon_0 - \epsilon}{\epsilon_0}x = \frac{\epsilon_0 - \epsilon}{\epsilon_0}y + \frac{\epsilon_0 - \epsilon}{\epsilon_0}z = \frac{\epsilon_0 - \epsilon}{\epsilon_0}y + \frac{\epsilon}{\epsilon_0}z'$, since both $y, z' \in \Theta$, $\frac{\epsilon}{\epsilon_0} + \frac{\epsilon_0 - \epsilon}{\epsilon_0} = 1$, $\frac{\epsilon_0 - \epsilon}{\epsilon_0}x$ is simply a convex combination

of y and z'. As Θ is a convex set, we deduce $\frac{\epsilon_0-\epsilon}{\epsilon_0}x = \frac{1}{1+\frac{\epsilon}{\epsilon_0-\epsilon}}x \in \Theta$, which means $x \in (1+\frac{\epsilon}{\epsilon_0-\epsilon})\Theta$, thus $\Theta + \frac{\epsilon}{\epsilon_0-\epsilon}\Theta \subset (1+\frac{\epsilon}{\epsilon_0-\epsilon})\Theta$.

Combining both sides, we have $\Theta + \frac{\epsilon'}{\epsilon_0}\Theta = (1+\frac{\epsilon'}{\epsilon_0})\Theta$, thus $V(\Theta + \frac{\epsilon'}{\epsilon_0}\Theta) = V((1+\frac{\epsilon'}{\epsilon_0})\Theta)$.

With this relation, we can further deduce the lower bound of our ratio, $\frac{V(\Theta)}{V(\text{Conv}(q_1,...,q_K))} \geq \frac{V(\Theta)}{V(\Theta+\epsilon\mathcal{B})} \geq \frac{V(\Theta)}{V(\Theta+\frac{\epsilon}{\epsilon_0-\epsilon}\Theta)} = \frac{V(\Theta)}{V((1+\frac{\epsilon}{\epsilon_0-\epsilon})\Theta)} = \frac{V(\Theta)}{V(\frac{\epsilon_0}{\epsilon_0-\epsilon}\Theta)} = (1-\frac{\epsilon}{\epsilon_0})^M$, The last inequality can be seen as a result of multivariate integration, which we further elaborate below. For simplicity, we denote λ as $1+\frac{\epsilon}{\epsilon_0-\epsilon}$ in the sequel.

Since by volume of a convex set, we mean the Lebesgue integration of constant 1 over the convex set. and all convex sets we discuss here are close and bounded, the Lebesgue integration on them can be well defined. So $V(\Theta) = \int_\Theta 1\, d\boldsymbol{q} = \int_\Theta 1\, dq_1 \dots dq_M$.. On the other hand, $V(\lambda\Theta) = \int_{\lambda\Theta} 1\, d\boldsymbol{q}' = \int_{\lambda\Theta} 1\, dq_1' \dots dq_M'$. Next we perform a change of variable, we substitute $\boldsymbol{q}'$ by $\lambda\boldsymbol{q}$. Since $\boldsymbol{q}' \in \lambda\Theta$, $\boldsymbol{q} \in \Theta$. Then the second integral becomes $V(\lambda\Theta) = \int_\Theta 1 d\lambda\boldsymbol{q} = \int_\Theta 1\, d\lambda q_1 \dots d\lambda q_M = \lambda^M \int_\Theta 1 dq_1 \dots dq_M = \lambda^M V(\Theta)$. Thus $\frac{V(\Theta)}{V(\lambda\Theta)} = \frac{V(\Theta)}{\lambda^M V(\Theta)} = \frac{1}{\lambda^M}$.
Multiply with the probability such that all $\hat{q}_k$ are within the ϵ range of the true mixture type, we have the finally concluded that the probability for $\sum_{k=1}^K \alpha_k \boldsymbol{q}_k \in \text{Conv}(q_1, ..., q_K)$ is lower bounded by $(1-\frac{\epsilon}{\epsilon_0})^M(1-\delta)^K$, completing the proof. Q.E.D..

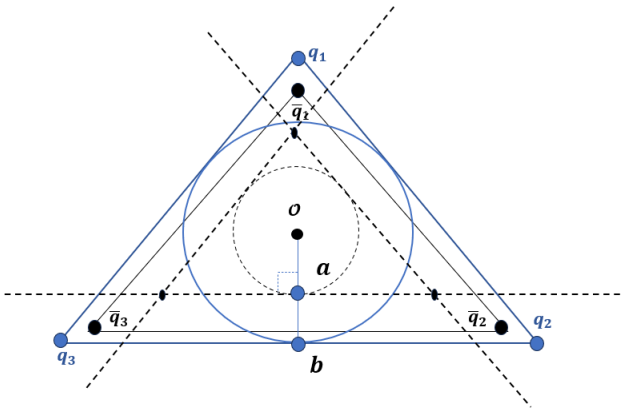

Figure 12 Ω **and** Θ

A.5. Proof of Lemma 4

Since Lemma 4 is closely related to the main theorem, we will first spend this subsection discussing about the proof of Lemma 4. With the no-purchase option, we have $\sum_{i=1}^M g_i \leq 1$. Hence we can regard $\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_j^t - g_j \leq \frac{\epsilon}{\sqrt{M}}, \forall i$ are mutually independent. On the other hand, we note that $\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}^t - \sum_{k=1}^K \alpha_k \boldsymbol{q}_k\|_2 \leq \epsilon$ if for every product j, $\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_j^t - g_j \leq \frac{\epsilon}{\sqrt{M}}$. Therefore, suppose the probability for $\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_j^t - g_j \leq \frac{\epsilon}{\sqrt{M}}$ to hold is $1-\delta'$ for every product j, then we have $\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}^t - \sum_{k=1}^K \alpha_k \boldsymbol{q}_k\|_2 \leq \epsilon$ holds with a probability of $(1-\delta')^M$. When δ' is very small, this probability can be approximated as $1-M\delta'$.

The following efforts are to bound the probability for $\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_j^t - g_j \leq \frac{\epsilon}{\sqrt{M}}$ to hold. According to Assumption 1, $\sqrt{T}(\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_j^t - g_j) \sim \mathcal{N}(0, \sigma^2)$ for every j. By the concentration theorem for normal random variables, we have:

$$\mathbb{P}(|\sqrt{T}(\frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_j^t - g_j)| \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2\sigma^2}} \tag{19}$$

Therefore, $\mathbb{P}(|(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_j^t - g_j)| \geq \epsilon) = \mathbb{P}(|\sqrt{T}(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_j^t - g_j)| \geq \sqrt{T}\epsilon) \leq e^{-\frac{T\epsilon^2}{2\sigma^2}}$. Let $e^{-\frac{T\epsilon^2}{2\sigma^2}}$ equal $\delta^{'}$, then we have $T = O(\frac{1}{\epsilon^2}log(\frac{1}{\delta'}))$. In other words, with $T = O(\frac{1}{\epsilon^2}log(\frac{1}{\delta'}))$ many samples, we have $|(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_j^t - g_j)| \geq \epsilon$ holds with a probability of $\delta^{'}$. By replacing ϵ by $\frac{\epsilon}{\sqrt{M}}$ in the above analysis, we have with $T = O(\frac{M}{\epsilon^2}log(\frac{1}{\delta'}))$ many samples, $\mathbb{P}(|(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_j^t - g_j)| \geq \frac{\epsilon}{\sqrt{M}}) \leq \delta'$, which implies

$$\mathbb{P}(|(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_j^t - g_j)| \leq \frac{\epsilon}{\sqrt{M}}) \geq 1-\delta'.$$

In summary, with $T = O(\frac{M}{\epsilon^2}log(\frac{1}{\delta'}))$ many samples, $\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq \epsilon$ holds with a probability of $1-M\delta'$. Setting $\delta = M\delta^{'}$, we have $\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq \epsilon$ happens with probability at least $1-\delta$ if the sample size $T = O(\frac{M}{\epsilon^2}log(\frac{M}{\delta}))$

Note that for any $\epsilon \geq 0$, Lemma 3 and Theorem 2 show that if we choose an appropriate number of T and subsample L, we can have $\sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k \in \mathrm{Conv}(\mathcal{Q})$ holds with a high probability. Therefore $\mathcal{L}(\mathbf{g}^{\texttt{OPT}}) \leq \mathcal{L}(\sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k)$ by definition of optimality for $\mathbf{g}^{\texttt{OPT}}$. In addition, we have $\sqrt{2\mathcal{L}(\sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k)} = \|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq \epsilon$ with a probability of at least $1-\delta$ according to analysis above. Thus we have $\mathcal{L}(\mathbf{g}^{\texttt{OPT}}) \leq \mathcal{L}(\sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k) \leq \frac{\epsilon^2}{2}$. Therefore,

$$\begin{aligned}
&\|\mathbf{g}^{\texttt{OPT}} - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \\
&\leq \|\mathbf{g}^{\texttt{OPT}} - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t\|_2 + \|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \qquad (20)\\
&= \sqrt{2\mathcal{L}(\mathbf{g}^{\texttt{OPT}})} + \sqrt{2\mathcal{L}(\sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k)} \leq 2\epsilon,
\end{aligned}$$

where the first inequality comes from the triangle inequality. (20) shows that for any $\epsilon \geq 0$, we can achieve $\|\mathbf{g}^{\texttt{OPT}} - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq \epsilon$ with a high probability given that we have adequate T and L.

Now Lemma 4 is a straightforward result by combining Lemma 1 and (20). Specifically, Lemma 1 implies $\|\mathbf{g}^{\texttt{OPT}} - \mathbf{g}^{\texttt{SSRFW}}\|_2 = \|\mathbf{g}^{\texttt{OPT}} - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t - \mathbf{g}^{\texttt{SSRFW}}\|_2 \leq \|\mathbf{g}^{\texttt{OPT}} - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t\|_2 + \|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}^t - \mathbf{g}^{\texttt{SSRFW}}\|_2 = \sqrt{2\mathcal{L}(\mathbf{g}^{\texttt{OPT}})} + \sqrt{2\mathcal{L}(\mathbf{g}^{\texttt{SSRFW}})} \leq \sqrt{2\mathcal{L}(\mathbf{g}^{\texttt{OPT}})} + \sqrt{2\mathcal{L}(\mathbf{g}^{\texttt{OPT}}) + O(\frac{1}{\hat{K}})} \leq \epsilon + \sqrt{\epsilon^2 + O(\frac{1}{\hat{K}})} \leq (1+\sqrt{2})\epsilon$, when iteration number $\hat{K}$ is big enough such that $\frac{4}{\hat{K}+2} \leq \epsilon^2$. The first inequality follows from triangle inequality, the seconde inequality follow from Lemma 1 and the third inequality follows from $\mathcal{L}(\mathbf{g}^{\texttt{OPT}}) \leq \frac{\epsilon^2}{2}$ as we deduced above. (20) shows that $\|\mathbf{g}^{\texttt{OPT}} - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq \epsilon$ with high probability when we have adequate T and L. Combining these two together, we have $\|\mathbf{g}^{\texttt{SSRFW}} - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq \|\mathbf{g}^{\texttt{OPT}} - \mathbf{g}^{\texttt{SSRFW}}\|_2 + \|\mathbf{g}^{\texttt{OPT}} - \sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\|_2 \leq (2+\sqrt{2})\epsilon$ with a high probability, which concludes Lemma 4.

Q.E.D.

A.6. Proof of Theorem 1

Finally, with all the properties, we can prove the main result Theorem1.

As illustrated in Figure 1, we want to show with probability $\geq 1-\delta$ we have the following three statements

1. $\|\mathbf{g}^{\texttt{SSRFW}} - \sum_k \alpha_k\boldsymbol{q}_k\| \leq \epsilon$.
2. $\forall\ \hat{k}, \exists\ k = \pi(\hat{k})$ s.t. $\|\hat{\boldsymbol{q}}_{\hat{k}} - \boldsymbol{q}_k\| \leq \epsilon$.

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

The statement 1 is directly from Lemma 4. Statement 2 is implied from Property 1: $\forall\ \hat{k}$, $\hat{\boldsymbol{q}}_{\hat{k}}\in\mathcal{E}(\mathrm{Conv}(\mathcal{Q}))$ and Property 3: $\forall\ \bar{\boldsymbol{q}}\in\mathcal{Q},\exists\ k$ s.t. $\|\bar{\boldsymbol{q}}-\boldsymbol{q}_k\|\le\epsilon$. Since our $\mathcal{Q}$ is a finite set, we know that the extreme points of $\mathrm{Conv}(\mathcal{Q})$ must be an element of $\mathcal{Q}$. Thus every output $\hat{\boldsymbol{q}}_{\hat{k}}$ belongs to $\mathcal{Q}$, and we are able to combine properties 1 and 3.

Next, we prove Statement 3. Denote $\hat{K}$ as the number of mixtures output by the `SSRFW` algorithm. Statement 1 implies

$$\left\|\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\hat{\boldsymbol{q}}_{\hat{k}}-\sum_{k=1}^{K}\alpha_k\boldsymbol{q}_k\right\|\le\epsilon. \tag{21}$$

According to Statement 2, for any $\hat{k}$, there exists a unique $k=\pi(\hat{k})$ such that $\left\|\hat{\boldsymbol{q}}_{\hat{k}}-\boldsymbol{q}_{\pi(\hat{k})}\right\|\le\epsilon$. We can write $\hat{\boldsymbol{q}}_{\hat{k}}=\boldsymbol{q}_{\pi(\hat{k})}+\hat{\boldsymbol{\epsilon}}_{\hat{k}}$ where $\|\hat{\boldsymbol{\epsilon}}_{\hat{k}}\|\le\epsilon$. Then we can rewrite $\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\hat{\boldsymbol{q}}_{\hat{k}}$ as $\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\sum_{k:\pi(\hat{k})=k}(\boldsymbol{q}_k+\hat{\boldsymbol{\epsilon}}_{\hat{k}})$, which can be reorganized as $\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\sum_{k:\pi(\hat{k})=k}\boldsymbol{q}_k+\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\hat{\boldsymbol{\epsilon}}_{\hat{k}}=\sum_{k=1}^{K}\boldsymbol{q}_k\sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}}+\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\hat{\boldsymbol{\epsilon}}_{\hat{k}}$. Hence inequality (21) can be reorganzied as follows: Rearranging Eqn. (21) gives

$$\left\|\sum_{k=1}^{K}\boldsymbol{q}_k\left(\sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}}-\alpha_k\right)+\sum_{\hat{k}=1}^{\hat{K}}\hat{\alpha}_{\hat{k}}\hat{\boldsymbol{\epsilon}}_{\hat{k}}\right\|\le\epsilon \tag{22}$$

By triangle inequality, we get

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

Since $\boldsymbol{q}_k$ is an arbitrary non-zero vector, we must have $|\sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}}-\alpha_k|\le 2\epsilon,\ \forall\ k$, which completes Statement 3 by replacing ϵ by $\frac{\epsilon}{2}$ in the above analysis.

Finally, we prove that the mapping $\pi(\cdot)$ is a surjective mapping. Statement 3 shows that $\left|\sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}}-\alpha_k\right|<\epsilon,\forall\ k$, the quantity $\sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}}$ becomes 0 if we cannot find $\hat{k}$ such that $\pi(\hat{k})=k$, which implies $\alpha_k<\epsilon$. Therefore, under the assumption that $\epsilon\le\alpha_1$, we always find $\hat{k}$ such that $\pi(\hat{k})=k$ for any k, which suggests $\pi(\cdot)$ is a surjective mapping. This implies that $\hat{K}$ is at least as large as K. Q.E.D.

A.6.1. Proof of Corollary 1

Proof : Using the definition, we know $\hat{q}_{\hat{k}0}=1-\sum_{m=1}^{M}\hat{q}_{\hat{k}m}$. Thus $\left|\hat{q}_{\hat{k}0}-q_{\pi(\hat{k})0}\right|=\left|(1-\sum_{m=1}^{M}\hat{q}_{\hat{k}m})-(1-\sum_{m=1}^{M}q_{\pi(\hat{k})m})\right|=\left|\sum_{m=1}^{M}\hat{q}_{\hat{k}m}-q_{\pi(\hat{k})m}\right|\le\sqrt{M\cdot\sum_{m=1}^{M}(\hat{q}_{\hat{k}m}-q_{\pi(\hat{k})m})^2}=\sqrt{M}\left\|\hat{\boldsymbol{q}}_{\hat{k}}-\boldsymbol{q}_{\pi(\hat{k})}\right\|$, where the inequality follows from Cauchy-Schwarz inequality. By Theorem 1, we know $\left\|\hat{\boldsymbol{q}}_{\hat{k}}-\boldsymbol{q}_{\pi(\hat{k})}\right\|\le\epsilon$ with high probability. Thus $\left|\hat{q}_{\hat{k}0}-q_{\pi(\hat{k})0}\right|\le\sqrt{M}\epsilon$ with high probability. Q.E.D.

A.7. Proofs in Section 4.2

In this subsection, we give the proofs of several results in Section 4.2 concerning the impossible results of our sample complexity.

A.7.1. Proof of Proposition 2

Proof : Let us denote an algorithm as $\mathcal{A}$ that achieves $\frac{\epsilon}{4}$-δ learnability, suppose we have two different MMNL ground-truth mixture types, namely $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two sets of samples taken respectively from distribution $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$. We use $\mathcal{A}(\mathcal{D})$ to represent the output of algorithm $\mathcal{A}$ given a set of sample data $\mathcal{D}$. By the definition of $\frac{\epsilon}{4}$-δ learnability, we have $\|\mathcal{A}(\mathcal{D}_1)-\boldsymbol{q}_1\|_2\le\frac{\epsilon}{4}$ and $\|\mathcal{A}(\mathcal{D}_2)-\boldsymbol{q}_2\|_2\le\frac{\epsilon}{4}$ with a

probability of at least $1-\delta$. In this case, we know that if $\boldsymbol{q}_1=\boldsymbol{q}_2$, by triangle inequality, $\|\mathcal{A}(\mathcal{D}_1)-\mathcal{A}(\mathcal{D}_2)\|_2 \le \|\boldsymbol{q}_1-\mathcal{A}(\mathcal{D}_1)\|_2+\|\boldsymbol{q}_1-\boldsymbol{q}_2\|_2+\|\mathcal{A}(\mathcal{D}_2)-\boldsymbol{q}_2\|_2 = \|\boldsymbol{q}_1-\mathcal{A}(\mathcal{D}_1)\|_2+\|\mathcal{A}(\mathcal{D}_2)-\boldsymbol{q}_2\|_2 \le \frac{\epsilon}{2}$ with a probability of at least $1-\delta$. Similarly, if $\|\boldsymbol{q}_1-\boldsymbol{q}_2\|_2\ge\epsilon$, applying triangle inequalities, we have $\epsilon \le \|\boldsymbol{q}_1-\boldsymbol{q}_2\|_2 = \|\boldsymbol{q}_1-\mathcal{A}(\mathcal{D}_1)+\mathcal{A}(\mathcal{D}_1)-\mathcal{A}(\mathcal{D}_2)+\mathcal{A}(\mathcal{D}_2)-\boldsymbol{q}_2\|_2 \le \|\boldsymbol{q}_1-\mathcal{A}(\mathcal{D}_1)\|_2+\|\mathcal{A}(\mathcal{D}_1)-\mathcal{A}(\mathcal{D}_2)\|_2+\|\mathcal{A}(\mathcal{D}_2)-\boldsymbol{q}_2\|_2$. Hence, $\|\mathcal{A}(\mathcal{D}_1)-\mathcal{A}(\mathcal{D}_2)\|_2 \ge \epsilon-\|\boldsymbol{q}_1-\mathcal{A}(\mathcal{D}_1)\|_2-\|\mathcal{A}(\mathcal{D}_2)-\boldsymbol{q}_2\|_2 \ge \frac{\epsilon}{2}$ with a probability of at least $1-\delta$. We then let the algorithm output **PASS** if $\|\mathcal{A}(\mathcal{D}_1)-\mathcal{A}(\mathcal{D}_2)\|_2 \le \frac{\epsilon}{2}$ and **FAIL** if $\|\mathcal{A}(\mathcal{D}_1)-\mathcal{A}(\mathcal{D}_2)\|_2 \ge \frac{\epsilon}{2}$. In summary, if $\boldsymbol{q}_1=\boldsymbol{q}_2$, the algorithm outputs **PASS** with a probability of $1-\delta$; if $\|\boldsymbol{q}_1-\boldsymbol{q}_2\|_2\ge\epsilon$, the algorithm output **FAIL** with a probability of $1-\delta$. Hence it is a $\mathcal{L}_2$ distance test (ϵ,δ,$\boldsymbol{q}_1$,$\boldsymbol{q}_2$) according to Definition 3. Q.E.D.

A.7.2. Proof of Theorem 3 Proposition 2 has connected any algorithm that archives $\frac{\epsilon}{4}$-δ learnability to a $\mathcal{L}_2$ distance test (ϵ,δ,$\boldsymbol{q}_1$,$\boldsymbol{q}_2$) given any two distributions $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$. To build the impossibility results, we only need to construct a set of specific distributions and establish the minimal number of samples to find a $\mathcal{L}_2$ distance test for these distributions. Specifically, consider two specific distributions $\boldsymbol{h}_1=(\frac{1}{2},\frac{1}{2})$, $\boldsymbol{h}_2=(\frac{1}{2}-\frac{\epsilon}{\sqrt{2}},\frac{1}{2}+\frac{\epsilon}{\sqrt{2}})$. Clearly $\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2=\epsilon$. Then we can apply Theorem 6 in Baignères et al. (2004) to build the following lemma.

LEMMA 7. *(Theorem 6 in Baignères et al. (2004)) The error probability denoted by $\mathcal{P}_e$, which is defined as the probability for the $\mathcal{L}_2$ distance test (ϵ,δ,$\boldsymbol{h}_1$,$\boldsymbol{h}_2$) to output **PASS** can be approximated by $\Phi(-\frac{\sqrt{d}}{2})$, where Φ is the CDF of the standard normal distribution and d is a real number that determines the minimal number of samples that are needed to find a $\mathcal{L}_2$ distance test (ϵ,δ,$\boldsymbol{h}_1$,$\boldsymbol{h}_2$), which is $T=\frac{d}{2\epsilon^2}$.*

Lemma 7 is a specific application of Theorem 6 in Baignères et al. (2004). In the original paper Baignères et al. (2004), the authors have proved a more general result involving more than two distributions. Now we are ready to prove Theorem 3.

Proof :

According to Lemma 7, the error probability $\mathcal{P}_e$ for $\mathcal{L}_2$ distance test (ϵ,δ,$\boldsymbol{h}_1$,$\boldsymbol{h}_2$) to output **PASS** is $\Phi(-\frac{\sqrt{d}}{2})$, where Φ is the cumulative distribution function of standard normal random variable. On the other hand, by definition of $\mathcal{L}_2$ distance test (ϵ,δ,$\boldsymbol{h}_1$,$\boldsymbol{h}_2$), the probability of outputting **FAIL** should be at least $1-\delta$, indicating the probability of outputting **PASS** (which is exactly $\mathcal{P}_e$) is less than δ. Hence, we have $\mathcal{P}_e=\Phi(-\frac{\sqrt{d}}{2})\le\delta$. Note

$$\Phi(-\frac{\sqrt{d}}{2}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\frac{\sqrt{d}}{2}} e^{-\frac{y^2}{2}}dy \tag{23}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{\frac{\sqrt{d}}{2}}^{\infty} e^{-\frac{y^2}{2}}dy \tag{24}$$

$$\le \frac{\sqrt{2}}{\sqrt{\pi d}}\exp(-\frac{d}{8}) \tag{25}$$

$$= O(\exp(-\frac{\epsilon^2 T}{4})) \tag{26}$$

The first inequality follows from the tail inequality of standard normal distribution : $\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{y^2}{2}}dy \le \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{x}\exp(-\frac{x^2}{2})$, for $x>0$. The third equality comes from $T=\frac{d}{2\epsilon^2}$. Hence $\Phi(-\frac{\sqrt{d}}{2})\le\delta$ holds if $O(\exp(-\frac{\epsilon^2 T}{4}))\le\delta$, which implies $T=O(\frac{1}{\epsilon^2}log(\frac{1}{\delta}))$. This concludes the proof of lower bound on sample complexity. Q.E.D.

A.8. Proof of Theorem 4

Proof: The two problems mentioned in Theorem 4 can be written as:

$$\begin{aligned}
\max_{\boldsymbol{p}} \ & \textstyle\sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{w_{kj}-ln(q_{kj})+ln(q_{k0})}{\beta_{k1}} q_{kj} \\
\text{s.t.} \ & \textstyle\sum_{j=0}^{M} q_{kj} = 1, \forall k = 1, ..., K, \\
& \tfrac{1}{\beta_{k1}} \left(w_{kj} - ln(q_{kj}) + ln(q_{k0})\right) = p_{jk}, \forall k = 1, \ldots, K, \\
& p_{jk} = p_{j1}, \forall k = 1, \ldots, K, \\
& c_j \le p_{jk} \le u_j, \forall k = 1, \ldots, K,
\end{aligned} \tag{27}$$

and:

$$\begin{aligned}
\max_{\boldsymbol{p}} \ & \textstyle\sum_{\hat{k}=1}^{\hat{K}} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} \frac{w'_{kj}-ln(\hat{q}_{\hat{k}j})+ln(\hat{q}_{\hat{k}0})}{\hat{\beta}_{\hat{k}1}} \hat{q}_{\hat{k}j} \\
\text{s.t.} \ & \textstyle\sum_{j=0}^{M} \hat{q}_{\hat{k}j} = 1, \forall \hat{k} = 1, ..., \hat{K}, \\
& \tfrac{1}{\hat{\beta}_{\hat{k}1}} \left(w'_{kj} - ln(\hat{q}_{\hat{k}j}) + ln(\hat{q}_{\hat{k}0})\right) = p_{jk}, \forall \hat{k} = 1, \ldots, \hat{K}, \\
& p_{jk} = p_{j1}, \forall \hat{k} = 1, \ldots, \hat{K}, \\
& c_j \le p_{jk} \le u_j, \forall \hat{k} = 1, \ldots, \hat{K},
\end{aligned} \tag{28}$$

where (27) is the problem using the true MMNL model and (28) is the problem using parameters learned from SSFW algorithm. Since we only consider the optimal value of two problems, we can set price $\boldsymbol{p}$ as decision variables for simplicity. In this case, we can rewrite the problem as :

$$\begin{aligned}
\max_{\boldsymbol{p}} \ & \textstyle\sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} p_j q_{kj} \\
\text{s.t.} \ & q_{kj} = q_{k0} exp(w_{kj} - \beta_{k1} p_j), \forall j = 1, ..., M, \forall k = 1, ..., K, \\
& q_{k0} = \tfrac{1}{1+\sum_i exp(w_{ki} - \beta_{k1} p_i)}, \forall k = 1, ..., K, \\
& c_j \le p_j \le u_j, \forall k = 1, \ldots, K,
\end{aligned} \tag{29}$$

and:

$$\begin{aligned}
\max_{\boldsymbol{p}} \ & \textstyle\sum_{\hat{k}=1}^{\hat{K}} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j \hat{q}_{\hat{k}j} \\
\text{s.t.} \ & \hat{q}_{\hat{k}j} = \hat{q}_{\hat{k}0} exp(w'_{kj} - \hat{\beta}_{\hat{k}1} p_j), \forall j = 1, ..., M, \forall \hat{k} = 1, ..., \hat{K}, \\
& \hat{q}_{\hat{k}0} = \tfrac{1}{1+\sum_i exp(w'_{ki} - \hat{\beta}_{\hat{k}1} p_i)}, \forall \hat{k} = 1, ..., \hat{K}, \\
& c_j \le p_j \le u_j, \forall \hat{k} = 1, \ldots, \hat{K},
\end{aligned} \tag{30}$$

We can see from the formulation that the feasible points of (29) are also feasible for (30). We denote $\{\tilde{p}\}$ as the optimal solution of (29). Note that for any pricing solution $\boldsymbol{p}$, we have:

$$\begin{aligned}
& \Big| \sum_{\hat{k},\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j \hat{q}_{\hat{k}j} - \alpha_k \sum_{j=1}^{M} p_j q_{kj} \Big| \\
\le & \Big| \sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j \hat{q}_{\hat{k}j} - \sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j q_{kj} \Big| + \Big| \sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j q_{kj} - \alpha_k \sum_{j=1}^{M} p_j q_{kj} \Big| \\
\le & \Big| \sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j (\hat{q}_{\hat{k}j} - q_{kj}) \Big| + \Big| \Big(\sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} - \alpha_k \Big) \sum_{j=1}^{M} p_j q_{kj} \Big|
\end{aligned}$$

By the results of SSFW algorithm (Theorem 1), for any $\epsilon > 0$, we can construct $\{\hat{\alpha}_{\hat{k}}\}$ and $\{\hat{\boldsymbol{q}}_{\hat{k}}\}$, such that $|\sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} - \alpha_k| \le \epsilon$, $\|\hat{\boldsymbol{q}}_{\hat{k}} - \boldsymbol{q}_k\| < \epsilon, \forall\ \hat{k}$. Hence we have the following:

$$\Big| \sum_{\hat{k},\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j \hat{q}_{\hat{k}j} - \alpha_k \sum_{j=1}^{M} p_j q_{kj} \Big| \le \sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} \sum_{j=1}^{M} p_j \epsilon + \epsilon \sum_{j=1}^{M} p_j q_{kj}$$

$$\leq Mu_{max}\epsilon \sum_{\hat{k}:\pi(\hat{k})=k} \hat{\alpha}_{\hat{k}} + Mu_{max}\epsilon$$

This inequality holds for any pricing solution $\boldsymbol{p}$. Based on this inequality and applying a triangle inequality we can deduce that:

$$\sum_{k=1}^{K}\alpha_k\sum_{j=1}^{M}\tilde{p}_j q_{kj} \leq \sum_{k=1}^{K}(\sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}}\sum_{j=1}^{M}\tilde{p}_j\hat{q}_{\hat{k}j}) + \sum_{k=1}^{K}Mu_{max}\epsilon \sum_{\hat{k}:\pi(\hat{k})=k}\hat{\alpha}_{\hat{k}} + \sum_{k=1}^{K}Mu_{max}\epsilon \tag{31}$$

$$\leq \sum_{k=1}^{\hat{K}}\alpha_k\sum_{j=1}^{M}\tilde{p}_j\hat{q}_{\hat{k}j} + (K+1)Mu_{max}\epsilon \tag{32}$$

$$\leq Z(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{q}},\hat{\boldsymbol{\alpha}}) + (K+1)Mu_{max}\epsilon \tag{33}$$

Since $\sum_{k=1}^{K}\alpha_k\sum_{j=1}^{M}\tilde{p}_j q_{kj} = Z(\boldsymbol{\beta},\boldsymbol{q},\boldsymbol{\alpha})$, we obtain $Z(\boldsymbol{\beta},\boldsymbol{q},\boldsymbol{\alpha}) \leq Z(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{q}},\hat{\boldsymbol{\alpha}}) + (K+1)Mu_{max}\epsilon$, and this completes the proof. Q.E.D.

Appendix B: Piece-wise Linear Approximation for Pricing with MMNL

In this section, we provide a piece-wise linear approximation (PLA) for the pricing with MMNL model and analyze the theoretical guarantees for this method. We denote R as $Z(\boldsymbol{\beta},\boldsymbol{q},\boldsymbol{\alpha})$, $\hat{R}$ as $Z(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{q}},\hat{\boldsymbol{\alpha}})$, and R', $\hat{R}'$ as the corresponding optimal value of PLA formation of $Z(\boldsymbol{\beta},\boldsymbol{q},\boldsymbol{\alpha})$, $Z(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{q}},\hat{\boldsymbol{\alpha}})$ respectively in the following discussion. The proofs of all theoretical guarantees are placed in Section

B.1. Piece-wise Linear Approximation (PLA)

Suppose the price constraints specify only the interval for each price. Specifically, for each product j, there is an upper bound u_j, and lower bound c_j for its price p_j, i.e., $c_j \leq p_j \leq u_j$. For analysis simplicity, we introduce a set of dummy variables p_{jk} to indicate the price of product j charged for type k customers. Then $p_{jk} = p_j$ for all k. For notation simplicity, we use w_{kj} to represent $\boldsymbol{\beta}_k^{\mathsf{T}}\boldsymbol{z}_j$. Then (5) can be specialized as (34).

$$\begin{aligned}
\max_{\boldsymbol{q}}\ & \textstyle\sum_{k=1}^{K}\alpha_k\sum_{j=1}^{M}\frac{w_{kj}-ln(q_{kj})+ln(q_{k0})}{\beta_{k1}}q_{kj} \\
\text{s.t.}\ & \textstyle\sum_{j=0}^{M} q_{kj} = 1, \forall k = 1,...,K, \\
& \tfrac{1}{\beta_{k1}}\left(w_{kj} - ln(q_{kj}) + ln(q_{k0})\right) = p_{jk}, \forall k = 1,\ldots,K, \\
& p_{jk} = p_{j1}, \forall k = 1,\ldots,K, \\
& c_j \leq p_{jk} \leq u_j, \forall k = 1,\ldots,K.
\end{aligned} \tag{34}$$

Next, we use piece-wise linear functions $\hat{\omega}(x)$ and $\hat{\nu}(x)$ to approximate $ln(x), xln(x)$, respectively. According to Magnanti and Stratila (2004), Thakur (1978), Kontogiorgis (2000), Güder and Morris (1994) and Goh and Yan (2022), a non-linear function $f(x)$ of market share can be approximated by a carefully designed piece-wise linear function if $x > \epsilon_0$, where ϵ_0 is a very small positive number. Hence we make an additional assumption to get the performance guarantee of the PLA.

Assumption 2. $q_{jk} > \epsilon_0$.

This assumption is essentially to assume a strictly positive choice probability for each product, which is mild since if the choice probability is almost zero, the product can be removed from the assortment.

Applying the construction algorithm in Goh and Yan (2022), we can easily find $\hat{\omega}(x)$ and $\hat{\nu}(x)$ such that $|\hat{\omega}(x) - ln(x)| \leq \frac{\epsilon_1}{2}, |\hat{\nu}(x) - xln(x)| \leq \epsilon_2$ for any pre-specified positive small numbers (accuracy tolerance) ϵ_1 and ϵ_2. Then (34) can be approximated by (35).

$$\begin{aligned}
\max_{\boldsymbol{q}} \ & \textstyle\sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{q_{kj} w_{kj} - \hat{\nu}(q_{kj}) + q_{kj}\hat{\omega}(q_{k0})}{\beta_{k1}} \\
\text{s.t.} \ & \textstyle\sum_{j=0}^{M} q_{kj} = 1, \forall k = 1, ..., K, \\
& \tfrac{1}{\beta_{k1}} \left(w_{kj} - \hat{\omega}(q_{kj}) + \hat{\omega}(q_{k0})\right) = p_{jk}, \forall k = 1, \ldots, K, \\
& |p_{jk} - p_{j1}| \leq \tfrac{\epsilon_1}{\beta_{k1}} + \tfrac{\epsilon_1}{\beta_{11}}, \forall k = 1, \ldots, K, \\
& c_j - \tfrac{\epsilon_1}{\beta_{k1}} \leq p_{jk} \leq u_j + \tfrac{\epsilon_1}{\beta_{k1}}, \forall k = 1, \ldots, K,
\end{aligned} \tag{35}$$

In (35), we have further relaxed the price constraints for the computational efficiency consideration. Later, in the next section, we will construct a feasible pricing solution to (34) based on the optimal solution to (35) and demonstrate the performance of the constructed pricing solution. Note that problem (35) can be formulated as an MIP. The detailed formula is as follows.

We denote the breakpoints of $ln(x)$ as $\{l\}_{i=1}^{Jl}$, the break points of $xln(x)$ as $\{v\}_{i=}^{Jv}$. Then we have the following MIP formulation:

$$\begin{aligned}
\max_{\boldsymbol{q}} \ & \textstyle\sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{q_{kj} w_{kj} - \hat{\nu}(q_{kj}) + q_{kj}\hat{\omega}(q_{k0})}{\beta_{k1}} \\
\text{s.t.} \ & \textstyle\sum_{j=0}^{M} q_{kj} = 1, \forall k = 1, ..., K, \\
& \tfrac{1}{\beta_{k1}} \left(w_{kj} - \hat{\omega}(q_{kj}) + \hat{\omega}(q_{k0})\right) = p_{jk}, \forall k = 1, \ldots, K, \\
& |p_{jk} - p_{j1}| \leq \tfrac{\epsilon_1}{\beta_{k1}} + \tfrac{\epsilon_1}{\beta_{11}}, \forall k = 1, \ldots, K, \\
& c_j - \tfrac{\epsilon_1}{\beta_{k1}} \leq p_{jk} \leq u_j + \tfrac{\epsilon_1}{\beta_{k1}}, \forall k = 1, \ldots, K, \\
& \textstyle\sum_{i=1}^{Jl} \lambda_{kji} l_i = q_{kj}, \forall k, j \\
& \textstyle\sum_{i=1}^{Jv} \mu_{kji} v_i = q_{kj}, \forall k, j \\
& \textstyle\sum_{i=1}^{Jl} \lambda_{kji} ln(l_i) = \hat{\omega}(q_{kj}), \forall k, j \\
& \textstyle\sum_{i=1}^{Jv} \mu_{kji} v_i ln(v_i) = \hat{\nu}(q_{kj}), \forall k, j \\
& \textstyle\sum_{i=1}^{Jl} \lambda_{kji} = 1, \forall k, j \\
& \textstyle\sum_{i=1}^{Jv} \mu_{kji} = 1, \forall k, j \\
& \lambda_{kjJl} \leq y1_{kjJl}, \lambda_{kj1} \leq y1_{kj1}, \forall k, j \\
& \mu_{kjJv} \leq y2_{kjJv}, \mu_{kj1} \leq y2_{kj1} \forall k, j \\
& \lambda_{kji} \leq y1_{kji-1} + y1_{kji} \forall k, j, i = 2, ..., Jl - 1 \\
& \mu_{kji} \leq y2_{kji-1} + y2_{kji} \forall k, j, i = 2, ..., Jv - 1
\end{aligned} \tag{36}$$

where $\{y1\}, \{y2\}$ are binary numbers and $\{\lambda\}, \{\mu\} \geq 0$

B.2. Constructed PLA Pricing Policy and Its Performance

Denote $\{\boldsymbol{q}'\}$ as the optimal solution of problem (35) and p'_{jk} as the corresponding price computed in (35). From the constraint in (35), we have:

$$p'_{jk} = \tfrac{1}{\beta_{k1}} \left(w_{kj} - \hat{\omega}(q'_{kj}) + \hat{\omega}(q'_{k0})\right) \tag{37}$$

We choose $p'_j = \max_k p'_{jk}$ as the price of the jth item. Then we can use $\{p'\}$ to compute the corresponding objective value of (34), which is denoted by R'. Denote the optimal value of (34) as R and we establish the relation between R' and R in the following theorem.

THEOREM 5. *For any given $\epsilon_1 > 0, \epsilon_2 > 0$, suppose functions $\hat{\omega}(x)$ and $\hat{\nu}(x)$ are piece-wise linear approximation of $ln(x)$ and $xln(x)$ such that $|\hat{\omega}(x) - ln(x)| \le \epsilon_1/2, |\hat{\nu}(x) - xln(x)| \le \epsilon_2$. Then we have $R' \ge e^{-2\epsilon_1 - \frac{2\epsilon_1\beta_{max}}{\beta_{min}}} \left(R - \frac{M(\epsilon_1+2\epsilon_2)}{\beta_{min}}\right)$, where $\beta_{min} = \min_{k,j} \beta_{kj}, \beta_{max} = \max_{k,j} \beta_{kj}$.*

The bound in Theorem 5 demonstrates that when ϵ_1, ϵ_2 are small enough, R' will converge to the true optimal value R, since we always have $R \ge R'$. Specifically, suppose we have 3 different products and $\beta_{min} = 0.00142, \beta_{max} = 1$. We choose $\epsilon_1 = \epsilon_2 = 10^{-5}$, then we can have the bound $R \ge R' \ge 0.98\,(R - 0.042)$. Also, we can see that ϵ_2 does not play an important role in the above theorem, which means we can relax the choice of ϵ_2, gaining more computational efficiency.

The performance guarantee established in Theorem 5 is based on the true parameters (K, α, q_{kj}) in the MMNL model. We further evaluate the performance of the pricing solution obtained from the PLA using the estimated MMNL from our SSFW algorithm.

COROLLARY 2. *For any given precision level ϵ in the SSFW algorithm, let $\hat{\alpha}, \hat{\boldsymbol{q}}$ denote the output of SSFW algorithm. Then the relation between PLA using learned MMNL model whose objective value is denoted by $\hat{R}'$ and the optimal value of the true MMNL denoted by R is $\hat{R}' \ge e^{-2\epsilon_1 - \frac{2\epsilon_1\hat{\beta}_{max}}{\hat{\beta}_{min}}} \left(R - \frac{M(\epsilon_1+2\epsilon_2)}{\hat{\beta}_{min}} - (K+1)Mu_{max}\epsilon\right)$, where $\hat{\beta}_{min} = \min_{\hat{k},j} \hat{\beta}_{\hat{k}j}, \hat{\beta}_{max} = \max_{\hat{k},j} \hat{\beta}_{\hat{k}j}$ are learned from `SSRFW` , $u_{max} = \max_j u_j$.*

Similar to theorem 5, we can demonstrate the bound in Corollary 2. Let $\hat{\beta}_{min} = 0.00142, \hat{\beta}_{max} = 1$ and assume $M = 10, K = 5, u_{max} = 1000$. In this case, if we have $\epsilon_1 = \epsilon_2 = 10^{-5}$ and $\epsilon = 10^{-4}$. Then we have $\hat{R}' \ge 0.98(R - 6.2)$.

To conclude the session, it is worthwhile to mention that pricing with mixed MNL models remains an open question (Hanson and Martin (1996) and Li et al. (2019)). The current literature lacks solutions for precisely solving the pricing problem with an arbitrary mixed MNL model. A seminar work by Li et al. (2019) proposes a gradient descent algorithm to approximate the pricing problem, but they do not provide a theoretical guarantee for the optimality of the obtained pricing solution from their algorithm. However, we emphasize our contribution to the pricing literature by first providing a provable performance guarantee for our proposed approximation method (see Theorem 5), based on which, we can further analyze the effect of estimation error on the optimality of the pricing solution (see Corollary 2).

B.3. Proofs of PLA performance guarantees

B.3.1. Proof of Theorem 5

First we consider an intermediate problem as follow:

$$\begin{aligned}
\max_{\boldsymbol{q}} \ & \textstyle\sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{w_{kj} - ln(q_{kj}) + ln(q_{k0})}{\beta_{k1}} q_{kj} \\
\text{s.t.} \ & \textstyle\sum_{j=0}^{M} q_{kj} = 1, \forall k = 1, ..., K, \\
& \tfrac{1}{\beta_{k1}} \left(w_{kj} - \hat{\omega}(q_{kj}) + \hat{\omega}(q_{k0})\right) = p_{jk}, \forall k = 1, \ldots, K, \\
& |p_{jk} - p_{j1}| \le \tfrac{\epsilon_1}{\beta_{k1}} + \tfrac{\epsilon_1}{\beta_{11}}, \forall k = 1, \ldots, K, \\
& c_j - \tfrac{\epsilon_1}{\beta_{k1}} \le p_{jk} \le u_j + \tfrac{\epsilon_1}{\beta_{k1}}, \forall k = 1, \ldots, K,
\end{aligned} \tag{38}$$

Since $|\hat{\omega}(x) - ln(x)| \leq \epsilon_1/2, |\hat{\nu}(x) - xln(x)| \leq \epsilon_2$, we can see that every feasible solution of (34) is a feasible solution of (38), denote the optimal value of (38) as R_m, we thus have $R \leq R_m$. Moreover, we should notice that the feasible region of (38) and the feasible region of (35) are identical, for simplicity, we denote the feasible region of (35) as $\hat{\Omega}_q$. For any $\{q\} \in \hat{\Omega}_q$, we have:

$$\begin{aligned}
&|\frac{w_{kj} - ln(q_{kj}) + ln(q_{k0})}{\beta_{k1}} q_{kj} - \frac{w_{kj} - \hat{\nu}(q_{kj}) + q_{kj}\hat{\omega}(q_{k0})}{\beta_{k1}}| \\
&\leq |\frac{q_{kj}ln(q_{kj}) - \hat{\nu}(q_{kj})}{\beta_{k1}}| + |\frac{q_{kj}ln(q_{k0}) - \omega(\hat{q_{k0}})}{\beta_{k1}}| \\
&\leq \frac{\epsilon_2}{\beta_{min}} + q_{kj}\frac{\epsilon_1}{2\beta_{min}}.
\end{aligned}$$

We denote the optimal solution of (38) as $\{q^m\}$, then we have the following result:

$$\begin{aligned}
R \leq R_m &= \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{w_{kj} - ln(q_{kj}^m) + ln(q_{k0}^m)}{\beta_{k1}} q_{kj}^m \\
&\leq \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{w_{kj} - \hat{\nu}(q_{kj}^m) + q_{kj}^m \hat{\omega}(q_{k0}^m)}{\beta_{k1}} + \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{\epsilon_2}{\beta_{min}} + q_{kj}^m \frac{\epsilon_1}{2\beta_{min}} \\
&\leq R^\star + \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{\epsilon_2}{\beta_{min}} + q_{kj}^m \frac{\epsilon_1}{2\beta_{min}},
\end{aligned}$$

where $R^\star$ is the optimal value of problem (35). Since $\sum_{k=1}^{K} \alpha_k = 1$ and $q_{kj}^m \leq 1$, we have:

$$R \leq R^\star + \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{\epsilon 1}{\beta_{min}} + \frac{\epsilon_1}{2\beta_{min}} \tag{39}$$

$$\leq R^\star + \frac{M\left(2\epsilon_2 + \epsilon_1\right)}{2\beta_{min}} \tag{40}$$

Next, we need to compare the difference between $R^\star$ and R'. Denote $\{\boldsymbol{q}'\}$ as the optimal solution of (35) and $\{p'\}_{j=1}^{M}$ as the price made according to our policy. Note that $p'_j = max_k p'_{jk}$, we have:

$$R^\star = \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{q'_{kj} w_{kj} - \hat{\nu}(q'_{kj}) + q'_{kj}\hat{\omega}(q'_{k0})}{\beta_{k1}} \tag{41}$$

Since $p'_{jk} = \frac{1}{\beta_{k1}}\left(w_{kj} - \hat{\omega}(q'_{kj}) + \hat{\omega}(q'_{k0})\right)$, we have :

$$\begin{aligned}
|p'_{jk}q'_{kj} - \frac{q'_{kj}w_{kj} - \hat{\nu}(q'_{kj}) + q'_{kj}\hat{\omega}(q'_{k0})}{\beta_{k1}}| &= |\frac{\hat{\omega}(q'_{kj})q'_{kj} - \hat{\nu}(q'_{kj})}{\beta_{k1}}| \\
&\leq |\frac{\hat{\omega}(q'_{kj})q'_{kj} - q'_{kj}ln(q'_{kj})}{\beta_{k1}}| + |\frac{\hat{\nu}(q'_{kj}) - q'_{kj}ln(q'_{kj})}{\beta_{k1}}| \\
&\leq \frac{q'_{kj}\epsilon_1}{2\beta_{min}} + \frac{\epsilon_2}{\beta_{min}} \\
&\leq \frac{\epsilon_1}{2\beta_{min}} + \frac{\epsilon_2}{\beta_{min}}
\end{aligned}$$

Combining with (41), we can get:

$$\begin{aligned}
R^\star &\leq \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} p'_{jk}q'_{jk} + \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} \frac{\epsilon_1}{2\beta_{min}} + \frac{\epsilon_2}{\beta_{min}} \\
&\leq \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} p'_j q'_{jk} + \frac{M\left(2\epsilon_2 + \epsilon_1\right)}{2\beta_{min}}
\end{aligned}$$

Let $q'_{kj-appro}$ denote the probability corresponding to p'_j, that is to say, $q'_{kj-appro} = q'_{k0-appro} e^{w_{kj}-\beta_{k1}p'_j}$ for $j = 1, ..., M$ and $q'_{k0-appro} = \frac{1}{1+\sum_i exp(w_{ki}-\beta_{k1}p'_i)}$ Then we first build the following lemma:

LEMMA 8. *Using the same notation above, we have* $q'_{k0} \le q'_{k0-appro} e^{\epsilon_1}$ *and* $q'_{kj} \le q'_{kj-appro} e^{2\epsilon_1 + \frac{2\epsilon_1 \beta_{max}}{\beta_{min}}}$.

Applying lemma 8, we have:

$$R^\star \le \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{M} p'_j q'_{kj-appro} e^{2\epsilon_1 + \frac{2\epsilon_1\beta_{max}}{\beta_{min}}} + \frac{M(2\epsilon_2+\epsilon_1)}{2\beta_{min}} \tag{42}$$

$$= e^{2\epsilon_1 + \frac{2\epsilon_1\beta_{max}}{\beta_{min}}} R' + \frac{M(2\epsilon_2+\epsilon_1)}{2\beta_{min}} \tag{43}$$

Combining (43) and $R \le R^\star + \frac{M(2\epsilon_2+\epsilon_1)}{2\beta_{min}}$, we can have:

$$R' \ge e^{-2\epsilon_1 - \frac{2\epsilon_1\beta_{max}}{\beta_{min}}} \left(R - \frac{M(2\epsilon_2+\epsilon_1)}{\beta_{min}} \right)$$

which completes the proof.

Proof of Lemma 8

Proof: First of all, we have $q'_{k0-appro} = 1/1 + \sum_i exp(w_{ki} - \beta_{k1}p'_i)$, since $p'_i \ge p'_{ik}$ for every k, we can get:

$$q'_{k0-appro} \ge \frac{1}{1+\sum_i exp(w_{ki} - \beta_{k1}p'_{ik})} \tag{44}$$

By definition, $p'_{ik} = \frac{w_{kj} - \hat{\omega}(q'_{ki}) + \hat{\omega}(q'_{k0})}{\beta_{k1}}$, so we have:

$$q'_{k0-appro} \ge \frac{1}{1+\sum_i exp(\hat{\omega}(q'_{ki}) - \hat{\omega}(q'_{k0}))} \tag{45}$$

By construction, we have $\hat{\omega}(q'_{ki}) \le ln(q'_{ki}) + \frac{\epsilon_1}{2}$, so $exp(\hat{\omega}(q'_{ki})) \le e^{\frac{\epsilon_1}{2}} q'_{ki}$ combining with the above inequality:

$$\begin{aligned} q'_{k0-appro} &\ge \frac{exp(\hat{\omega}(q'_{k0}))}{exp(\hat{\omega}(q'_{k0})) + exp(\frac{\epsilon_1}{2}) \sum_i q'_{ki}} \\ &\ge \frac{e^{\hat{\omega}(q'_{k0})}}{e^{\hat{\omega}(q'_{k0})} + e^{\frac{\epsilon_1}{2}}(1-q'_{k0})} \\ &\ge \frac{q'_{k0}/e^{\frac{\epsilon_1}{2}}}{e^{\frac{\epsilon_1}{2}} q'_{k0} + e^{\frac{\epsilon_1}{2}}(1-q'_{k0})} \end{aligned}$$

So we have $q'_{k0} \le e^{\epsilon_1} q'_{k0-appro}$.

Next we consider $q'_{ki-appro} = q'_{k0-appro} e^{w_{kj} - \beta_{k1}p'_j}$. We first write p'_{jk} as :

$$p'_{jk} = \frac{w_{kj} - ln(q'_{kj}) + ln(q'_{k0}) + \epsilon_1(q'_{k0}) - \epsilon_1(q'_{kj})}{\beta_{k1}} \tag{46}$$

where $|\epsilon_1(x)| \le \epsilon_1$ In this case, we have:

$$\begin{aligned} q'_{kj} &= q'_{k0} exp(w_{kj} - \beta_{k1}p'_{jk} + \epsilon_1(q'_{k0}) - \epsilon_1(q'_{kj})) \\ &\le e^{\epsilon_1} q'_{k0-appro} exp(w_{kj} - \beta_{k1}p'_{jk} + \epsilon_1) \\ &= e^{2\epsilon_1} q'_{k0-appro} exp(w_{kj} - \beta_{k1}p'_{jk}) \end{aligned}$$

Since $p'_j = \max_k p'_{jk}$, there exists a k_0, such that $p'_j = p'_{jk_0}$. By the constraints of (35) we have $p'_{jk_0} \le p'_{ik} + \frac{2\epsilon_1}{\beta_{min}}, -p'_{jk_0} \le -p'_{ik} + \frac{2\epsilon_1}{\beta_{min}}$. Combining with the results above, we have:

$$\begin{aligned} q'_{kj} &\le e^{2\epsilon_1} q'_{k0-appro} exp(w_{kj} - \beta_{k1}(p'_j - \frac{2\epsilon_1}{\beta_{min}})) \\ &\le e^{2\epsilon_1 + 2\epsilon_1 \frac{\beta_{max}}{\beta_{min}}} q'_{kj-appro} \end{aligned}$$

which completes the proof.

B.3.2. Proof of Corollary 2

Proof: Corollary 2 follows easily from Theorem 5 and Theorem 4. From Theorem 4, we know $\hat{R} \geq R - (K+1)Mu_{max}\epsilon$. And from Theorem 5, we know using PLA formation to solve pricing problem gives us bound $\hat{R}' \geq e^{-2\epsilon_1 - \frac{2\epsilon_1 \hat{\beta}_{max}}{\hat{\beta}_{min}}} \left(\hat{R} - \frac{M(2\epsilon_2+\epsilon_1)}{\hat{\beta}_{min}}\right)$, replacing $\hat{R}$ with R, we get $\hat{R}' \geq e^{-2\epsilon_1 - \frac{2\epsilon_1 \hat{\beta}_{max}}{\hat{\beta}_{min}}} \left(R - \frac{M(2\epsilon_2+\epsilon_1)}{\hat{\beta}_{min}} - (K+1)Mu_{max}\epsilon\right)$, which completes the proof.

References

Baignères, Thomas, Pascal Junod, Serge Vaudenay. 2004. How far can we go beyond linear cryptanalysis? *Advances in Cryptology - ASIACRYPT 2004, 10th International Conference on the Theory and Application of Cryptology and Information Security, Jeju Island, Korea, December 5-9, 2004, Proceedings*, *Lecture Notes in Computer Science*, vol. 3329. Springer, 432–450. doi:10.1007/978-3-540-30539-2_31. URL `https://iacr.org/archive/asiacrypt2004/33290427/33290427.pdf`.

Dvoretzky, A., J. Kiefer, J. Wolfowitz. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics* **27**(3) 642–669. URL `http://www.jstor.org/stable/2237374`.

Goh, Youhui, Zhenzhen Yan. 2022. Minimizing the number of pieces for piecewise linear approximation in separable concave minimization. Tech. rep., Working Paper.

Güder, Faruk, James G Morris. 1994. Optimal objective function approximation for separable convex quadratic programming. *Mathematical programming* **67** 133–142.

Hanson, Ward, Kipp Martin. 1996. Optimizing multinomial logit profit functions. *Management Science* **42**(7) 992–1003.

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

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

Kontogiorgis, Spyros. 2000. Practical piecewise-linear approximation for monotropic optimization. *INFORMS Journal on Computing* **12**(4) 324–340.

Li, Hongmin, Scott Webster, Nicholas Mason, Karl Kempf. 2019. Product-line pricing under discrete mixed multinomial logit demand: winner—2017 m&som practice-based research competition. *Manufacturing & Service Operations Management* **21**(1) 14–28.

Magnanti, Thomas L, Dan Stratila. 2004. Separable concave optimization approximately equals piecewise linear optimization. *International Conference on Integer Programming and Combinatorial Optimization*. Springer, 234–243.

Thakur, Lakshman S. 1978. Error analysis for convex separable programs: the piecewise linear approximation and the bounds on the optimal objective value. *SIAM Journal on Applied Mathematics* **34**(4) 704–714.