Light RUMs

Flavio Chierichetti [1] **Ravi Kumar** [2] **Andrew Tomkins** [2]

Abstract

A *Random Utility Model* (RUM) is a distribution on permutations over a universe of items. For each subset of the universe, a RUM induces a natural distribution of the *winner* in the subset: choose a permutation according to the RUM distribution and pick the maximum item in the subset according to the chosen permutation. RUMs are widely used in the theory of discrete choice.

In this paper we consider the question of the (lossy) compressibility of RUMs on a universe of size n, i.e., the minimum number of bits required to approximate the winning probabilities of each slate. Our main result is that RUMs can be approximated using $\widetilde{O}(n^2)$ bits, an exponential improvement over the standard representation; furthermore, we show that this bound is optimal.

En route, we sharpen the classical existential result of McFadden & Train (2000) by showing that the minimum size of a mixture of multinomial logits required to approximate a general RUM is $\widetilde{\Theta}(n)$.

1. Introduction

Random utility models, or RUMs, are the most influential and well-studied class of user behavior models in the field of discrete choice (see Train, 2003, for an overview). A model in the RUM family is a predictor for the item a user will choose when presented with a slate of options. The prediction has a specific form: the user arrives with a utility in mind for each item of the universe, drawn from a joint distribution over utility vectors. For any slate of options, the user behaves *rationally* by selecting the highest-utilty option available. Much of the work in the area covers subclasses of RUMs in which the utility distribution has a specific form, but in this paper we consider the general class.

A seminal work of McFadden & Train (2000) showed that any RUM may be approximated by a mixture of multinomial logit (MNL) models. Models of this form, MNL mixtures, have seen significant recent study, as in (Shazeer et al., 2017; Yang et al., 2018). However, the construction of McFadden & Train (2000) relied on unboundedly many mixture components, leaving little understanding of the correct complexity. Partial progress on this question was attained by Chierichetti et al. (2018a), who showed that quadratically many mixture components or quadratically many permutations suffice to approximate any RUM. In this paper we continue to study efficient representations for approximating any RUM.

We begin with some intuition about approximating RUMs. Any RUM encodes exponentially many distributions over a universe of n items, in the sense that each of the exponentially many slates (subsets of the universe) induces a distribution over the item of the slate that a random user will prefer. These distributions are not arbitrary, as the RUM imposes some structure relating the distributions of nearby slates. Nonetheless, exactly representing the RUM requires exponentially many bits. The question therefore is whether the combination of the restrictions given by the form of RUMs plus the ability to introduce a small and controllable error in the approximation will allow a more concise representation of the entire RUM.

An answer to this question will reveal how much practical information a RUM actually carries, as well as how concisely it can be specified and communicated. As in sketching, metric approximation, and related areas, the information content captures some aspect of the representative power of the model. While RUMs have seen significant investment over many decades, leading to the 2000 Nobel prize being awarded to Daniel McFadden "for his development of theory and methods for analyzing discrete choice," computational and information-theoretic properties of the model have not seen the same level of scrutiny until more recently. Hence, fundamental questions such as the amount of information required to approximate a RUM remain unresolved to now.

Our Results. This paper closes the question to within logarithmic factors, as follows. For a universe of n items, any RUM may be approximated arbitrarily closely on all slates as either a mixture of $\Theta(n)$ permutations or as a mixture

[*]Equal contribution [1]Dipartimento di Informatica, Sapienza University of Rome, Italy [2]Google, Mountain View, CA, USA. Correspondence to: Flavio Chierichetti <flavio@di.uniroma1.it>, Ravi Kumar <ravi.k53@gmail.com>, Andrew Tomkins <atomkins@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

of $\widetilde{\Theta}(n)$ MNLs, and both these bounds are tight up to logarithmic factors. Either type of model may be represented using $\widetilde{\Theta}(n^2)$ bits; such a representation is also tight, in the strict sense that no representation of asymptotically fewer bits is sufficient to approximate a generic RUM, no matter the form of the representation.

We show additionally that the mixture of MNLs representation may be viewed as strictly more powerful than the mixture of permutations in the technical sense that any mixture of t permutations may be well-approximated by a mixture of t MNLs, but the converse is not true: even mixtures of a single MNL may require $\Omega(n)$ permutations to approximate. Finally, we show that a mixture of $\widetilde{\Theta}_n(k)$ permutations is sufficient to approximate any RUM model on slates of size at most k, a common setting, and this result also is tight to polylogarithmic factors in n.

We perform some experiments to explore the implications of these theoretical results. In our first set of experiments we study a dataset in which users provide their total ordering of different sushi variants, essentially encoding a complete RUM. This allows us to study exactly how well the construction in our upper bound approximates a RUM in a practical setting. We show that the quality of approximation is almost perfectly predicted by our theoretical results, and that a representation based on just 1% of the data provides an accurate approximation of the overall RUM.

In our second set of experiments we consider the setting of Ragain & Ugander (2016), who present an interesting non-rational choice model that is incomparable to the class of RUMs. In this setting we are able to compare the non-rational choice models learned by Ragain & Ugander (2016) against RUMs we learn based on a simple linear program. RUMs perform well, outperforming the MNL mixtures model of Ragain & Ugander (2016), and in some cases outperforming the non-rational choice model, as well. This suggests that, at least in some settings, our findings on expressive power of permutation mixtures may point to new algorithmic approaches to RUM discovery.

The paper is structured as follows. Section 2 introduces the notation. Sections 3 and 4 give, respectively, the upper and lower bound on the bit complexity of arbitrary representations. Section 5 gives the corresponding bounds for MNL mixtures. Section 6 compares the RUM, and the mixture of MNLs, representations. Section 7 extends the results to the setting of slates of bounded size. Section 8 gives our experimental results. The Supplementary Material contains each proof missing from the main body of this paper, along with an exponential lower bound on the bit complexity of exact representations of RUMs, and comparisons of RUMs with several other choice models.

2. Preliminaries

Distributions. Throughout the paper, we deal with discrete probability distributions. Let $\mathrm{supp}(D)$ denote the support of a discrete distribution D. We use $x \sim D$ to denote that $x \in \mathrm{supp}(D)$ is sampled according to D. For a generic x, we use $D(x)$ to denote the probability that D assigns to x; in particular, if $x \in \mathrm{supp}(D)$, then $D(x) > 0$, otherwise $D(x) = 0$.

For $S \subseteq \mathrm{supp}(D)$, we use $D(S)$ to denote the probability that $x \sim D$ is in S, i.e., $D(S) = \sum_{x \in S} D(x)$. When this creates no ambiguity, for a generic set S, we use $D(S)$ to denote $D(S \cap \mathrm{supp}(D))$.

The *total variation distance* between D and D' is equal to

$$|D - D'|_{\mathrm{tv}} = \frac{1}{2} \sum_{\substack{x \in \mathrm{supp}(D) \\ \cup\, \mathrm{supp}(D')}} |D(x) - D'(x)| = \frac{1}{2} |D - D'|_1 .$$

The total variation distance between D and D' is also equal to the maximum, over all the events ξ, of the absolute difference between the probabilities of ξ in D and in D', i.e.,

$$|D - D'|_{\mathrm{tv}} = \max_{S \subseteq \mathrm{supp}(D) \cup \mathrm{supp}(D')} |D(S) - D'(S)| .$$

Permutations and RUMs. Let $[n] = \{1, \ldots, n\}$, $2^{[n]}$ be the power set of $[n]$, $\binom{[n]}{k}$ be the set of subsets of $[n]$ of size k. Let $\mathbf{S}_n$ be the set of *permutations* of the set $[n]$. For a given permutation $\pi \in \mathbf{S}_n$ and for $i \in [n]$, we let $\pi(i) \in [n]$ be the *value* (or *position*) of item i in π. E.g., if $\pi = (2 \prec 3 \prec 1)$, then $\pi(2) = 1, \pi(3) = 2, \pi(1) = 3$.

In this paper we use the term *slate* to denote any non-empty subset of $[n]$. Given $\pi \in \mathbf{S}_n$ and a slate $T \subseteq [n]$, let

$$\pi(T) = \arg\max_{i \in T} \pi(i),$$

i.e., the maximum item in T according to π, aka, the *winner*.

A *RUM* on $[n]$ is a probability distribution D over $\mathbf{S}_n$.[1] We drop the quantifier "on $[n]$" when it is obvious from the context. Given a slate $T \subseteq [n]$, we use D_T to denote the distribution of the random variable $\pi(T)$ for $\pi \sim D$, i.e., the distribution of the winner in the slate T with a random permutation from D. Note that $\mathrm{supp}(D_T) \subseteq T$.

[1] RUMs are typically presented in terms of noisy item evaluations made by users. Each item $i \in [n]$ is assumed to have some base value V_i; each user samples a noise vector $(E_1, \ldots, E_n)$ from a joint noise distribution, and observes the utility of $i \in [n]$ to be $U_i = V_i + E_i$. The user then chooses an item "rationally" as the option with the highest utility U_i between the available ones (breaking ties, if they exist, u.a.r.). As the utilities are random, the family of resulting models is named "Random Utility Models," or RUMs. In a second definition, the user first sorts all the items decreasingly according to their observed utilities U_i (breaking ties, if they exist, u.a.r.), obtaining a permutation. Then, given a slate, a rational user will choose its item with highest rank in the permutation. These two definitions of RUMs are equivalent (see, e.g., Chierichetti et al., 2018a).

MNLs and MNL Mixtures. A *Multinomial Logit* (aka, *MNL*) is a widely used kind of RUM. In an MNL L, one associates a positive weight $w_i > 0$ to each item $i \in [n]$.[2] To produce a random permutation, one samples the n elements, one after the other without replacement, with probability proportional to their respective MNL weights. It is not hard to see that, with this RUM, the probability that i wins in S is exactly $L_S(i) = \frac{w_i}{\sum_{j \in S} w_j}$, for each $i \in S \subseteq [n]$.

A *mixture of t MNLs*, also called a *mixed logit*, is given by a sequence $L^{(1)}, \ldots, L^{(t)}$ of MNLs and a mixture distribution p over $[t]$. To determine the winner of a slate S, we first sample $i \sim p$ and then use the MNL $L^{(i)}$ to sample the winner of S. Since a mixture of RUMs is a RUM, it also holds that a mixture of MNLs is a RUM (see, e.g., McFadden & Train, 2000).

Approximating RUMs. To define an approximation notion for RUMs, we first define a *distance* between RUMs D, D':

$$\begin{aligned}\mathrm{dist}(D, D') &= \max_{\varnothing \neq S \subseteq [n]} |D_S - D'_S|_{\mathrm{tv}} \\ &= \max_{\varnothing \neq S' \subset S \subseteq [n]} |D_S(S') - D'_S(S')| \,.\end{aligned}$$

I.e., the distance is the maximum, over the slates S, of the total variation distance of the winner distributions of S with D and D'. Equivalently, it is the maximum over $S' \subseteq S$ of the absolute difference of the probabilities, with D and D', that the random winner in slate S is in S'. E.g., if $\mathrm{dist}(D, D') \leq \epsilon$, S is a slate of movies, and $S' \subseteq S$ is its subset of dark comedies, then the probability that the movie chosen from S is a dark comedy changes by no more than ϵ from D to D'. Conversely, if $\mathrm{dist}(D, D') \geq \epsilon$, then there is a slate S and one of its subsets $S' \subseteq S$ such that the probability that the winner of S is in S' changes by at least ϵ from D to D'.

Since a RUM is a probability distribution over permutations, it can be represented by $O(n!)$ real numbers, each giving the probability of a permutation. Clearly, this representation is prohibitive. If one is allowed to approximate a RUM, is a more succinct representation possible?

3. An Efficient Representation

In this section we show that $O(n^2 \log n)$ bits are sufficient to approximately represent a RUM. The algorithm we will give to produce the representation will sample repeatedly the distribution D on $\mathbf{S}_n$ underlying the RUM.

Theorem 1. *Let $0 < \epsilon, \delta < 1$. There is a polynomial time algorithm that, given any distribution D on $\mathbf{S}_n$, produces a multiset M of $O\left(\epsilon^{-2} \cdot \left(n + \ln \delta^{-1}\right)\right)$ permutations such that, with probability at least $1 - \delta$, the uniform distribution $\widetilde{D}$ on M guarantees that $\mathrm{dist}(D, \widetilde{D}) \leq \epsilon$.*

Proof. The algorithm will first sample $t = \left\lceil \frac{n \cdot \ln 3 + \ln \frac{2}{\delta}}{2\epsilon^2} \right\rceil$ independent permutations $\pi_1, \ldots, \pi_t$ from D. After this first step, the algorithm fixes $\widetilde{D}$ to be the uniform distribution on the multiset of these samples, i.e., $\widetilde{D}$ chooses $i \in [t]$ uniformly at random (*u.a.r.*), and returns π_i.

We now prove that, with high probability, $\mathrm{dist}(D, \widetilde{D}) \leq \epsilon$. Consider any slate $S \subseteq [n]$, and any of its non-empty subsets $S' \subseteq S$. Let $\widetilde{D}_S(S') = \sum_{s \in S'} \widetilde{D}_S(s)$ be the probability that, using the distribution $\widetilde{D}$, the winner in the slate S belongs to S'. Then, $\widetilde{D}_S(S') \in [0, 1]$ is a random variable. Clearly, $\mathrm{E}\left[\widetilde{D}_S(S')\right] = \sum_{s \in S'} \mathrm{E}\left[\widetilde{D}_S(s)\right] = \sum_{s \in S'} D_S(s) = D_S(S')$, which is the probability that the winner in the slate S with distribution D is in S'. Since

$$|D_S - \widetilde{D}_S|_{\mathrm{tv}} = \max_{\varnothing \neq S' \subseteq S} |D_S(S') - \widetilde{D}_S(S')| \,,$$

the claim is proved if we show that, with probability at least $1 - \delta$, for each $\varnothing \neq S' \subseteq S \subseteq [n]$ it holds that $|D_S(S') - \widetilde{D}_S(S')| \leq \epsilon$. Indeed, by a Chernoff–Hoeffding bound (Hoeffding, 1963),

$$\begin{aligned}&\Pr\left[|D_S(S') - \widetilde{D}_S(S')| \geq \epsilon\right] \leq 2 \cdot e^{-2\epsilon^2 t} \\ &\leq 2 \cdot e^{-2\epsilon^2 \cdot \frac{n \ln 3 + \ln \frac{2}{\delta}}{2\epsilon^2}} = 2 \cdot e^{-n \ln 3 + \ln \frac{\delta}{2}} = \delta \cdot 3^{-n}.\end{aligned}$$

There are at most 3^n pairs of slates $S' \subseteq S \subseteq [n]$ (the generic item either belongs to S', or to $S \setminus S'$, or to $[n] \setminus S$), thus we get

$$\begin{aligned}&\Pr\left[\exists \varnothing \neq S' \subseteq S \subseteq [n] : |D_S(S') - \widetilde{D}_S(S')| \geq \epsilon\right] \\ &\leq \delta \cdot 3^{-n} \cdot 3^n = \delta.\end{aligned}$$

□

Note that the above upper bound improves the one in (Chierichetti et al., 2018a) from $O(n^2)$ to $O(n)$. An immediate consequence of this improvement is that any RUM can be approximately represented using $O(n^2 \log n)$ bits.

Corollary 2. *For each $0 < \epsilon < 1$, and for each RUM D, one can build a data structure using $O(\epsilon^{-2} \cdot n^2 \cdot \log n)$ bits that one can use to return, for each slate $S \subseteq [n]$, a distribution $\widetilde{D}_S$ satisfying $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \epsilon$.*

Proof. The algorithm of Theorem 1 provides, with probability at least $1/2$, a multiset of $O(n/\epsilon^2)$ permutations of $[n]$ such that the RUM $\widetilde{D}$ that chooses a permutation u.a.r. in the multiset, satisfies the approximation requirement of the statement. The generic permutation of $[n]$ can be represented with $O(n \log n)$ bits; thus, $\widetilde{D}$ can be represented with $O(\epsilon^{-2} \cdot n^2 \log n)$ bits. □

4. A Lower Bound on the Representation Size

In this section we prove an $\Omega(n^2)$ lower bound on the number of bits required to sketch a RUM. This shows that the

[2] In a machine learning setting, this weight is often the result of a linear or non-linear combination of features of the item, and is produced as the exponentiation of a computed logit.

representation obtained in Section 3 is near-optimal.

The cornerstone of our lower bound is a class of $2^{\Omega(n^2)}$ pairwise-distant RUMs that will be constructed randomly. Before introducing this class, we introduce the key notion of a *sieve RUM*.

Definition 3 (Sieve RUM)**.** *Given a sequence $\sigma = (S_1, \ldots, S_s)$ of sets such that $S_i \subseteq [n-s]$ for each $i \in [s]$, we define $D^{(\sigma)}$, the* sieve RUM with signature σ*, as follows:*

- *for $i \in [s]$, let $\pi_i \in \mathbf{S}_n$ have (i) the items of S_i in its top $|S_i|$ positions, sorted increasingly, (ii) item $n-s+i$ at position $|S_i|+1$, and (iii) the remaining items in the bottom $n-|S_i|-1$ positions, also sorted increasingly;*
- *the sieve RUM $D^{(\sigma)}$ will choose i u.a.r. from $[s]$, and will return permutation π_i.*

To build the class of RUMs, and prove our lower bound, we will independently sample exponentially many sieve RUMs from the following distribution.

Definition 4 (Random Sieve Distribution)**.** *Let s be given. For each $i \in [s]$, let S_i be a i.i.d. and u.a.r. subset of $[n-s]$. Let $\sigma = (S_1, \ldots, S_s)$. Return the sieve RUM $D^{(\sigma)}$ as in Definition 3.*

As a first step in our lower bound proof, we show that given two independent random sieve RUMs, with extremely high probability, there exists at least one slate where the two induced distributions are far in total variation distance. To do so, we will prove that (i) there exists a class of $\Theta(n)$ slates such that, for each slate S in that class, the probability that two random sieve RUMs have close winner distributions on S is $2^{-\Theta(n)}$ and (ii) the behavior of a random sieve RUM on a generic subclass is independent of its behavior on any disjoint subclass. Thus, the probability that two random sieve RUMs behave similarly on each of the $\Theta(n)$ slates in the class can be upper bounded by $\left(2^{-\Theta(n)}\right)^{\Theta(n)} = 2^{-\Theta(n^2)}$.

Our lower bound proof will then be concluded by a union bound argument: if one samples $2^{\Theta(n^2)}$ random sieve RUMs, with large enough probability any two of the sampled RUMs will behave dissimilarly on at least one slate.

Let $\mathrm{H}(x) = x \log_2 \frac{1}{x} + (1-x)\log_2 \frac{1}{1-x}$ denote the *binary entropy* of $x \in (0,1)$.

Lemma 5. *Suppose that D' and D'' are sieve RUMs sampled independently from the distribution in Definition 4, with $s = (1-\alpha)\cdot n$, for $0 < \alpha \leq 1/2$. Then,*

$$\Pr\left[\forall \varnothing \neq T \subseteq [n] : |D'_T - D''_T|_{\mathrm{tv}} \leq \frac{1-\alpha}{2}\right] \leq 2^{-\left(1-\mathrm{H}\left(\frac{1-\alpha}{2}\right)\right)\alpha(1-\alpha)n^2}.$$

Proof. We will in fact show that the event in the statement of the lemma holds, with small enough probability, even if we consider all and only the slates $T_1, \ldots, T_{n-s}$, where $T_k = \{k\} \cup ([n] \setminus [n-s])$ for $k \in [n-s]$.

Given the generic slate T_k and a generic signature $\sigma = (S_1, \ldots, S_s)$, for any $i \in [s]$, it holds that $n-s+i \in \mathrm{supp}(D^{(\sigma)}_{T_k})$ if and only if $k \notin S_i$. Recall that $[n] \setminus [n-s] \subseteq T_k$ and hence if π_j is sampled from the sieve RUM, one of its $|S_j|+1$ top-most items will be the winner: indeed, the $(|S_j|+1)$th item of π_j is $n-s+j$ (which is in T_k) and thus either that item or one of those preceding it (those of the set $S_j \subseteq [n-s]$), will be the winner.

Thus, for $n-s+i$ to be chosen in the slate T_k, it must hold that (i) π_i is sampled and (ii) $k \notin S_i$. In particular, if $k \in S_i$ we will have that $D^{(\sigma)}_{T_k}(n-s+i) = 0$; if $k \notin S_i$, then $D^{(\sigma)}_{T_k}(n-s+i) = 1/s$.

For a given σ, define the $(n-s)\times s$ matrix M_σ whose (k,i)th entry is $D^{(\sigma)}_{T_k}(n-s+i)$. Recall that the random sampling of $\sigma = (S_1, \ldots, S_s)$ is such that, for each $k \in [n-s]$ and $i \in [s]$, an independent fair coin flip determines whether $k \in S_i$. Thus, under our sampling of σ, each entry of M_σ will be chosen independently and u.a.r. in $\{0, 1/s\}$.

Let σ' (resp., σ'') be the (random) signature corresponding to D' (resp., D'') and let $M' = M_{\sigma'}$, $M'' = M_{\sigma''}$.

Pick any $k \in [n-s]$. Observe that

$$\begin{aligned} 2 \cdot \left|D'_{T_k} - D''_{T_k}\right|_{\mathrm{tv}} &= \sum_{i \in T_k} \left|D'_{T_k}(i) - D''_{T_k}(i)\right| \\ &= \left|D'_{T_k}(k) - D''_{T_k}(k)\right| + \delta_k, \end{aligned}$$

where we let

$$\delta_k = \sum_{i \in [n]\setminus[n-s]} \left|D'_{T_k}(i) - D''_{T_k}(i)\right|.$$

Hence, $\left|D'_{T_k} - D''_{T_k}\right|_{\mathrm{tv}} \geq \delta_k/2$. Now, consider the event

$$\xi_k = \text{“}\delta_k \leq 1-\alpha\text{”}.$$

Observe that $\left|D'_{T_k} - D''_{T_k}\right|_{\mathrm{tv}} \leq \frac{1-\alpha}{2}$ implies $\delta_k \leq 1-\alpha$, that is, it implies ξ_k. Moreover, whether ξ_k happens is a function of the kth rows of M' and M''. Since the entries of M' and M'' are chosen i.i.d., the events $\xi_1, \ldots, \xi_{n-s}$ are mutually independent.

Note also that for each $i \in [n]\setminus[n-s]$, $\left|D'_{T_k}(i) - D''_{T_k}(i)\right|$ is chosen i.i.d. and u.a.r. in $\{0, 1/s\}$, i.e., $s \cdot \delta_k \sim \mathrm{Bin}(s, \frac{1}{2})$.[3] Thus,

$$\begin{aligned} \Pr[\xi_k] &= \Pr\left[\mathrm{Bin}\left(s, \frac{1}{2}\right) \leq s \cdot \frac{1-\alpha}{2}\right] \\ &\leq 2^{-s} \cdot 2^{\mathrm{H}\left(\frac{1-\alpha}{2}\right)\cdot s} = 2^{-\left(1-\mathrm{H}\left(\frac{1-\alpha}{2}\right)\right)\cdot s}, \end{aligned}$$

since $\sum_{i=0}^{\lfloor \beta t \rfloor} \binom{t}{i} \leq 2^{t\cdot \mathrm{H}(\beta)}$ for $\beta \leq 1/2$.

[3] The binomial distribution $\mathrm{Bin}(n,p)$ has support $\{0, 1, \ldots, n\}$, and it is defined as $\Pr[\mathrm{Bin}(n,p) = k] = \binom{n}{k} p^k (1-p)^{n-k}$.

Putting these together, we have

$$\Pr\left[\forall \varnothing \neq T \subseteq [n] : |D'_T - D''_T|_{\mathrm{tv}} \leq \frac{1-\alpha}{2}\right]$$
$$\leq \Pr\left[\forall k \in [n-s] : \left|D'_{T_k} - D''_{T_k}\right|_{\mathrm{tv}} \leq \frac{1-\alpha}{2}\right]$$
$$\leq \Pr\left[\bigwedge_{k=1}^{n-s} \xi_k\right] = \prod_{k=1}^{n-s} \Pr[\xi_k] \leq 2^{-\left(1-\mathrm{H}\left(\frac{1-\alpha}{2}\right)\right)s(n-s)},$$

where we used the mutual independence of the events $\xi_1, \ldots, \xi_{n-s}$. The claim follows from $s = (1-\alpha)n$. □

We now use the strong probability guarantee of Lemma 5, and a union bound, to produce $2^{\Omega(n^2)}$ pairwise distant RUMs from the distribution of Definition 4.

Theorem 6. *For each $0 < \alpha \leq 1/2$, there is a set $\mathcal{D}$ of RUMs on $[n]$ such that (i) $|\mathcal{D}| = 2^{\left\lfloor \left(1-\mathrm{H}\left(\frac{1-\alpha}{2}\right)\right)\frac{\alpha(1-\alpha)}{2}\cdot n^2 \right\rfloor}$ and (ii) for each $\{D', D''\} \in \binom{\mathcal{D}}{2}$, $\mathrm{dist}(D', D'') > \frac{1-\alpha}{2}$.*

Proof. Let $\mathcal{D} = \left\{D^{(\sigma_1)}, \ldots, D^{(\sigma_t)}\right\}$ be a multiset of

$$t = 2^{\left\lfloor \left(1-\mathrm{H}\left(\frac{1-\alpha}{2}\right)\right)\frac{\alpha(1-\alpha)}{2}\cdot n^2 \right\rfloor},$$

RUMs sampled i.i.d. from the distribution in Definition 4.

We apply Lemma 5 on each of the $\binom{t}{2} \leq \frac{t^2}{2}$ pairs of sampled RUMs, together with a union bound, to obtain:

$$\Pr\left[\exists \{j, j'\} \in \binom{[t]}{2}, \forall \varnothing \neq T \subseteq [n] : \left|D_T^{(\sigma_j)} - D_T^{(\sigma_{j'})}\right|_{\mathrm{tv}} \leq \frac{1-\alpha}{2}\right]$$
$$\leq 2^{-(1-\mathrm{H}((1-\alpha)/2))\alpha(1-\alpha)n^2} \cdot t^2/2 \leq 1/2.$$

Thus, $\mathcal{D}$ has pairwise distinct elements (and is then a set) at distance larger than $\frac{1-\alpha}{2}$ from each other, with probability at least $1/2$ — i.e., with positive probability, $\mathcal{D}$ has properties (i) and (ii). □

We conclude by showing the representation lower bound: representing a RUM to within a maximum total variation distance bounded below $1/4$ requires $\Omega(n^2)$ bits.

Corollary 7. *Fix some $0 < \alpha \leq 1/2$. A data structure for a generic RUM D that can be used, for each slate $S \subseteq [n]$, to return a distribution $\widetilde{D}_S$ satisfying $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \frac{1-\alpha}{4}$, requires at least $\frac{\alpha^3}{6} \cdot n^2 - 1$ bits.*

Proof. It is well-known that $\mathrm{H}\left(\frac{1-x}{2}\right) \leq 1 - \frac{x^2}{\ln 4}$, for $x \in [-1, 1]$, (see, e.g., Calabro, 2009). Theorem 6 guarantees that, for each small enough α, there exists a set $\mathcal{D}$ of RUMs, such that for each $\{D, D'\} \in \binom{\mathcal{D}}{2}$, there exists a slate $T = T_{D,D'} \subseteq [n]$ such that $|D_T - D'_T|_{\mathrm{tv}} > \frac{1-\alpha}{2}$, and with

$$\lg_2 |\mathcal{D}| \geq \frac{\alpha(1-\alpha)}{2} \cdot \left(1 - \mathrm{H}\left(\frac{1-\alpha}{2}\right)\right) \cdot n^2 - 1$$
$$\geq \frac{\alpha(1-\alpha)}{2} \cdot \frac{\alpha^2}{\ln 4} \cdot n^2 - 1$$
$$\geq \frac{\alpha}{4} \cdot \frac{\alpha^2}{\ln 4} \cdot n^2 - 1 > \frac{\alpha^3}{6} \cdot n^2 - 1,$$

since $0 < \alpha \leq 1/2$ and $4 \ln 4 < 6$.

Let $D \in \mathcal{D}$. Suppose we have a data structure Δ that, for each slate $S \subseteq [n]$, can provide a distribution $\widetilde{D}_S$ over S, such that $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \frac{1-\alpha}{4}$. We show that Δ uniquely determines $D \in \mathcal{D}$. Indeed, for each $D' \in \mathcal{D} \setminus \{D\}$ there exists at least one slate $T = T_{D,D'} \subseteq [n]$ such that

$$\frac{1-\alpha}{2} < |D_T - D'_T|_{\mathrm{tv}} \leq |D_T - \widetilde{D}_T|_{\mathrm{tv}} + |\widetilde{D}_T - D'_T|_{\mathrm{tv}}$$
$$\leq \frac{1-\alpha}{4} + |\widetilde{D}_T - D'_T|_{\mathrm{tv}},$$

where the second inequality is the triangle inequality. The above inequalities then entail that $|\widetilde{D}_T - D'_T|_{\mathrm{tv}} > \frac{1-\alpha}{2} - \frac{1-\alpha}{4} = \frac{1-\alpha}{4}$. Hence, D is the only RUM of $\mathcal{D}$ that, for all slates S, guarantees $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \frac{1-\alpha}{4}$. It follows that Δ can be used to uniquely identify each RUM in $\mathcal{D}$. Thus, by counting, Δ uses at least $\lg_2 |\mathcal{D}| > \frac{\alpha^3 n^2}{6} - 1$ bits. □

5. The Efficiency of MNL Mixtures

In this section we study the question of how well succinct MNL mixtures can approximate a RUM. We start by observing that the distribution of winners of a permutation can be well-approximated by those of an MNL.

Observation 8. *Let D be a RUM supported on a single permutation. Then, for each $0 < \epsilon < 1$, there exists an MNL L such that $\mathrm{dist}(L, D) \leq \epsilon$.*

Proof. Let $\mathrm{supp}(D) = \{\pi\}$. We construct an MNL L by assigning a weight of ϵ^r to the item with rank r in π, for each $r \in [n]$. For any slate $S \subseteq [n]$, let $i = \pi(S)$ be the top element of S with the ordering of π; then, $D_S(i) = 1$. If r is the rank of i in π, then

$$L_S(i) \geq \frac{\epsilon^r}{\sum_{j=r}^{\infty} \epsilon^j} = \frac{1}{\sum_{j=0}^{\infty} \epsilon^j} = \frac{1}{\frac{1}{1-\epsilon}} = 1 - \epsilon.$$

Since $D_S(i) = 1$, we have $|L_S - D_S|_{\mathrm{tv}} \leq \epsilon$. □

Theorem 1 and Observation 8 immediately yield:

Corollary 9. *For each RUM D, there exists a uniform mixture L of $O(n/\epsilon^2)$ MNLs such that $\mathrm{dist}(L, D) \leq \epsilon$.*

In the remainder of this section we will prove an almost matching lower bound: one cannot approximate a generic RUM to within some constant error using a mixture of only $o(n/\log n)$ MNLs. There are two ingredients in this proof: the representational lower bound for RUMs (Corollary 7) and a compression result for MNL mixtures that we show below (Theorem 12). The latter is of independent interest.

To show MNL mixtures can be compressed we proceed as

follows. First, we show that the weights in any MNL can be reduced to weights whose consecutive ratios can be represented with $O(\log n)$ bits each, at the cost of a small error in the winner distributions. Consequently, any MNL can be approximated by a representation that uses only $O(n \log n)$ bits. Then, we use this result to show that a mixture of t MNLs can be approximated using only $O(nt \log n)$ bits.

The argument is concluded by applying Corollary 7 which ensures that, for a mixture of t MNLs to approximate a generic RUM, one must have $nt \log n \geq \Omega(n^2)$, and thus $t \geq \Omega\left(\frac{n}{\log n}\right)$.

We begin by proving our main MNL compression result.

Theorem 10. *For any MNL L and for each $0 < \epsilon \leq 1$, there is an MNL $\widetilde{L}$ that can be represented with $O(n \log \frac{n}{\epsilon})$ bits such that $\mathrm{dist}(L, \widetilde{L}) \leq \epsilon$.*

Proof. Let $[n] = \{i_1, \ldots, i_n\}$ and assume w.l.o.g. that $1 = w_{i_1} \leq \cdots \leq w_{i_n}$ are the weights of the items in the MNL L. For each $j = 2, \ldots, n$, define $\rho_j = \left\lceil \log_{1+\frac{\epsilon}{2n}} \min\left(\frac{w_{i_j}}{w_{i_{j-1}}}, \frac{2n}{\epsilon}\right)\right\rceil$. Note that ρ_j is a non-negative integer of value at most $O\left(\log_{1+\frac{\epsilon}{2n}} \frac{n}{\epsilon}\right) = O\left(\frac{n}{\epsilon} \log \frac{n}{\epsilon}\right)$ and hence can be represented using $O\left(\log \frac{n}{\epsilon}\right)$ bits. Thus, the full sequence of $\rho_2, \ldots, \rho_n$, and the ordering $i_1, i_2, \ldots, i_n$, can be represented with $O(n \log \frac{n}{\epsilon})$ bits.

We define the MNL $\widetilde{L}$ using $\rho_2, \ldots, \rho_n$ and $i_1, \ldots, i_n$: the weight of item $i_j \in [n]$ in $\widetilde{L}$ is $\widetilde{w}_{i_j} = (1 + \frac{\epsilon}{2n})^{\sum_{t=2}^{j} \rho_t}$. To prove $\mathrm{dist}(L, \widetilde{L}) \leq \epsilon$, we need two key properties of these new weights, stated next.

Lemma 11. *(i) If for some $j < j'$ it holds $w_{i_j}/w_{i_{j'}} < \frac{\epsilon}{2n}$, then $\widetilde{w}_{i_j}/\widetilde{w}_{i_{j'}} < \frac{\epsilon}{2n}$. (ii) If for some $j < j'$ it holds $w_{i_j}/w_{i_{j'}} > \frac{\epsilon}{2n}$, then $\left(1 - \frac{\epsilon}{2}\right) \cdot \frac{w_{i_j}}{w_{i_{j'}}} \leq \frac{\widetilde{w}_{i_j}}{\widetilde{w}_{i_{j'}}} \leq \frac{w_{i_j}}{w_{i_{j'}}}$.*

Now, consider a slate $S \subseteq [n]$ with $S = \{s_1, \ldots, s_{|S|}\}$ and $w_{s_1} \leq \cdots \leq w_{s_{|S|}}$. We aim to show $|L_S - \widetilde{L}_S|_{\mathrm{tv}} \leq \epsilon$.

If $|S| = 1$, then $L_S = \widetilde{L}_S$ and $|L_S - \widetilde{L}_S|_{\mathrm{tv}} = 0$. Otherwise, let $j^* \geq 1$ be the smallest integer such that $w_{s_{j^*}} \geq \frac{\epsilon}{2n} \cdot w_{s_{|S|}}$. Now, we write $|L_S - \widetilde{L}_S|_{\mathrm{tv}} = (\alpha_S + \beta_S)/2$, where

$$\alpha_S = \sum_{j=1}^{j^*-1} |L_S(s_j) - \widetilde{L}_S(s_j)|, \tag{1}$$

$$\beta_S = \sum_{j=j^*}^{|S|} |L_S(s_j) - \widetilde{L}_S(s_j)|. \tag{2}$$

We begin by upper bounding α_S. First,

$$\sum_{j=1}^{j^*-1} L_S(s_j) \leq \sum_{j=1}^{j^*-1} L_{\{s_j, s_{|S|}\}}(s_j) = \sum_{j=1}^{j^*-1} \frac{w_{s_j}}{w_{s_{|S|}} + w_{s_j}}$$

$$\leq \sum_{j=1}^{j^*-1} \frac{w_{s_j}}{w_{s_{|S|}}} < \sum_{j=1}^{j^*-1} \frac{\epsilon}{2n} < \frac{\epsilon}{2}, \tag{3}$$

and next,

$$\sum_{j=1}^{j^*-1} \widetilde{L}_S(s_j) \leq \sum_{j=1}^{j^*-1} \widetilde{L}_{\{s_j, s_{|S|}\}}(s_j)$$

$$\leq \sum_{j=1}^{j^*-1} \frac{\widetilde{w}_{s_j}}{\widetilde{w}_{s_{|S|}}} < \sum_{j=1}^{j^*-1} \frac{\epsilon}{2n} < \frac{\epsilon}{2}, \tag{4}$$

where the penultimate step is from the definition of j^* and by applying Lemma 11(i). Using (1), (3), (4), and by applying the triangle inequality, we obtain $\alpha_S < \epsilon$.

We now upper bound β_S. To do this, we write

$$\widetilde{L}_S(s_j) = \frac{\widetilde{w}_{s_j}}{\sum_{\ell=1}^{|S|} \widetilde{w}_{s_\ell}} = \frac{\widetilde{w}_{s_j}/\widetilde{w}_{s_{|S|}}}{\sum_{\ell=1}^{|S|} (\widetilde{w}_{s_\ell}/\widetilde{w}_{s_{|S|}})},$$

and apply Lemma 11(ii) to get upper and lower bounds

$$\widetilde{L}_S(s_j) \geq \frac{\left(1 - \frac{\epsilon}{2}\right) \cdot \frac{w_{s_j}}{w_{s_{|S|}}}}{\sum_{\ell=1}^{|S|} \frac{w_{s_\ell}}{w_{s_{|S|}}}} = \left(1 - \frac{\epsilon}{2}\right) \cdot \frac{w_{s_j}}{\sum_{\ell=1}^{|S|} w_{s_\ell}},$$

and

$$\widetilde{L}_S(s_j) \leq \frac{w_{s_j}/w_{s_{|S|}}}{\left(1 - \frac{\epsilon}{2}\right) \cdot \sum_{\ell=1}^{|S|} \frac{w_{s_\ell}}{w_{s_{|S|}}}} = \frac{1}{1 - \frac{\epsilon}{2}} \cdot \frac{w_{s_j}}{\sum_{\ell=1}^{|S|} w_{s_\ell}}.$$

Using these, we obtain

$$|L_S(s_j) - \widetilde{L}_S(s_j)|$$

$$\leq \max\left(1 - \left(1 - \frac{\epsilon}{2}\right), \frac{1}{1 - \frac{\epsilon}{2}} - 1\right) \cdot \frac{w_{s_j}}{\sum_{\ell=1}^{|S|} w_{s_\ell}}$$

$$= \frac{\epsilon/2}{1 - \epsilon/2} \cdot \frac{w_{s_j}}{\sum_{\ell=1}^{|S|} w_{s_\ell}} \leq \epsilon \cdot \frac{w_{s_j}}{\sum_{\ell=1}^{|S|} w_{s_\ell}},$$

where we used $\epsilon \in (0, 1)$. Now,

$$\beta_S = \sum_{j=j^*}^{|S|} |L_S(s_j) - \widetilde{L}_S(s_j)| \leq \epsilon \cdot \frac{\sum_{j=j^*}^{|S|} w_{s_j}}{\sum_{\ell=1}^{|S|} w_{s_\ell}} \leq \epsilon.$$

Finally, $|L_S - \widetilde{L}_S|_{\mathrm{tv}} = (\alpha_S + \beta_S)/2 \leq \epsilon$. □

We next show that a mixture of t MNLs can be approximated by a mixture that admits an $O(tn \log \frac{n}{\epsilon})$ bit representation.

Theorem 12. *For any mixture L of t MNLs and for each $0 < \epsilon \leq 1$, there is a mixture $\widetilde{L}$ of t MNLs that can be represented with $O(tn \log \frac{n}{\epsilon})$ bits and $\mathrm{dist}(L, \widetilde{L}) \leq \epsilon$.*

Using these, we can now prove the lower bound on the size of a mixture of MNLs that approximates a RUM.

Corollary 13. *Fix any small enough constant $\alpha > 0$. There is a RUM D such that for each mixture L of $o(n/\log n)$ MNLs, it holds that $\mathrm{dist}(L, D) > \frac{1-2\alpha}{4}$.*

Proof. Let L be a generic mixture of $t = o(n/\log n)$ MNLs. By Theorem 12, there is a mixture $\widetilde{L}$ such

that $\mathrm{dist}(L, \widetilde{L}) \leq \alpha/4$ and $\widetilde{L}$ can be represented with $O(tn \log \frac{n}{\alpha}) = o(n^2)$ bits.

By contradiction, suppose that for each RUM D, one could find a mixture $L^{(D)}$ of $o(n/\log n)$ MNLs that approximates D to within total variation distance $\frac{1-2\alpha}{4}$ on each slate S. Then, one would be able to approximate D_S to within total variation distance $\frac{1-2\alpha}{4} + \frac{\alpha}{4} = \frac{1-\alpha}{4}$ for each slate $S \subseteq [n]$ by storing only the representation of $\widetilde{L}^{(D)}$, i.e., $o(n^2)$ bits. But, this contradicts Corollary 7 and hence it is impossible that the mixture $L^{(D)}$ exists for each RUM D. □

6. RUMs vs MNL Mixtures

In this section we compare RUMs and MNL mixtures having supports of the same size. We will see that, for large supports, they are near-equivalent. On the other hand, small mixtures of MNLs are more powerful than small-support RUMs.

Earlier, we have shown that a RUM supported on $O(n)$ permutations (Theorem 1), and a mixture of $O(n)$ MNLs (Corollary 9), are both sufficient to approximate any RUM. We also proved that no mixture of $o(n/\log n)$ MNLs can approximately represent a generic RUM (Corollary 13).

We begin this section by proving that there are simple RUMs that can only be approximated by RUMs with $\Omega(n)$ support. In particular, let U be the uniform RUM, i.e., the RUM that chooses a permutation u.a.r. from $\mathbf{S}_n$. Clearly $U_S(i) = 1/|S|$ for any slate $S \subseteq [n]$ and for each $i \in S$.[4]

Theorem 14. *For any RUM $\widetilde{U}$ with $|\mathrm{supp}(\widetilde{U})| = o(n)$, it holds that $\mathrm{dist}(U, \widetilde{U}) \geq 1 - o(1)$.*

Thus, in the worst case, both RUMs and MNL mixtures require a support of size $\widetilde{\Theta}(n)$ to approximately represent the generic RUM. This equivalence, though, does not translate uniformly to the whole space of RUMs. In particular, U is equivalent to an MNL that assigns the same weight of 1 to each item. Thus, U is a RUM that can be perfectly represented with a *single* MNL, but that can only be approximated by RUMs having $\Omega(n)$ permutations in their support. Conversely, by Observation 8, any RUM supported on k permutations can be approximately represented by a mixture of k MNLs.

7. The Case of Small Slates

In many practical settings, only the behavior of the RUM on slates of small sizes matters. In this section we consider this case and extend many of our results. First, we modify the distance notion to focus on small slates:

$$\mathrm{dist}_k(D, D') = \max_{\varnothing \neq S \subseteq [n], |S| \leq k} |D_S - D'_S|_{\mathrm{tv}}\,.$$

We show that in order to approximate the winner distributions but only for slates of size at most k, the number of permutations needed can be shrunk from $\Theta(n)$ to $O(k \log n)$.

Theorem 15. *Let $0 < \epsilon, \delta < 1$. There is a polynomial time algorithm that, given any distribution D on $\mathbf{S}_n$, produces a multiset M of $O\left(\epsilon^{-2} \cdot \left(k \log n + \ln \delta^{-1}\right)\right)$ permutations such that, with probability at least $1 - \delta$, the uniform distribution D' on M guarantees that $\mathrm{dist}_k(D, D') \leq \epsilon$.*

The following is immediate and analogous to Corollary 2.

Corollary 16. *For each $0 < \epsilon < 1$ and for each RUM D, one can build a data structure using $O(\epsilon^{-2} \cdot nk \log^2 n)$ bits that one can use to return, for each slate $S \subseteq [n]$ with $|S| \leq k$, a distribution $\widetilde{D}_S$ satisfying $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \epsilon$.*

We conclude with a near-matching $\Omega(nk)$ bit lower bound.

Theorem 17. *Fix some $0 < \alpha \leq 1/2$. A data structure for any RUM D that can be used, for each slate $S \subseteq [n]$ with $|S| \leq k$, to return a distribution $\widetilde{D}_S$ satisfying $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \frac{1-\alpha}{4}$, requires at least $(1 - O(k^{-2}))\frac{\alpha^3}{6} nk$ bits.*

8. Experimental Results

The experiments were coded in Python and used IBM cplex.[5] We ran them on a 8-Core i9 MacBook Pro with 64GiB of RAM; the total running time of all runs of all experiments was under two hours.

Approximation vs Size. In the first experiment, we aim to understand the relation between the maximum (and the average) total variation distance of the approximating RUM and the number of permutations in its support.

We consider the Sushi 3A dataset (Kamishima, 2003). The dataset is composed of 5,000 permutations of $n = 10$ fixed types of sushi, where the ith permutation represents the user i's preference order of the 10 types. A uniform distribution on the dataset defines a RUM D. To compress D, for each $t \in [25]$, we sample i.i.d. $t \cdot n$ permutations from D and produce a RUM $\widetilde{D}^{(t)}$ as in Theorem 1.

We computed two errors for each such $\widetilde{D}^{(t)}$: the maximum and the average total variation distance $|D_S - \widetilde{D}^{(t)}_S|_{\mathrm{tv}}$ over all slates S. The results, each averaged over 1,000 runs of the sampling algorithm, are in Figure 1. The error, averaged over all the slates, is below 0.1 already with $5n = 50$ samples (i.e., 1% of the original data). The maximum error over the slates is below 0.1 already with $23n$ samples (4.6% of the original data).

[4] We point out that the distribution over permutations of U is the one supporting the min-hash (or *shingles*) sketch (Broder, 1997).

[5] www.ibm.com/analytics/cplex-optimizer

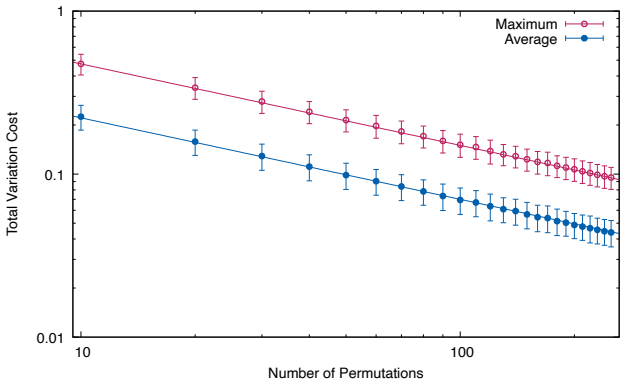

Figure 1. The results for Sushi 3A. The data points represent the costs averaged over 1,000 runs of the sampling algorithm; the error bars represent one standard deviation. The solid lines are power laws with exponent $-1/2$. The $c\epsilon^{-2}n$ prediction of Theorem 1 ($x = cy^{-2}n$, or $y = \sqrt{cn/x}$) fits the data quite precisely.

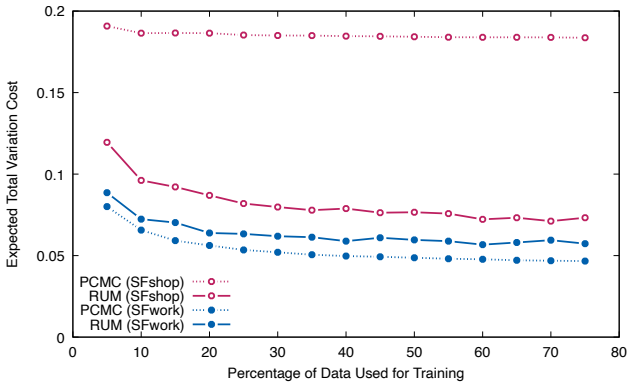

Figure 2. The expected total variation distance on the SFshop and SFwork datasets (averaged over 100 random partitions of the datasets) between M^{te} and the RUM model produced by our LP, and between M^{te} and the PCMC model of Ragain & Ugander (2021), as the percentage of the dataset used for training ranges from 5% to 75%.

Choices Representation. In our second experiment, we compare the quality of the predictions given by a RUM representation with that of the PCMC model[6] of Ragain & Ugander (2016; 2021), for the SFwork and the SFshop datasets (Koppelman & Bhat, 2006). These datasets represent the choices between transportation alternatives made by people that were to travel to and from their workplace (SFwork), and a shopping center (SFshop). SFwork contains 5,029 events, each of which is composed of a slate shown to a user, together with the user's choice in that slate; here, $n = 6$. SFshop is similar, with 3,157 events and $n = 8$.

To measure the quality of a prediction, we follow (Ragain & Ugander, 2021), and use the expected total variation distance[7], which we now define. Given a multiset $M = \{(S_1, i_1), \ldots, (S_t, i_t)\}$, with $i_j \in S_j$, of (slate, winner) pairs, one computes the empirical distribution μ_M over the slates as follows: $\mu_M(S)$ equals the fraction of the pairs of M that have S as their slate. Then, given the same M and a slate S such that $\mu_M(S) > 0$, one computes the empirical distribution of the winner of S: $M_S(i)$ is the ratio of the number of pairs of M that have S as their slate and i as their winner, to the number of pairs of M that have S as their slate. To test the quality of a model D against M, we compute the expected total variation distance:

$$\text{dist}_M(D) = E_{S\sim\mu_M}\left[|D_S - M_S|_{\text{tv}}\right].$$

[6] A PCMC model is defined by an $n \times n$ matrix Q representing a continuous time Markov chain, with $Q_{i,j} + Q_{j,i} > 0$ for each $\{i, j\} \in \binom{[n]}{2}$. Given a slate S, the distribution of the winner of S is the (unique) stationary distribution of the continuous-time Markov chain on state space S and transition rates $q_{i,j} = Q_{i,j}$ for each $i \in S$ and $j \in S \setminus \{i\}$.

[7] Ragain & Ugander (2021) plot the expected ℓ_1 distance, which is exactly twice the expected total variation distance.

As in (Ragain & Ugander, 2021), for each dataset M, we split the events: a u.a.r. part M^{tr} of the dataset, conditioned on having a fixed size (in the 5%–75% range), was used for training; a u.a.r. part M^{te} of the dataset disjoint from M^{tr}, conditioned on having size 25%, was used for testing.

We use M^{tr} to produce the RUM $\widetilde{D}$ that minimizes the expected total variation distance $\text{dist}_{M^{\text{tr}}}(\widetilde{D})$ to the training data M^{tr}, using the following linear program (LP):

$$\begin{cases} \min \frac{1}{2} \cdot \sum_{S\subseteq[n]} \sum_{i\in S} (\mu_{M^{\text{tr}}}(S) \cdot \delta_{S,i}) \\ -\delta_{S,i} \le M^{\text{tr}}_S(i) - \sum_{\substack{\pi\in\mathcal{S}_n \\ \pi(S)=i}} p_\pi \le \delta_{S,i} \quad \forall i \in S \in \text{supp}(\mu_{M^{\text{tr}}}) \\ \sum_{\pi\in\mathcal{S}_n} p_\pi = 1 \\ p_\pi \ge 0 \qquad \forall \pi \in \mathbf{S}_n \end{cases}$$

The solution to this LP directly gives a RUM $\widetilde{D}$: the RUM will sample a permutation π with probability p_π.

After having computed the best RUM for M^{tr}, we test it on M^{te}. Figure 2 plots, for each dataset, the $\text{dist}_{M^{\text{te}}}(\widetilde{D})$ of our RUM and the $\text{dist}_{M^{\text{te}}}(P)$ of the PCMC model P of Ragain & Ugander (2021), also trained on M^{tr}. Each point represents the expected total variation distance, averaged over 100 random partitions of the dataset.

For SFshop, approximating with a RUM gives a much better error than that obtained with PCMC (the average of the ratios of the PCMC error and of the RUM error is ≈ 2.29). For SFwork, the RUM gives a slightly worse approximation than the PCMC one (the average of the PCMC/RUM error ratios is ≈ 0.84). Thus RUM is competitive against PCMC, representation-wise; also, the RUMs we learned have a smaller expected total variation error than that of the

mixture of MNL models of Ragain & Ugander (2021) for both datasets (see their Figure 2, and the Figure 2 of their Supplementary Material; recall that the distance they plot is twice ours).

When using the full dataset for both training and testing, the RUM representations given by the LP have a dist_M error of 0.026 for SFwork[8], and a dist_M error of 0.027 for SFshop[9].

9. Related Work

Discrete choice theory is a well-established research topic in economics; see the excellent book by Train (2003). We only cover work that is directly relevant to our focus.

Farias et al. (2009) considered the problem of approximating the choices made by users with a RUM model on bounded support. They proposed an ℓ_0-minimization formulation of the problem, and showed that it can be optimized efficiently, under assumptions on its optimum. Our positive results can be seen as robust versions of their conditional result.

Several papers (Chierichetti et al., 2018a;b; Negahban et al., 2018; Oh & Shah, 2014; Soufiani et al., 2012; Tang, 2020) have considered the problem of learning the behavior of a general RUM or of restricted RUM models, in both the passive and active learning settings. Our work, on the other hand, addresses the representation complexity of RUMs.

Very recently, some deterministic choice models have been considered in the ML literature (Rosenfeld et al., 2020). These models, while interesting, provably cannot approximate RUMs. In fact, in the Supplementary Material, we show that deterministic models, as well as the models of (Ragain & Ugander, 2016; Seshadri et al., 2019), provably cannot represent general RUMs.

10. Conclusions

In this paper we consider the representational complexity of approximating an arbitrary RUM on n items. We obtain near-optimal bounds of $\widetilde{\Theta}(n^2)$ bits needed to represent any RUM. We also show a similarly tight bound of $\widetilde{\Theta}(n)$ on the size of MNL mixtures for approximating the generic RUM.

Besides the immediate question of how to close the small gap between our upper and lower bounds, our work opens up other research avenues. For example, it would be interesting to consider the ℓ_∞ version of our question, i.e., if we only want to ϵ-additively approximate the probability that the generic item wins in the generic slate. While our upper bound applies here, the lower bound does not. Another question: can a succinct representation be learned with a deep network?

[8]This RUM model for SFwork is supported on $t = 16$ permutations. The model can then be represented with $t \cdot (n-1) = 80$ integers (used for storing the permutations in the RUM's support) and $t - 1 = 15$ floating point numbers (used for storing the probabilities that the RUM assigns to those t permutations).

[9]This RUM model for SFshop is supported on $t = 11$ permutations. The model can then be represented with $t \cdot (n-1) = 77$ integers and $t - 1 = 10$ floating point numbers.

Acknowledgements

We thank Johan Ugander and the anonymous reviewers for their comments and suggestions.

Flavio Chierichetti was supported in part by the PRIN project 2017K7XPAN, by a Google Focused Research Award, by BiCi — Bertinoro international Center for informatics, and by the "Dipartimenti di Eccellenza" grant awarded to the Dipartimento di Informatica at Sapienza.

References

Broder, A. Z. On the resemblance and containment of documents. In *SEQUENCES*, pp. 21–29, 1997.

Calabro, C. *The Exponential Complexity of Satisfiability Problems*. PhD thesis, UCSD, 2009.

Chierichetti, F., Kumar, R., and Tomkins, A. Discrete choice, permutations, and reconstruction. In *SODA*, pp. 576–586, 2018a.

Chierichetti, F., Kumar, R., and Tomkins, A. Learning a mixture of two multinomial logits. In *ICML*, pp. 961–969, 2018b.

Farias, V. F., Jagabathula, S., and Shah, D. A data-driven approach to modeling choice. In *NIPS*, pp. 504–512, 2009.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *JASA*, 58, 1963.

Kamishima, T. Nantonac collaborative filtering: Recommendation based on order responses. In *KDD*, pp. 583–588, 2003.

Koppelman, F. S. and Bhat, C. A self instructing course in mode choice modeling: Multinomial and nested logit models, 2006. US Department of Transportation, Federal Transit Administration.

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

Negahban, S., Oh, S., Thekumparampil, K. K., and Xu, J. Learning from comparisons and choices. *JMLR*, 19(1): 1478–1572, 2018.

Oh, S. and Shah, D. Learning mixed multinomial logit model from ordinal data. In *NIPS*, pp. 595–603, 2014.

Ragain, S. and Ugander, J. Pairwise choice Markov chains. In *NIPS*, pp. 3198–3206, 2016.

Ragain, S. and Ugander, J. Pairwise choice Markov chains. In *arXiv:1603.02740*, 2021.

Rosenfeld, N., Oshiba, K., and Singer, Y. Predicting choice with set-dependent aggregation. In *ICML*, pp. 8220–8229, 2020.

Seshadri, A., Peysakhovich, A., and Ugander, J. Discovering context effects from raw choice data. In *ICML*, pp. 5660–5669, 2019.

Seshadri, A., Ragain, S., and Ugander, J. Learning rich rankings. In *NeurIPS*, 2020.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Soufiani, H. A., Parkes, D. C., and Xia, L. Random utility theory for social choice. In *NIPS*, pp. 126–134, 2012.

Tang, W. Learning an arbitrary mixture of two multinomial logits. *arXiv*, 2007.00204, 2020.

Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003.

Yang, Z., Dai, Z., Salakhutdinov, R., and Cohen, W. W. Breaking the softmax bottleneck: A high-rank RNN language model. In *ICLR*, 2018.