Reconciling the Accuracy-Diversity Trade-off in Recommendations

Kenny Peng
Cornell Tech
New York, NY, USA
kennypeng@cs.cornell.edu

Manish Raghavan
Massachusetts Institute of Technology
Cambridge, MA, USA
mragh@mit.edu

Emma Pierson
Cornell Tech
New York, NY, USA
emma.pierson@cornell.edu

Jon Kleinberg*
Cornell University
Ithaca, NY, USA
kleinberg@cornell.edu

Nikhil Garg*
Cornell Tech
New York, NY, USA
ngarg@cornell.edu

ABSTRACT

When making recommendations, there is an apparent trade-off between the goals of *accuracy* (to recommend items a user is most likely to want) and *diversity* (to recommend items representing a range of categories). As such, real-world recommender systems often explicitly incorporate diversity into recommendations, at the cost of accuracy.

We study the accuracy-diversity trade-off by bringing in a third concept: user utility. We argue that accuracy is misaligned with user utility because it fails to incorporate a user's consumption constraints: at any given time, users can typically only use at most a few recommended items (e.g., dine at one restaurant, or watch a couple of movies). In a theoretical model, we show that utility-maximizing recommendations—when accounting for consumption constraints—are naturally diverse due to diminishing returns of recommending similar items. Therefore, while increasing diversity may come at the cost of accuracy, it can also help align accuracy-based recommendations toward the more fundamental objective of user utility. Our theoretical results yield practical guidance into how recommendations should incorporate diversity to serve user ends.

CCS CONCEPTS

• **Information systems** → **Information retrieval diversity**; **Recommender systems**.

KEYWORDS

Accuracy-diversity trade-off, recommender systems, theoretical modeling

*Joint Senior Author.

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05…$15.00
https://doi.org/10.1145/3589334.3645625

ACM Reference Format:
Kenny Peng, Manish Raghavan, Emma Pierson, Jon Kleinberg, and Nikhil Garg. 2024. Reconciling the Accuracy-Diversity Trade-off in Recommendations. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3589334.3645625

1 INTRODUCTION

Machine learning-based approaches in recommender systems are well-suited to optimize for *accuracy*, the percentage of recommended items a user likes. At the same time, recommender systems are often designed to incorporate *diversity*, since users empirically prefer to be shown items from a range of categories [7, 37, 42, 52]. In practice, however, these goals are in tension. To counter accuracy-maximization's tendency toward homogeneity, real-world recommender systems inject diversity into recommendations using a range of heuristics [23, 40]. A wide literature proposes methods to address the apparent "accuracy-diversity trade-off" (e.g., [1, 3, 5, 6, 8, 14, 18, 19, 22, 24, 25, 27, 28, 31, 34, 36, 38, 39, 44, 47, 53, 56, 58, 59]).

Despite the trade-off's practical importance, a deeper understanding of the underlying tension is missing. Without a principled understanding of the trade-off, attempts to diversify recommendations have difficulty moving beyond intuition—and difficulty articulating what they are accomplishing at a deeper level.

Reconciling the trade-off. In this paper, we conceptualize and analyze the relationship between accuracy and diversity by observing that there are in fact three fundamentally distinct quantities of interest in these problems: accuracy, diversity, and user utility. None of these quantities serves as a proxy for any other, and a true understanding of the trade-offs requires understanding how all three of them interact. Moreover, the heart of the problem is really about maximizing one of these three—namely user utility, since this is what users actually experience. In particular, we argue that accuracy, in general, misrepresents user utility, and that by considering a better-conceived measure of utility, the trade-off with diversity dissipates. In turn, our results inform how diversity supports utility-maximizing recommendations.

Accuracy does correspond directly to a model of user utility in a very specific situation: when users obtain value from *all* recommended items—that is, with binary utility, value 1 for each item liked and 0 for each disliked. We argue that under more reasonable and general assumptions, however, accuracy is misaligned with

user utility because it does not consider *consumption constraints*—limits on the number of recommended items a user can use. At a given time, a user can only watch one movie, dine at one restaurant, or purchase one new TV; a job recruiter can only extend interviews to a handful of recommended candidates.

A more precise measure of user utility accounts for consumption constraints and therefore focuses on the value the user obtains from the *best* items they are recommended. Given a "unit consumption constraint" and binary value for each item, this reduces to the probability that the user is shown *at least one* item they like. Once accounting for consumption constraints, we show that user utility is in fact aligned with and supported by diversity; a preference for diversity arises endogenously in our model. As a consequence, efforts to navigate the accuracy-diversity trade-off can be viewed not as balancing two genuinely competing desiderata, but rather as using diversity to steer accuracy-maximizing recommendations towards utility-maximizing recommendations.

To see why consumption constraints induce diversity, consider a thought experiment by Steck [54] in the design of recommendations. A user on a movie-streaming service is in the mood for comedy 80 percent of the time, and action 20 percent of the time. How many movies of each genre should we recommend? The items the user will like with the highest probability are mostly comedy, meaning that the accuracy-maximizing set of recommendations is likely to be fairly homogeneous. But now suppose that we aim to maximize the probability the user likes at least one recommended movie—accounting for a "unit consumption constraint." Now, recommending only comedy movies is suboptimal. If the user is in the mood for action, they will be left without any options; meanwhile, it is not beneficial to recommend many additional comedy movies, since the user only cares that they have a single good movie to watch. In this way, accounting for consumption constraints intuitively induces diverse recommendations.

A model of recommendations. To make this intuition precise—that user utility is aligned with diversity—and to understand when and to what extent it holds, we need to analyze the diversity of accuracy- and utility-maximizing recommendations. (From here on, utility refers to a measure that accounts for consumption constraints.) A primary contribution of our work is a stylized-but-rich model of recommendations that is analytically tractable in this respect.

In our model, items of varying quality belong to discrete types, and each user has a probability distribution over types. In a given session, the user is in the mood for one of these types, where the type is drawn from the distribution. (Uncertainty of a user's mood can arise either due to genuine stochasticity, or limitations in the recommender's inferential abilities.) This model lends itself to an interpretable measure of diversity, where a set of recommendations is diverse if it represents items from many types roughly equally.

We derive results in an asymptotic regime where the number of recommendations grows large, obtaining precise characterizations of accuracy- and utility-maximizing recommendations as a function of model parameters that control the quality of items within and across types. We show in computational experiments that our theoretical findings hold more generally.

Diminishing returns drive our results and proof technique. Our results connect the composition of recommendation sets with the rate of diminishing returns when recommending more items of a given type (with respect to accuracy or utility). With large diminishing returns in one type, after recommending a few items of that type, a recommender becomes incentivized to recommend other types. Roughly speaking, utility induces sharper diminishing returns than accuracy, and thus more diversity. The key steps in our proofs are to (1) precisely characterize the asymptotic behavior of these diminishing returns under different parameterizations of our model, and (2) to show how this behavior determines the asymptotic representation of item types in optimal recommendations.

Overview of results. In a basic setting (Theorem 1), the model confirms our intuition in a striking way. In this setting, accuracy-maximizing recommendations are completely homogeneous (representing only items from one type); yet, by accounting for consumption constraints, utility-maximizing recommendations are completely diverse (representing each type with equal proportion) in the limit. This uncovers a surprising fact—that even if the user prefers one type with higher probability than another, the optimal set of recommendations may contain an equal proportion of each.

In a more general setting (Theorems 2a and 2b), we consider differences in item quality within and across types, accounting for the idiosyncratic properties of recommendation settings.

Theorem 2a shows that accuracy-maximizing recommendations become more diverse when item quality decays at a faster rate (i.e., the recommender quickly begins to run out of "good options"). This accords with our conceptual understanding that larger diminishing returns induce more diversity. However, when this rate of decay is "reasonable" (in a sense made precise in Section 3.2), accuracy-maximizing recommendations remain relatively homogeneous—roughly speaking, they represent types "less than proportionally."

Theorem 2b shows that utility-maximizing recommendations are generally diverse. More specifically, however, we show that when there is no decay in item quality, the representation of a type varies *inversely* with the quality of items within that type. This holds empirically whenever the rate of decay is small. While perhaps counterintuitive, this is explained by the need to recommend more items from such a type to ensure that the user likes at least one. We isolate this case in Corollary 3, which we call the "milk and ice cream theorem," since it helps explain the paradoxical empirical fact that while consumers are more likely to buy milk, grocery stores devote much more aisle space to ice cream.

When the rate of decay is "severe," Theorems 2a and 2b collectively show that accuracy- and utility-maximizing recommendations coincide, and are diverse.

Implications. Our results lay out the specific ways in which diversity supports user utility, and thus inform how diversity should be incorporated into recommender systems. In particular:

- Maximizing user utility—properly conceived as incorporating consumption constraints—is often aligned with showing users a diverse set of items. Thus, to the extent that real-world systems do not show diverse recommendation sets, our results suggest that they are also failing to optimize user utility. Notably, this is true even before considering other ways in which diversity factors into utility (e.g., an intrinsic preference for diversity).

- Our results suggest principled approaches to diversify recommendations in a way that also optimizes utility. In particular, our results show that when users have consumption constraints, the relative likelihood a user prefers a specific type of item does not asymptotically affect the optimal representation of that item. Therefore, systems should recommend items relatively equally from a user's possible set of interests—even the niche interests.
- When the platform can estimate quality within a type (how often consumers like a specific ice cream flavor, conditional on wanting ice cream), the platform should recommend *more* items from types in which individual items are *less likely* to be satisfactory.

Related Work. A significant line of work has developed methods to increase diversity in recommendations (e.g., [2, 30]). Here, we provide an overview of two types of approaches and discuss how our work creates a synthesis that goes beyond what either of these two lines of work seeks to achieve. A first approach directly incorporates diversity into recommendations, for example by maximizing a weighted combination of individual item quality (e.g., accuracy) and item diversity or dissimilarity (e.g., [11, 54]). This approach explicitly navigates the accuracy-diversity trade-off by choosing how much to weigh diversity, at the cost of accuracy. A second approach optimizes objectives that are more implicitly connected to diversity. In other words, diversity is not guaranteed a priori in this approach. In search and information retrieval, maximizing the likelihood of a relevant result has been associated with diversity [4, 46], since effective search results must account for different intents of queries ("pandas" can refer to an animal or a Python package). Other objectives, such as alpha-nDCG [13], penalize redundancy in items shown. As in our model of user utility, these objectives result in diminishing returns when showing similar items.

Our work departs from past work by analyzing the amount of diversity induced by an objective. We focus on a simple measure of user utility that accounts for a user's consumption constraints. Our analysis requires instantiating a model of user preferences and item qualities, and estimating the optimal amount of diversity in specific settings. By doing this, we explicitly show that incorporating diversity into recommendations optimizes for user utility objectives (and vice versa), connecting the two approaches described above.

Paper Outline. In Section 2, we introduce our model. In Section 3, we introduce our theoretical results, first in a basic setting (Section 3.1) and then in a general setting (Section 3.2). In Section 4, we give an overview of our proof technique, sketching how we are able to derive our asymptotic results. In Section 5, we test our theoretical predictions in a range of computational experiments. In particular, we conduct a semi-synthetic experiment in which items and user preferences lie in a continuous space as estimated via matrix factorization, relaxing the assumption that there are a finite number of item and preference types. In Section 6, we conclude. Extended related work and additional details for computational experiments are included in the appendix.

2 MODEL

2.1 Specifying a recommendation setting

A recommender is tasked with recommending a fixed number of items to a user. There are m types of items, and each item belongs to exactly one type. At recommendation time, a user prefers exactly one of these m types of items. In our exposition, we will treat m as fixed and omit notation that depends on m.

Types are indexed by $[m] = \{1, 2, \cdots, m\}$ and we let a user's type preference be given by a random variable T, such that $\Pr[T = t] = p_t$ (so that $\sum_{t=1}^{m} p_t = 1$). When a user prefers type t (i.e., $T = t$), they only like items of type t. We assume that there are an arbitrarily large number of items of each type, and that conditional on $T = t$, a user likes the i-th item of type t independently with probability $q_{t,i}$. Without loss of generality, we let $q_{t,1} \geq q_{t,2} \geq \cdots$.

Formally, we let the random variable $V_{t,i}$ indicate if the user likes the i-th item of type t, so that

$$\Pr[V_{t,i} = 1] = \Pr[T = t] \Pr[V_{t,i} \mid T = t] = p_t q_{t,i}. \tag{1}$$

Note that the random variables $V_{t,i}$ are independent conditional on T. A **recommendation setting** is thus characterized by:

(1) **type probabilities** $p_1, p_2, \cdots, p_t$;
(2) **conditional item probabilities** $q_{t,1}, q_{t,2}, \cdots$ for $t \in [m]$.

In what follows, we assume that a recommendation setting is specified, and will omit dependencies of certain quantities on the recommendation setting.

(Remark: Under standard measures of accuracy, as we consider here, items have binary values. However, it is possible to consider a setup in which values can be distributed according to arbitrary distributions.)

2.2 Choosing a set of recommendations

We now focus on the task of selecting n items to recommend. In this case, a set of recommendations can be identified by an ordered tuple $S = (n_1, n_2, \cdots, n_m)$, where the recommender recommends the i-th item of type t for all $t \in [m]$ and $i \in [n_t]$. In other words, S represents the set of recommendations with the top n_t items from each type.[1] We will let $\mathcal{S}_n := \left\{(n_1, n_2, \cdots, n_m) \in \mathbb{Z}^m_{\geq 0} : \sum_{t=1}^{m} n_t = n\right\}$ denote the set of recommendation sets of size n.

We consider two objectives by which to optimize a set of recommendations $S = (n_1, n_2, \cdots, n_m) \in \mathcal{S}_n$:

- **Accuracy**: The expected proportion of items in S that the user likes, given by

$$\mathsf{acc}(S) := \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{m}\sum_{i=1}^{n_t} V_{t,i}\right] \tag{2}$$

$$= \frac{1}{n}\sum_{t=1}^{m} p_t \sum_{i=1}^{n_t} q_{t,i}. \tag{3}$$

  acc is the standard notion of accuracy commonly used to evaluate machine learning models. By the linearity of expectation, it is maximized by selecting the items with the highest $\mathbb{E}[V_{t,i}] = p_t q_{t,i}$—i.e., the individual items the user is most likely to like.

[1] In what follows, it will be clear that the recommender should only recommend the top items from each type. For example, the recommender would never recommend the first, second, and fourth items of a type, but not the third.

Table 1: Key notation in our model

p_t	a *type probability*; a user prefers type t with probability p_t
$q_{t,i}$	a *conditional item probability*; conditional on preferring type t, a user likes the i-th item of type t with probability $q_{t,i}$; in our general setting, we parameterize $q_{t,i}$ as $q_{t,i} = q_t(i+\beta)^{-\alpha}$
$V_{t,i}$	indicator random variable for if user likes i-th item of type t
α	the *rate of decay* of item quality within a type
q_t	a *relative type quality*; q_t determines the relative quality of items in t compared to other types
S_n	the set of n recommendations that maximizes $\texttt{acc}$, the expected proportion of items a user likes
$S_{n,1}$	the set of n recommendations that maximizes $\texttt{util}_1$, the probability a user likes at least one item
$r_t(S)$	the proportion of items in S of type t

- **Utility** (w/ unit consumption constraint): The probability that a user likes at least one item in S, given by

$$\texttt{util}_1(S) := \Pr\left[V_{t,i} = 1 \text{ for some } t \in [m], i \in [n_t]\right] \tag{4}$$

$$= 1 - \sum_{t=1}^{m} p_t \prod_{i=1}^{n_t} (1 - q_{t,i}). \tag{5}$$

  $\texttt{util}_1$ aligns with a user's satisfaction when they only intend to use one of the recommended items, as is common. In this case—e.g., when the goal is to find one restaurant to dine at, one movie to watch, or one website to visit—what matters is if the user likes at least one recommended item. (In some cases, users may use multiple items; we experimentally consider a more general form of utility in Section 5.1.)

In our analysis, we characterize the accuracy- and utility-maximizing recommendation sets, given by the following notation.

Definition 1 (S_n and $S_{n,1}$)**.** Given a specified recommendation setting, we let S_n and $S_{n,1}$ denote the recommendation sets of size n that maximize $\texttt{acc}$ and $\texttt{util}_1$, respectively:[2]

$$S_n := \arg\max_{S \in \mathcal{S}_n} \texttt{acc}(S) \tag{6}$$

$$S_{n,1} := \arg\max_{S \in \mathcal{S}_n} \texttt{util}_1(S). \tag{7}$$

To understand the diversity of S_n and $S_{n,1}$, we consider how well-represented items of each type are.

Definition 2 (Representation)**.** For $S = (n_1, n_2, \cdots, n_m)$, define

$$r_t(S) = \frac{n_t}{\sum_{u=1}^{m} n_u}, \tag{8}$$

the *representation* of type t in S.

Intuitively, sets with relatively equal representation across types are more diverse. In the following section, we will characterize—in terms of the type probabilities and conditional item probabilities—$r_t(S_n)$ and $r_t(S_{n,1})$ across several regimes.

[2]It is sometimes possible for multiple sets of recommendations to maximize these objectives. In this case, our results hold when selecting any of these sets.

3 RESULTS

We now introduce our theoretical results, which characterize the composition of the accuracy- and utility-maximizing sets S_n and $S_{n,1}$. We begin in Section 3.1 by considering a basic setting that starkly contrasts the objectives $\texttt{acc}$ and $\texttt{util}_1$; the first has a strong trade-off with diversity, while the second is entirely aligned with diversity. In Section 3.2, we characterize the representation of item types, $r_t(S_n)$ and $r_t(S_{n,1})$, in a significantly more general setting, where we focus on the effects of different properties of the recommendation setting (i.e., providing comparative statics).

3.1 A Basic Setting

We start with a simple case of our model where we let the type probabilities $p_1, p_2, \cdots, p_m$ vary but hold the conditional item probabilities $q_{t,i} = q$ fixed for some $q \in (0, 1)$. This setting provides a clear illustration of the drastic effect incorporating a consumption constraint can have.

To provide a concrete backdrop, suppose that there are m types of movie genres. Then the probability a user is in the mood for genre t is p_t. These type probabilities p_t can vary, so that a user is more likely to be in the mood for some genres than others. Conditional on a user being in the mood for any genre t, any movie in that genre is liked by the user independently with probability q.

Theorem 1. *Given type probabilities $p_1, p_2, \cdots, p_m$ and conditional item probabilities $q_{t,i} = q$,*

$$r_t(S_n) = \mathbb{1}_{\{t = \arg\max_t p_t\}} \tag{9}$$

$$\lim_{n \to \infty} r_t(S_{n,1}) = \frac{1}{m}. \tag{10}$$

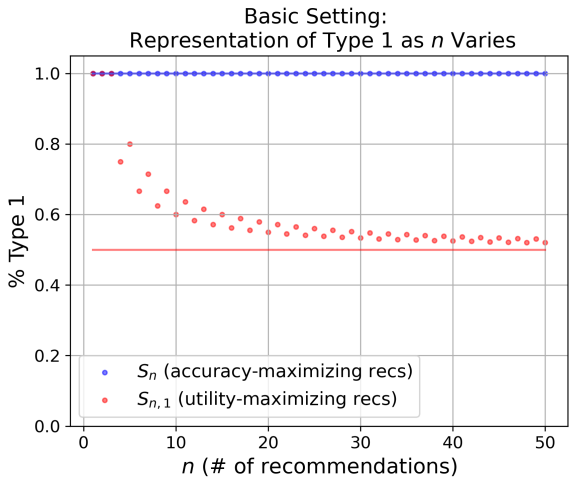
This result conveys a strong dichotomy. (9) says that recommendations maximizing $\texttt{acc}$ contain exclusively items from the genre the user is most likely to prefer, $\arg\max_t p_t$. This reflects the empirical existence of an accuracy-diversity trade-off: the accuracy-maximizing set of recommendations is fully homogeneous.

Meanwhile, (10) says that the set of recommendations that maximizes utility with a unit consumption constraint is fully diverse. Specifically, as the number of recommended items n grows large, the recommender should recommend an equal proportion of items from each genre.[3] In this way, accounting for a user's consumption constraint dissolves the apparent accuracy-diversity trade-off; maximizing the probability a user likes *at least one* recommended movie is fully aligned with recommending a diverse set of movies.

We now take a moment to convey the intuition behind Theorem 1. S_n maximizes $\texttt{acc}$, which is equivalent to maximizing the expected number of recommended items the user likes. By linearity of expectation, this is achieved when recommending the individual items the user likes with the highest probability. The probability a user likes any item in genre t is equal to $p_t q$. Therefore, the recommender should only recommend items from genre $\arg\max_t p_t$.

When the objective is to instead maximize $\texttt{util}_1$, the probability a user likes at least one item, recommending items from only one type is suboptimal. After recommending, say, many action movies, the probability the user is in the mood for action but does not like

[3]In fact, one can show that $r_t(S_{n,1}) = \frac{1}{m} + O\left(\frac{1}{n}\right)$, giving a relatively fast rate of convergence.



Details: A recommendation setting with $(p_1, p_2) = (0.8, 0.2)$, $q_{t,i} = 0.5$ for all t, i. We plot empirical results (dots) and asymptotic theoretical results (solid lines) from Theorem 1.

Figure 1: An illustration of Theorem 1. Even while a user prefers type 1 much more often than type 2, as the number of recommendations increases, utility-maximizing recommendations (red) represent both types roughly equally. Meanwhile, accuracy-maximizing recommendations (blue) remain fully homogeneous throughout.

any of the recommended action movies is small ($p_t(1-q)^{n_t}$), where n_t is the number of recommended action movies). At this point, recommending more action movies has diminishing returns, and one should hedge for the possibility that the user is in the mood for a different genre.

A surprising insight of Theorem 1 is that the type probabilities p_t do not play any role asymptotically for $S_{n,1}$; even when a user watches more action than romance, the optimal set of recommendations represents the genres equally. To give some intuition, let X be the event that a user does not like *any* recommended item. For an optimal set of recommendations S, $\Pr[X \mid T = t] = p_t q^{n_t}$ should be equalized across t; otherwise, there would be an incentive to recommend more items from a type where this probability is higher. If $p_1 > p_2$, $\Pr[X \mid T = 1] = \Pr[X \mid T = 2]$ when recommending only a constant number $\log_q \frac{p_1}{p_2}$ more items from type 1 than type 2. So asymptotically, the proportion of items recommended from each type is equal. Representation thus quickly converges to this asymptotic value as n increases; this is illustrated in Figure 1.

3.2 A General Setting

We turn to a more general case where we consider heterogeneous conditional item probabilities and analyze comparative statics. Again, we consider arbitrary type probabilities $p_1, p_2, \cdots, p_m$. Now, we parameterize conditional item probabilities in the following way:

$$q_{t,i} := q_t (i + \beta)^{-\alpha}, \tag{11}$$

for $\alpha, \beta \geq 0$. This parameterization models heterogeneity of conditional item probabilities both within and across types. Some comments on the parameters α, β, and $(q_1, q_2, \cdots, q_t)$:

- α is the *rate of decay* of item quality within a type. When $\alpha > 1$, this rate is extreme in the following sense: even if a user were recommended an infinite number of items in their preferred type, they (1) would only like a constant number of the items in expectation, and (2) with positive probability, would not like *any* of the recommended items.[4] Therefore, when the recommender has a reasonable "supply" of items, $\alpha \leq 1$ is realistic.
- β parameterizes the initial steepness, with higher β corresponding to lower initial steepness. β does not end up appearing in our estimates.
- $q_1, q_2, \cdots, q_m$ are the *relative type qualities*. If $q_t > q_{t'}$, then $q_{t,i} > q_{t',i}$ for all i. Users can be less likely to like items of a certain type, even conditioned on preferring that type. This has two equivalent interpretations: the user is more picky when they prefer type t', or the recommender has lower quality or more niche items in type t'.

We now give two main results, Theorem 2.A and Theorem 2.B which characterize $r_t(S_n)$ and $r_t(S_{n,1})$ in terms of these parameters.

Theorem 2.A (Accuracy-maximizing recommendations). *Given type probabilities $p_1, p_2, \cdots, p_m$ and conditional item probabilities $q_{t,i} := q_t(i + \beta)^{-\alpha}$,*

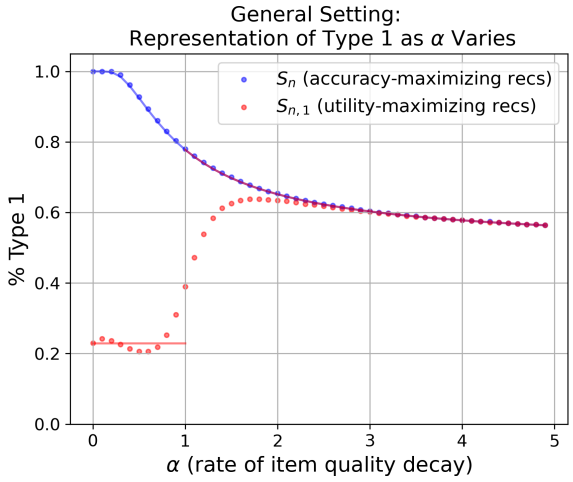
$$\lim_{n \to \infty} r_t(S_n) = \frac{(p_t q_t)^{1/\alpha}}{\sum_{u=1}^{m} (p_u q_u)^{1/\alpha}} \tag{12}$$

The key takeaway from Theorem 2.A is that accuracy-maximizing recommendations are more diverse when α is larger, i.e., when the quality of items in a type decays faster. Intuitively, this means that a recommender quickly runs out of high-quality items in a type, and thus benefits more from recommending items from other types. On the other hand, with small α, the recommender has access to many high-quality items within each type.

Let us consider three cases of Theorem 1 to understand the functional form in (12):

- As $\alpha \to 0$, and if $\arg\max_t p_t q_t$ is unique,[5]
$$\lim_{n \to \infty} r_t(S_n) \to \mathbb{1}_{\{t = \arg\max_t p_t q_t\}}, \tag{13}$$
meaning that only the type with the highest type probability is recommended.
- For $\alpha = 1$,
$$\lim_{n \to \infty} r_t(S_n) \propto p_t q_t, \tag{14}$$
meaning that a type is recommended in proportion to the probability a user likes items in that type.
- For $\alpha \to \infty$,
$$\lim_{n \to \infty} r_t(S_n) \to \frac{1}{m}, \tag{15}$$
meaning that an equal proportion of items from each type are recommended.

[4] Both facts boil down to the convergence of $\sum_{i=1}^{\infty} i^{-\alpha}$ when $\alpha > 1$.
[5] If there are multiple types with the maximum type probability, one may check that in the limit, an equal proportion of items are recommended from these types, and a zero proportion from other types.



Details: A setting with $(p_1, p_2) = (0.6, 0.4)$, $(q_1, q_2) = (0.7, 0.3)$, $\beta = 2$, and α varying. Empirical results (dots) are for $n = 500$ and theoretical estimates (solid lines) are from Theorem 2.A and 2.B. For $\alpha \in [0, 1]$, the plotted theoretical estimate for $S_{n,1}$ are based on the theoretical result for $\alpha = 0$.

Figure 2: An illustration of Theorem 2.A and 2.B in a setting with two item types. When $\alpha < 1$, utility-maximizing recommendations (red) represent type 1 less than type 2 because 2 has lower relative conditional item probability, even though a user prefers 1 more often than 2. Behavior changes at $\alpha = 1$, after which accuracy- and utility-maximizing recommendations coincide.

As α ranges from 0 to ∞, the amount of diversity in S_n smoothly interpolates from maximal homogeneity to proportional representation to maximal diversity. Notably, when $\alpha \leq 1$, diversity is in the range between homogeneity and proportional representation. This suggests that in practice, the accuracy-diversity trade-off is particularly severe when the recommender has access to many high-quality items of a type.

We next turn to utility-maximizing recommendations $S_{n,1}$, which account for a unit consumption constraint.

Theorem 2.B (Utility-maximizing recommendations). *Given type probabilities $p_1, p_2, \cdots, p_m$ and conditional item probabilities $q_{t,i} := q_t(i+\beta)^{-\alpha}$,*

$$\lim_{n\to\infty} r_t(S_{n,1}) = \begin{cases} \frac{\left(\log \frac{1}{1-q_t}\right)^{-1}}{\sum_{u=1}^{m}\left(\log \frac{1}{1-q_u}\right)^{-1}} & \alpha = 0 \\ \lim_{n\to\infty} r_t(S_{n,1}) = \frac{(p_t q_t)^{1/\alpha}}{\sum_{u=1}^{m}(p_u q_u)^{1/\alpha}} & \alpha > 1 \end{cases}. \tag{16}$$

The representation exhibits phase change at $\alpha = 1$. As mentioned in our discussion of the parameters, we expect $\alpha > 1$ to represent an "extreme setting." We thus focus on the case $\alpha = 0$, which we pull out as its own result. (Note also that Theorem 1 is obtained by taking $q_1 = q_2 = \cdots = q_m$ and $\alpha = 0$.)

Corollary 3 (The "milk and ice cream theorem"). *Given type probabilities $p_1, p_2, \cdots, p_m$ and conditional item probabilities $q_{t,i} = q_t(i+\beta)^{-\alpha}$, when $\alpha = 1$,*

$$\lim_{n\to\infty} r_t(S_{n,1}) = \frac{\left(\log \frac{1}{1-q_t}\right)^{-1}}{\sum_{u=1}^{m}\left(\log \frac{1}{1-q_u}\right)^{-1}}, \tag{17}$$

meaning that $r_t(S_{n,1})$ is larger for types t with lower q_t.

The corollary's name references a paradoxical fact of grocery stores: that even while a customer is much more likely to buy milk than ice cream, significantly more aisle space is devoted to ice cream. The paradox can be resolved by the corollary in the following way. Let milk be type 1 and ice cream be type 2. A customer is more likely to purchase milk than ice cream, so $p_1 > p_2$. However, the probability a given carton of ice cream will satisfy a customer trying to purchase milk is lower than the probability that a given bottle of milk will satisfy a customer trying to purchase milk.[6] This means that $q_1 > q_2$. Corollary 3 reveals that more items should be recommended from the type with lower q_t—and, in fact, that p_t is asymptotically irrelevant.

Corollary 3 demonstrates a broader insight into recommendations when the user has consumption constraints. Rather than focusing on type probabilities, it is more important to consider the conditional item values *within a type*—in particular, to recommend more items from types with low conditional item values. A good set of recommendations covers its bases across all possible preferences of the user, and "covering" a type requires more items when items in that type have low conditional item probabilities.

Computational experiments suggest that behavior remains similar when α is small but larger than 0 (see Figure 2). However, Theorem 2.B shows that the behavior of $S_{n,1}$ changes when $\alpha > 1$. In fact, referring back to Theorem 2.A, we have that in this regime,

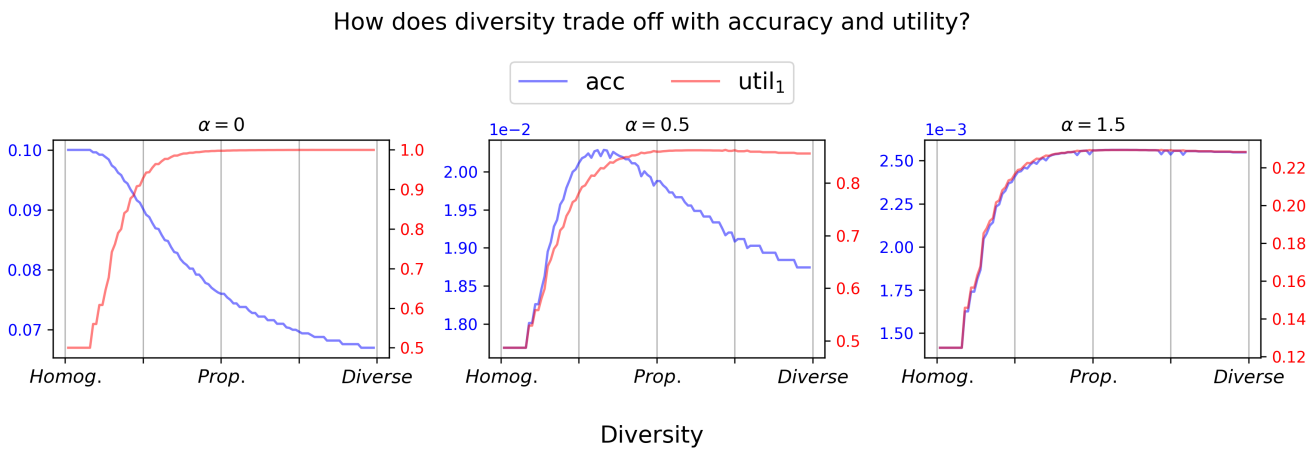
$$\lim_{n\to\infty} r_t(S_n) = \lim_{n\to\infty} r_t(S_{n,1}) = \frac{(p_t q_t)^{1/\alpha}}{\sum_{u=1}^{m}(p_u q_u)^{1/\alpha}}. \tag{18}$$

So when $\alpha > 1$, both S_n and $S_{n,1}$ become more diverse as α increases. To provide rough intuition for this equality, note that for large α, it is unlikely that a user will like more than one item in each type. Therefore, maximizing the likelihood a user likes at least one item ($\texttt{util}_1$) is roughly equivalent to maximizing the total number of recommended items a user likes (acc).

3.3 Summary of results

To summarize our theoretical results, while accuracy can exhibit a strong trade-off with diversity (especially when the rate of decay α is relatively small), our measure of utility that accounts for the capacity constraints of users does not exhibit a trade-off with diversity in the settings we study. This is exhibited in computational results shown in Figure 3, which shows how accuracy and utility vary as the level of diversity increases.

[6] We note that this fact is complicated by customers' increasingly diversified tastes for—and the increased availability of—different types of plant-based milks [45].



Details: Recommendation settings in which $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$ and $q_{t,i} = 0.2(i+1)^{-\alpha}$ for $\alpha \in \{0, 0.5, 1.5\}$.

Figure 3: In three settings with varying α, we plot how accuracy and utility trade off with diversity. We consider sets S with $n = 100$ items ranging from complete homogeneity (only type 1 recommended), to prop. representation ($r_t(S) = p_t$) to complete diversity ($r_t(S) = \frac{1}{m}$), and plot $\texttt{acc}(S)$ and $\texttt{util}_1(S)$. Throughout, $\texttt{util}_1$ is aligned with diversity, while $\texttt{acc}$ exhibits a strong trade-off with diversity when $\alpha = 0$, which becomes less severe for larger α. These results agree with Theorem 2.A and Theorem 2.B.

4 PROOF TECHNIQUE

We overview our proof technique, which enables us to precisely characterize the asymptotic behavior of recommendation sets, allowing for comparative statics. We begin with some high-level intuition. The basic idea is that maximizing functions of the form

$$\sum_{t=1}^{m} \lambda_t h(z_t), \tag{19}$$

subject to the constraint $(z_1, \cdots, z_m) \in \mathcal{S}^n$ is tractable when f is simple (a monomial, e.g.). Then, roughly speaking, one can solve $\lambda_1 h'(x_1) = \lambda_2 h'(x_2) = \cdots = \lambda_m h'(x_m)$, and show that the integer-valued optimum must be near the real-valued optimum. While the objectives $\texttt{acc}$ and $\texttt{util}_1$ do not take the exact form as above, we show that there are choices of λ_t where the objectives evaluate to

$$\sum_{t=1}^{m} \lambda_t h_t(z_t), \tag{20}$$

where even when $h_t(z_t)$ is complicated, $\lim_{z\to\infty} \frac{h_t(z)}{h(z)} = 1$ for a simple function h. We then show that under reasonable assumptions on h, the solution to (20) is approximated by that of (19) as $n \to \infty$.

5 COMPUTATIONAL EXPERIMENTS

We present results from a range of computational experiments, showing that our theoretical results generalize to practical settings.

5.1 Finite number of recommendations and beyond unit consumption constraints

We first focus on experiments in which n is finite, ranging from small to moderate. We also relax the assumption that users have *unit* consumption constraints and consider varying rates of decay $\alpha < 1$. Consider the following more general version of utility corresponding to a consumption constraint of k:

$$\texttt{util}_k(S) := \mathbb{E}\left[\max\left\{\sum_{t\in[m], i\in[n_t]} V_{t,i}, \quad k\right\}\right], \tag{21}$$

the number of items in S the user likes, capped at k. Recall that $V_{t,i}$ indicates if a user likes the i-th item of type t. Accordingly, let

$$S_{n,k} := \arg\max_{S\in\mathcal{S}_n} \texttt{util}_k(S), \tag{22}$$

where we recall that $\mathcal{S}_n$ is the set of all possible recommendation sets with n items. Our previous definitions of $\texttt{util}_1$ and $S_{n,1}$ agree with this more general definition. (When $k = 1$, (21) reduces to the probability $V_{t,i} = 1$ for at least one item.) We evaluate the diversity of $S_{n,k}$ for $k > 1$ in different settings, testing the robustness of our results to larger consumption constraints. (Obtaining theoretical results in this regime is an interesting open question.)

We focus on a recommendation setting where there are two item types with $p_1 = 0.7$ and $p_2 = 0.3$. We let $q_{t,i} = q_t(i+\beta)^{-\alpha}$ where we fix $q_t = 0.5$ and $\beta = 1$, and only consider $\alpha < 1$ (reasonable rates of decay). We compare our empirical results to the estimate given by Theorem 2.B, which tells us that when $\alpha = 0$, $\lim_{n\to\infty} r_1(S_{n,1}) = \frac{1}{2}$, meaning that both types are equally represented. Results are shown in Table 2. We observe that $r_1(S_{n,k})$ is near 0.5 whenever $\frac{k}{n}$ and α are small, but closer to 1 when k is near n. This suggests that our broad theory—that utility-maximizing recommendations tend to be diverse—holds even when the consumption constraint is larger than 1, as long as it is not too close to n. When users use almost all recommended items, however, utility-maximizing recommendations may not be diverse. We provide additional computational experiments in different settings in the appendix, as well as details on how we determine the empirically objective-maximizing recommendations.

$r_1(S_{n,k})$ **when** $(p_1, p_2) = (0.7, 0.3)$ **and** $q_{t,i} = 0.5(i+1)^{-\alpha}$.

	$\alpha = 0$				$\alpha = 0.2$				$\alpha = 0.5$				$\alpha = 0.9$			
n	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$
10	0.6	0.6	1.0	1.0	0.6	0.6	1.0	1.0	0.6	0.7	0.9	0.9	0.7	0.7	0.8	0.8
20	0.55	0.55	0.6	1.0	0.55	0.6	0.7	1.0	0.6	0.65	0.8	0.95	0.65	0.7	0.75	0.8
50	0.52	0.48	0.52	0.54	0.52	0.56	0.54	0.66	0.56	0.58	0.64	0.86	0.64	0.62	0.78	0.76
100	0.5	0.49	0.49	0.49	0.52	0.49	0.49	0.53	0.55	0.56	0.57	0.76	0.63	0.62	0.7	0.78

Table 2: Empirical estimates of $r_1(S_{n,k})$ based on 5000 iterations for each setting for each possible set of recommendations. Blue indicates close to equal representation (the theoretical result for $r_1(S_{n,k})$ for $\alpha = 0$), and red indicates less diversity.

5.2 Continuous Items and User Preferences

We now depart from our assumption that user preferences and items fall into discrete types, and instead represent both by embeddings on the unit d-dimensional sphere S^d. Given a user preference $t \in S^d$ and an item $v \in S^d$, we let the value of item v be equal to the dot product $\max(t \cdot v, 0)$. Thus, items that are closer to a user's true preference have higher value and items cannot have negative value.

For a set of n recommendations, we again let accuracy measure the average value of recommended items and utility measure the value of the best-recommended item (that of the highest value). Here, utility again models a user who chooses one item to use from a set of recommendations. We compare the performance of an accuracy-maximizing set of recommendations with a heuristically constructed set of diverse recommendations, plotting results in Figure 4. In alignment with our general findings, the accuracy-maximizing set of recommendations is homogeneous, while a diverse set of recommendations improves user utility. This suggests that our theoretical findings extend beyond our exact model.

Specifically, we use 10-dimensional embeddings trained using interaction data from GoodReads between 1000 users and 200 books. Embeddings are normalized to lie on S^{10}. We assume user preferences are drawn uniformly from the set of books they have interacted with (since these represent the range of interests the user has). We limit our experiments to the 206 users with at least 20 total interactions. For each user, we consider a "train set" of 10 of the user's past interactions. We then select n of the 190 remaining books to recommend. The accuracy-maximizing set is chosen to maximize accuracy when user preferences are assumed to be drawn from the train set. The diverse set is chosen by choosing the closest items to each of the books in the train set (thus, covering the full range of user interests). We evaluate recommendations by randomly drawing a user preference from the books they have interacted with that were not already included in the train set. Additional details for our experiment are given in the appendix.

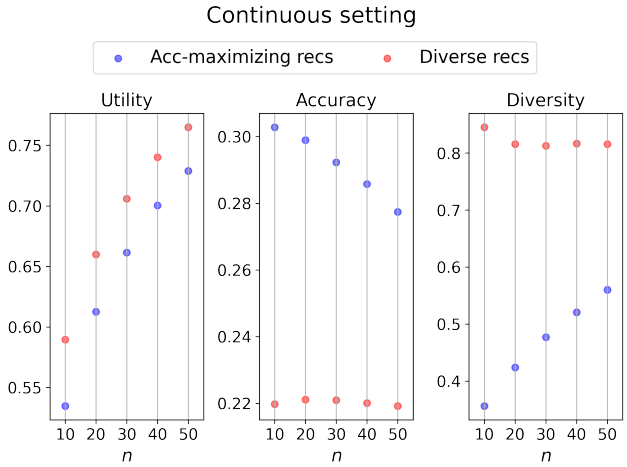


Figure 4: Book recommendations that maximize accuracy on a train set are more accurate in evaluation, but also achieve less utility and diversity as compared to a heuristically-chosen diverse set. Here, diversity is the average cosine distance between recommendations. The values plotted are averages over 100 trials for each of the 206 users we evaluated.

6 CONCLUSION

We introduced and analyzed a model of recommendations that reconciles the apparent accuracy-diversity trade-off. In particular, we showed that accuracy is misaligned with user utility, because it does not consider the consumption constraints of users. By accounting for these consumption constraints, we found that user utility is in fact aligned with and supported by diversity. As a consequence, navigating the accuracy-diversity trade-off can be viewed as a way of incorporating diversity to help align accuracy with the more fundamental goal of user utility. Our results provide insight into how diversity can be incorporated in this manner.

ACKNOWLEDGMENTS

This work was supported in part by a Digital Life Initiative doctoral fellowship, a Google Research Scholar award, NSF CAREER 2142419, a CIFAR Azrieli Global scholarship, a LinkedIn Research Award, the Abby Joseph Cohen Faculty Fund, a Vannevar Bush Faculty Fellowship, AFOSR grant FA9550-19-1-0183, a Simons Collaboration grant, a grant from the MacArthur Foundation, and a Meta research award. We thank members of the AI, Policy, and Practice (AIPP) group at Cornell and participants of the Marketplace Innovation Workshop for helpful feedback and discussion.

REFERENCES

[1] Himan Abdollahpouri, Zahra Nazari, Alex Gain, Clay Gibson, Maria Dimakopoulou, J. Anderton, Benjamin Carterette, Mounia Lalmas, and Tony Jebara. 2023. Calibrated Recommendations as a Minimum-Cost Flow Problem. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (2023).

[2] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911.

[3] Gediminas Adomavicius and YoungOk Kwon. 2012. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering* 24 (2012), 896–911.

[4] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/1498759.1498766

[5] Georgios Alexandridis, Georgios Siolas, and Andreas Stafylopatis. 2015. Accuracy Versus Novelty and Diversity in Recommender Systems: A Nonuniform Random Walk Approach. In *Recommendation and Search in Social Networks*.

[6] Bushra Alhijawi, Salam Fraihat, and Arafat A. Awajan. 2023. Multi-factor ranking method for trading-off accuracy, diversity, novelty, and coverage of recommender systems. *International Journal of Information Technology* 15 (2023), 1427 – 1433.

[7] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic Effects on the Diversity of Consumption on Spotify. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2155–2165. https://doi.org/10.1145/3366423.3380281

[8] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. 2015. Optimal Greedy Diversity for Recommendation. In *International Joint Conference on Artificial Intelligence*.

[9] Dimitris Bertsimas and Velibor V Mišic. 2015. Data-driven assortment optimization. *Management Science* 1 (2015), 1–35.

[10] William Brown and Arpit Agarwal. 2022. Diversified Recommendations for Agents with Adaptive Preferences. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=VVsNTPK1FBp

[11] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[12] Qinyi Chen, Negin Golrezaei, Fransisca Susan, and Edy Baskoro. 2022. Fair assortment planning. *arXiv preprint arXiv:2208.07341* (2022).

[13] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 659–666.

[14] Maurizio Ferrari Dacrema. 2021. Demonstrating the Equivalence of List Based and Aggregate Metrics to Measure the Diversity of Recommendations (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* (2021).

[15] James Mario Davis, Guillermo Gallego, and Huseyin Topaloglu. 2014. Assortment Optimization Under Variants of the Nested Logit Model. *Oper. Res.* 62 (2014), 250–273.

[16] Omar El Housni, Omar Mouchtaki, Guillermo Gallego, Vineet Goyal, Salal Humair, Sangjo Kim, Ali Sadighian, and Jingchen Wu. 2021. Joint assortment and inventory planning for heavy tailed demand. *Columbia Business School Research Paper Forthcoming* (2021).

[17] Omar El Housni and Huseyin Topaloglu. 2022. Joint assortment optimization and customization under a mixture of multinomial logit models: On the value of personalized assortments. *Operations Research* (2022).

[18] Farzad Eskandanian and Bamshad Mobasher. 2020. Using Stable Matching to Optimize the Balance between Accuracy and Diversity in Recommendation. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (2020).

[19] Ignacio Fernández-Tobías, Paolo Tomeo, Iván Cantador, T. D. Noia, and Eugenio Di Sciascio. 2016. Accuracy and Diversity in Cross-domain Recommendations for Cold-start Users with Positive-only Feedback. *Proceedings of the 10th ACM Conference on Recommender Systems* (2016).

[20] Guillermo Gallego and Huseyin Topaloglu. 2014. Constrained assortment optimization for the nested logit model. *Management Science* 60, 10 (2014), 2583–2601.

[21] Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1100–1111.

[22] Anupriya Gogna and Angshul Majumdar. 2017. Balancing accuracy and diversity in recommendations using matrix completion framework. *Knowl. Based Syst.* 125 (2017), 83–95.

[23] David Goldberg. 2014. Diversity in Search. (2014).

[24] Wenshuo Guo, Karl Krauth, Michael Jordan, and Nikhil Garg. 2021. The stereotyping problem in collaboratively filtered recommender systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–10.

[25] Xiaoyun He. 2022. Does Utilizing Online Social Relations Improve the Diversity of Personalized Recommendations? *Int. J. Strateg. Decis. Sci.* 13 (2022), 1–15.

[26] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[27] Zhipeng Hou and Jing Liu. 2020. A Two-phase Evolutionary Algorithm for Solving the Accuracy-diversity Dilemma in Recommendation. *2020 IEEE Congress on Evolutionary Computation (CEC)* (2020), 1–8.

[28] Elvin Isufi, Matteo Pocchiari, and Alan Hanjalic. 2021. Accuracy-diversity trade-off in recommender systems via graph convolutions. *Inf. Process. Manag.* 58 (2021), 102459.

[29] Srikanth Jagabathula. 2014. Assortment optimization under general choice. *Available at SSRN 2512831* (2014).

[30] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 935–944.

[31] Amin Javari and Mahdi Jalili. 2015. A probabilistic model to resolve diversity–accuracy challenge of recommendation systems. *Knowledge and Information Systems* 44 (2015), 609–627.

[32] Jon Kleinberg and Manish Raghavan. 2018. Selection problems in the presence of implicit bias. *arXiv preprint arXiv:1801.03533* (2018).

[33] Jon Kleinberg and Maithra Raghu. 2018. Team performance with test scores. *ACM Transactions on Economics and Computation (TEAC)* 6, 3-4 (2018), 1–26.

[34] Jon M. Kleinberg, Emily Ryu, and Éva Tardos. 2023. Calibrated Recommendations for Users with Decaying Attention. *ArXiv* abs/2302.03239 (2023).

[35] A Gürhan Kök and Marshall L Fisher. 2007. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* 55, 6 (2007), 1001–1021.

[36] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems* 123 (2017), 154–162.

[37] Jae kyeong Kim, Il Young Choi, and Qinglong Li. 2021. Customer Satisfaction of Recommender System: Examining Accuracy and Diversity in Several Types of Recommendation Approaches. *Sustainability* 13 (2021), 6165.

[38] Emanuel Lacić, Dominik Kowald, Markus Reiter-Haas, Valentin Slawicek, and E. Lex. 2017. Beyond Accuracy Optimization: On the Value of Item Embeddings for Student Job Recommendations. *ArXiv* abs/1711.07762 (2017).

[39] Jianguo Liu, Kerui Shi, and Qiang Guo. 2012. Solving the accuracy-diversity dilemma via directed random walks. *Physical review. E, Statistical, nonlinear, and soft matter physics* 85 1 Pt 2 (2012), 016118.

[40] Ivan Medvedev, Taylor Gordong, and Haotian Wu. 2019. Powered by AI: Instagram's Explore recommender system. (2019).

[41] Scott Page. 2008. *The Difference.* Princeton University Press.

[42] Sung-Hyuk Park and Sang Pil Han. 2013. From Accuracy to Diversity in Product Recommendations: Relationship Between Diversity and Customer Retention. *International Journal of Electronic Commerce* 18 (2013), 51 – 72.

[43] Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1929–1942.

[44] Bibek Paudel, Thilo Haas, and Abraham Bernstein. 2017. Fewer Flops at the Top: Accuracy, Diversity, and Regularization in Two-Class Collaborative Filtering. *Proceedings of the Eleventh ACM Conference on Recommender Systems* (2017).

[45] Victoria Petersen. 2022. Have We Reached Peak Plant Milk? Not Even Close. *The New York Times* (2022). https://www.nytimes.com/2022/02/28/dining/plant-based-milk.html

[46] Filip Radlinski, Robert D. Kleinberg, and Thorsten Joachims. 2008. Learning diverse rankings with multi-armed bandits. In *International Conference on Machine Learning*.

[47] Shaina Raza and Chen Ding. 2021. Deep Neural Network to Tradeoff between Accuracy and Diversity in a News Recommender System. *2021 IEEE International Conference on Big Data (Big Data)* (2021), 5246–5256.

[48] Steffen Rendle. 2012. Factorization Machines with libFM. *ACM Trans. Intell. Syst. Technol.* 3, 3, Article 57 (May 2012), 22 pages.

[49] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[50] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* 58, 6 (2010), 1666–1680.

[51] Paat Rusmevichientong, David Shmoys, Chaoxu Tong, and Huseyin Topaloglu. 2014. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management* 23, 11 (2014), 2023–2039.

[52] João Sá, Vanessa Queiroz Marinho, Ana Rita Magalhães, Tiago Lacerda, and Diogo Gonçalves. 2022. Diversity Vs Relevance: A Practical Multi-objective Study in Luxury Fashion Recommendations. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).
[53] Sinan Seymen, Himan Abdollahpouri, and Edward C. Malthouse. 2021. A Constrained Optimization Approach for Calibrated Recommendations. *Proceedings of the 15th ACM Conference on Recommender Systems* (2021).
[54] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.
[55] Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 86–94. https://doi.org/10.1145/3240323.3240369
[56] Shengqi Wu, Huaizhen Kou, Chao Lv, Wanli Huang, Lianyong Qi, and Hongya Wang. 2020. Service Recommendation with High Accuracy and Diversity. *Wirel. Commun. Mob. Comput.* 2020 (2020), 8822992:1–8822992:10.
[57] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).
[58] Mi Zhang and Neil J. Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In *ACM Conference on Recommender Systems*.
[59] Tao Zhou, Zoltan Kuscsik, Jianguo Liu, Matus Medo, Joseph R. Wakeling, and Yi-Cheng Zhang. 2008. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107 (2008), 4511 – 4515.

In the appendices, we provide extended related work (Appendix A) and additional computational experiments and experimental details (Appendix B). A longer version of the paper, with complete proofs, can be found at https://arxiv.org/abs/2307.15142. Code to reproduce the experiments and figures is available at https://github.com/kennylpeng/diverse_recommendations.

A EXTENDED RELATED WORK

Our work sits at the intersection of two broad sets of work. On the one hand are arguments that diversity is key to achieving efficiency. On the other are those that cast diversity as in conflict with efficiency or accuracy, but perhaps that diversity should nevertheless be pursued as an axiomatic good.

Broadly, our work seeks to understand this tension by sharply characterizing the *amount* of diversity in efficient solutions, as a function of key setting characteristics: user utilities and consumption constraints, and uncertainty in the item quality distribution. In particular, our results characterize *in what settings* the intuition regarding diversity being efficient holds, and in what settings they may be in conflict.

The (efficiency) benefits of diversity. The importance of diversity for efficiency is an old idea present across many fields; Page [41] synthesizes the conceptual and empirical arguments in support of this principle. Hong and Page [26] develop a model in which a randomly selected team of problem solvers outperforms a team of the individually best-performing agents, due to diversity in problem solving perspective (Kleinberg and Raghu [33] show that, in some settings, there exist *tests* under which selecting the best-performing agents again becomes optimal). Kleinberg and Raghavan [32] show that constraints promoting diversity can improve efficiency when they work to counteract a decision-maker's biases. Agrawal et al. [4] develop an algorithm to diversify search results, to minimize the risk of user dissatisfaction. We are particularly influenced by the work of Steck [54], who presents the intuition that recommendations should be *calibrated*: "When a user has watched, say, 70 romance movies and 30 action movies, then it is reasonable to expect the personalized list of recommended movies to be comprised of about 70% romance and 30% action movies as well." Guo et al. [24] show that collaborative filtering-based recommendations may not be able to effectively show users such a diverse set of content, harming efficiency.

More broadly, researchers studying various combinatorial optimization problems may find it obvious that homogeneous solutions can be sub-optimal; indeed, in classical problems like *maximum coverage*, redundancy is undesirable.

Our work particularly is intimately connected to the large literature on assortment optimization [9, 12, 15–17, 20, 29, 35, 50, 51]. That literature also considers consumption-constrained consumer item selections based on an intermediary's recommendations (e.g., that customers picks one item according to a multinomial choice model). The literature primarily devises *approximation algorithms* to find the optimal recommendation ("assortment") as a function of the consumer's choice model, platform objective, and the item distribution. In other words, an implicit premise of this literature is that the naive approach of presenting the items with highest individual expected values is sub-optimal, i.e., that optimal assortments are not completely 'homogeneous.' On the other hand, optimal assortments are not necessarily diverse; roughly speaking, the results of El Housni and Topaloglu [17] imply that a standard assortment approach (Mixed MNL) might produce solutions that are not "diverse" enough to satisfy multiple customer types, and so there is benefit to personalize to each type.[7][8] Our work contributes to this literature by (a) examining the implicit premise that optimal assortments are not homogeneous (i.e., when is the naive[9] approach sufficient?); and (b) showing the characteristics under which optimal assortments are not diverse.

Diversity and fairness as a contrast to efficiency and accuracy. On the other hand, many works start with the premise that—although diversity may conflict with efficiency or accuracy—it is an axiomatic good that should be pursued. For example, diversity is often considered to be inherently desirable from a fairness perspective and user satisfaction perspective. As a result, there is a wide body of work devoted to optimizing for various metrics of diversity. A common approach (taken, for example, in Carbonell and Goldstein [11] and Gimpel et al. [21]) is to consider an objective function that balances a weighted measure of "accuracy" or "relevance" with a measure of diversity. More recently, Brown and Agarwal [10] consider set recommendation for an agent with adaptive preferences, to ensure that consumption over time is diverse. Numerous metrics for diversity have been proposed—we refer the reader to Kunaver and Požrl [36] for a survey. Similarly, the fair ranking and recommendation literature (see Patro et al. [43] and Zehlike et al. [57] for recent surveys) considers metrics and methods for fairness in such problems. On the other hand, empirical work has demonstrated that such tradeoffs may be small in practice [49]. Such formulations imply that there is a tension between diversity and measures of accuracy.

[7] We thank the authors for highlighting this connection to us.
[8] Furthermore, as Chen et al. [12] recently characterize, standard assortment optimization approaches may be "unfair" to items in other ways.
[9] Note that *naive* is much simpler than the *greedy* approach studied in the literature, which picks items iteratively potentially as a function of previous items picked.

$r_1(S_{n,k})$ **when** $(p_1, p_2) = (0.6, 0.4)$, $(q_1, q_2) = (0.7, 0.3)$, **and** $q_{t,i} = q_t(i+1)^{-\alpha}$.

	$\alpha = 0$				$\alpha = 0.2$				$\alpha = 0.5$				$\alpha = 0.9$			
n	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$
10	0.3	0.4	0.8	1.0	0.4	0.5	1.0	1.0	0.5	0.6	1.0	1.0	0.6	0.8	0.9	0.9
20	0.3	0.3	0.4	0.8	0.3	0.4	0.55	1.0	0.4	0.5	0.8	1.0	0.55	0.7	0.8	0.8
50	0.26	0.18	0.28	0.34	0.26	0.3	0.32	0.5	0.3	0.38	0.56	0.88	0.44	0.58	0.72	0.8
100	0.24	0.11	0.16	0.26	0.25	0.19	0.25	0.34	0.26	0.29	0.39	0.68	0.39	0.43	0.73	0.85

Table 3: Empirical estimates of $r_1(S_{n,k})$ based on 5000 iterations for each setting (i.e., entry in table) for each possible set of recommendations. The table illustrates the effect of the varying relative type qualities q_1 and q_2. Because q_2 is lower than q_1, we would expect type 1 to be represented less than type 2 according to our theoretical prediction from Corollary 3, which applies directly to the case $\alpha = 0$, $k = 1$, and $n \to \infty$. We highlight when this is the case in blue.

$r_1(S_{n,k})$ **when** $(p_1, p_2) = (0.7, 0.3)$, $q_{t,i} = 0.5(i+1)^{-\alpha}$.

	$\alpha = 1$				$\alpha = 1.5$				$\alpha = 2.0$				$\alpha = 2.5$			
n	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$	$k=1$	$k=2$	$k=5$	$k=10$
10	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.5	0.6	0.5	0.6	0.6
20	0.65	0.7	0.75	0.65	0.65	0.75	0.65	0.6	0.6	0.65	0.45	0.4	0.6	0.5	0.75	0.6
50	0.64	0.66	0.76	0.72	0.64	0.62	0.68	0.74	0.6	0.68	0.64	0.6	0.58	0.6	0.32	0.42
100	0.64	0.57	0.75	0.72	0.64	0.63	0.64	0.72	0.61	0.58	0.46	0.62	0.59	0.39	0.44	0.54

Table 4: Empirical estimates of $r_1(S_{n,k})$ based on 5000 iterations for each setting (i.e., entry in table) for each possible set of recommendations. This table illustrates the regime in which $\alpha \geq 1$. In this regime, Theorem 2.A, which applies directly to the case $k = 1, \alpha > 1$, and $n \to \infty$, suggests that representation should be proportional to p_t when $\alpha = 1$ representation should approach equal as α increases. Our empirical results appear to roughly reflect this trend for all pairs (n, k).

B COMPUTATIONAL EXPERIMENTS

We provide additional details about the experiments we conduct in Section 5.

B.1 Finite number of recommendations and beyond unit consumption constraints

We explain how we computed empirically optimal sets in 5.1. For each recommendation setting, we determine the set that maximizes $\texttt{util}_k$ by manually computing

$$\max \left\{ \sum_{t \in [2], i \in [n_t]} V_{t,i}, \quad k \right\} \tag{23}$$

for sets with all possible combinations of item type representations. Specifically, in computing $S_{n,k}$, we consider the sets of the form $(i, n-i)$ for $i \in \{0, 1, \cdots, n\}$. For each of these sets S, we compute $\texttt{util}_k(S)$ directly, and take the average over 5000 iterations. We then choose the set with the maximum empirical expected utility, and display $r_1(S)$ in the Table 2. We consider additional settings, focusing on settings with varying q_t and $\alpha > 1$, in Table 3 and Table 4 respectively.

B.2 Continuous Items and User Preferences

We provide details for our experiment in Section 5.2 on GoodReads data [55]. We used a subset of interaction data from 1000 users and 200 books, which we used to compute embeddings.[10] We considered the 206 users that had at least 20 book interactions.

Let V be the set of all 200 book embeddings and V_i be the set of books the user i has interacted with. Suppose that V_{train} is the train set of 10 book embeddings randomly drawn from V_i. Suppose that $V_i' = V_i \setminus V_{\text{train}}$ is the remaining set of embeddings.

Recall that a user's value of an embedding v given their current preference v_{pref} is given by $u(v, v_{\text{pref}}) := \max\{0, v \cdot v_{\text{pref}}\}$. Given a set of n recommendations $S \subset S^d$, we evaluate it by drawing a random embedding v_{test} from $V_i' \cap V_i$ as the user's current preference and considering the objectives

$$\texttt{acc}(S) = \frac{1}{n} \sum_{v \in S} u(v, v_{\text{test}}) \tag{24}$$

$$\texttt{util}_1(S) = \max_{v \in S} u(v, v_{\text{test}}), \tag{25}$$

the naturally analogs of acc and $\texttt{util}_1$ we considered in our theoretical results.

[10] The main dataset can be found at https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/shelves. Embeddings were trained using the matrix factorization library libFM [48], which can be found at http://libfm.org/.

We constructed two sets of recommendations for each user given the training set V_{train}: the accuracy-maximizing set S_n and a diverse set S_{diverse}. We construct these sets as follows.

S_n is the set that maximizes average accuracy when user preferences are drawn from the train set:

$$\frac{1}{10} \sum_{v_{\text{pref}} \in V_{\text{train}}} \frac{1}{n} \sum_{v \in S_n} u(v, v_{\text{pref}}). \tag{26}$$

Computationally, we can determine this set S by choosing the n individual embeddings v in V' that maximize

$$\frac{1}{10} \sum_{v_{\text{pref}} \in V_{\text{train}}} u(v, v_{\text{pref}}). \tag{27}$$

We choose S_{diverse} using a heuristic method. We iterate over items in the train set, and select the closest item to the train set in terms of cosine distance that has not yet been selected. At iteration $i \in [n]$, we let v_{pref} be the $i \pmod{10}$-th item in V_{train} and $S_{\text{diverse},i-1}$ be the set of $i-1$ items selected so far. Then we construct $S_{\text{diverse},i}$ by adding the item in V' that maximizes

$$v \cdot v_{\text{pref}} \tag{28}$$

to the set $S_{\text{diverse},i-1}$, where $S_{\text{diverse},0} = \emptyset$. We then choose $S_{\text{diverse}} = S_{\text{diverse},n}$.

To evaluate the diversity of a set S of n recommendations, we use the average cosine distance between embeddings:

$$\frac{1}{\binom{n}{2}} \sum_{v,v' \in S} 1 - v \cdot v'. \tag{29}$$

The results we report are averages over all users and over 100 independently drawn training sets for each user.