Reconciling the accuracy-diversity trade-off in recommendations

Kenny Peng, Manish Raghavan, Emma Pierson, Jon Kleinberg, Nikhil Garg*

Draft: May 21, 2023

Abstract

In recommendation settings, 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 separately from accuracy. This approach, however, leaves a basic question unanswered: Why is there a trade-off in the first place?

We analyze a stylized model of recommendations reconciling this trade-off. Accounting for a user's capacity constraints (users do not typically make use of all the items that are recommended to them), optimal recommendations in our model are inherently diverse. Thus, accuracy and diversity appear misaligned because traditional accuracy metrics do not consider capacity constraints. Our model yields precise and interpretable characterizations of diversity in different settings, giving practical insights into the design of diverse recommendations.

1 Introduction

A large body of work in recommendations has developed methods to navigate an apparent trade-off between the goals of *accuracy* (to recommend items a user is most likely to want) and *diversity* (to recommend items from a range of categories) [1–23]. Real-world recommender systems use heuristics to directly incorporate diversity into recommendations [24, 25], and empirical evidence demonstrates that users prefer diverse recommendations [26–29].

A fundamental question remains: Why is there a trade-off in the first place? More specifically, why is "accuracy" unaligned with a user's true preference for diversity? Without a principled understanding of the accuracy-diversity trade-off, attempts to diversify recommendations have difficulty moving beyond a heuristic basis—and difficulty articulating what they are accomplishing at a deeper level.

In this work, we introduce and analyze a stylized model of recommendations that helps explain and reconcile the apparent accuracy-diversity trade-off. Our model studies this trade-off through the lens of a user's capacity constraints: users typically examine a list of recommendations and use only the top (highest value) options. (A person can watch only one movie in an evening, a recruiter can select only a handful of candidates to interview.) This lens has strong explanatory power:

- When the goal is to maximize the expected value of the *top* recommended items, diverse recommendations are often optimal in our model. By accounting for a user's capacity constraints, optimizing for what they are most likely to want is aligned with recommending a diverse set of items.
- Without accounting for capacity constraints, optimal recommendations in our model are homogeneous. This suggests that the trade-off between accuracy and diversity can be explained by commonly-used accuracy metrics not accounting for capacity constraints.

A strength of our model is that we can precisely and interpretably analyze the optimal *amount* of diversity in different settings, allowing us to isolate the effect of capacity constraints. This precision also yields practical insights about the role of diversity in recommendations. In Theorem 3, for example, we uncover natural settings where optimal recommendations overrepresent the category of item a user is *least* likely to want—a paradox that appears in grocery stores, where even though customers are less likely to buy ice cream than milk, stores allocate much more space to ice cream.

*Kenny Peng, Cornell Tech, kennypeng@cs.cornell.edu; Manish Raghavan, MIT; Emma Pierson, Cornell Tech; Jon Kleinberg, Cornell University; Nikhil Garg, Cornell Tech.

1.1 Model

There are m types of items indexed by $[m] = \{1, 2, \cdots, m\}$. A user prefers exactly one type of item, so that they prefer type $t \in [m]$ with probability p_t. The value of the $i-$th item of type t is a random variable $X_i^{(t)}$ if the user prefers type t and 0 otherwise (so its expected value is $p_t \mathbb{E}[X_i^{(t)}]$). The recommender knows only how $X_i^{(t)}$ are distributed, not their realizations. We refer to p_t as the ***likelihood*** of a type and the random variable $X_i^{(t)}$ as a ***conditional item value*** (the value of an item conditional on the user preferring the item's type).

Let $S_{n,k}$ be the set of n items that maximizes the expected total value of the k items with the highest realized values. We call $S_{n,k}$ the ***optimal*** set of n recommendations with respect to this objective. (We omit the dependency of $S_{n,k}$ on the other model parameters, as these will be clear from context.)

$S_{n,k}$ arises naturally from an assumption that the user can only use k items, so that they derive value from the k highest value recommended items. So $S_{n,1}$ results from an assumption that the user uses only one item, while $S_{n,n}$ results from an assumption that the user uses *every* recommended item.

$S_{n,n}$ maximizes the expected total value of all of the recommended items, so by linearity of expectation, $S_{n,n}$ contains the n items with the highest individual expected values. Meanwhile, $S_{n,1}$ maximizes the expected value of the highest-value item, an objective that—importantly—is not maximized by choosing the individual items with the highest expected values.

Our overarching technical result and interpretation. Our main results (summarized in Table 1) show across several settings that:

> $S_{n,k}$ is diverse for k fixed and n growing. Meanwhile, $S_{n,n}$ is homogeneous.

This technical result has an interpretation explaining the accuracy-diversity trade-off. Objectives that account for a user's capacity constraints naturally induce diversity, while objectives that do not account for capacity constraints can produce homogeneous recommendations. Thus, the observed trade-off is (partly) a consequence of common accuracy metrics not modeling capacity constraints.

One might suggest that our results show homogeneity is desirable when users can use all recommended items. We do not emphasize this interpretation since there are additional reasons outside of our model to incorporate diversity. Thus, our results are best interpreted as showing (1) that diversity arises under *minimal* assumptions, and (2) that standard objectives can (mistakenly) induce homogeneity.

Evaluating diversity. To evaluate diversity, we consider the representation of each type in $S_{n,k}$. In many of our results, representation falls on an interpretable continuum: from complete diversity (each type represented equally) to proportional diversity (each type represented proportionally to its likelihood p_t) to complete homogeneity (only the highest likelihood type represented). We formalize our approach to evaluating diversity in Section 2.

Discussion and limitations. Our model is purposefully stylized and minimal so that we can abstract a small set of ingredients common to a wide set of domains—namely, those where users prefer a certain category of items and a recommender has noisy estimates of user preferences and item values. Importantly, our model captures the capacity constraints of users: when a user is given a set of recommendations, they can typically use only the best few recommendations—a person can watch only one movie in an evening, a recruiter can select only a handful of candidates to interview.

Other reasons for diversity. Importantly, when we refer to *optimal* sets of recommendations, we mean optimal with respect to our stylized optimization problem—maximizing the expected sum of the k highest item values. In real-world contexts, there are many reasons to incorporate diversity that we do not consider here. For example, our model does not consider any explicit preferences users have for diversity. While such a preference may be empirically well-grounded, omitting such a preference makes our conclusions stronger: simply by modeling user capacity constraints, we show that diversity arises naturally in optimal recommendations *even when* our model does not explicitly value diversity.

Model generalizations. Real-world settings can differ from our model in natural ways. For example, users can prefer multiple types of items at a time and some items may fall under multiple types. We discuss

Table 1: Our main results broadly show: (1) capacity constraints induce diverse recommendations, and (2) without accounting for capacity constraints, recommendations tend to be homogeneous.

Setting	With capacity constraints	Without capacity constraints
Thm. 1. $X_i^{(t)} \sim \mathcal{D}$	As n grows large for fixed k, $S_{n,k}$ exhibits diversity depending on the tail behavior of $\mathcal{D}$. For non-heavy-tailed $\mathcal{D}$, $S_{n,k}$ is at least proportionally diverse.	$S_{n,n}$ contains only one type of item.
Thm. 2. $X_i^{(t)} \sim \text{Ber}(q_i)$ for $q_1, q_2, \cdots$ decaying by a power law (roughly, $q_i \propto i^{-\alpha}$).	For moderate amounts of decay ($\alpha < 1$), as n grows large, $S_{n,1}$ represents each item type equally.	For moderate amounts of decay ($\alpha < 1$), as n grows large, $S_{n,n}$ is less than proportionally diverse.
Thm. 3. $X_i^{(t)} \sim \text{Ber}(q_t)$ for $q_1, q_2, \cdots, q_m$.	For large n, $S_{n,1}$ represents each item type in proportion to $\frac{1}{\log \frac{1}{1-q_t}}$, so that items of *lower* success probability are recommended *more.*	$S_{n,n}$ contains only one type of item.

generalizations in Appendix C. Here, the assumptions that users prefer only one type and items each fall under one type allow us to analyze diversity in an interpretable way: characterizing how represented each item type is in comparison to the user's likelihood of preferring that type.

1.2 An illustrative example: Recovering Steck's standard of calibration

We now consider a simple instantiation of our model, motivated by a thought experiment suggested by the Netflix researcher Harald Steck [30]. A user watches romance movies 70 percent of the time and action movies the other 30 percent of the time. Steck raises a concern that an accuracy-maximizing algorithm can produce entirely homogeneous recommendations in this setting, and proposes a standard of *calibration*, where 70 percent of recommended movies here are romance and 30 percent are action. The example below reflects Steck's concern by showing that—before accounting for capacity constraints—optimal recommendations are homogeneous. Yet, by assuming a user can only watch one movie, the optimal recommendations in the example are in fact *calibrated*.

Example 1 (Recovering calibration)**.** Suppose there are two genres of movies, romance and action, indexed 1 and 2 respectively. A user prefers romance with probability p_1 and action with probability p_2, with $p_1 > p_2$. Movies from the user's preferred genre have values drawn i.i.d. from an exponential distribution $\text{Exp}(\lambda)$, while other movies have value 0. In the language of our model, $X_i^{(t)} \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$.

$S_{n,n}$ maximizes the expected total value of recommended items, so by linearity of expectation it contains the items of highest individual expected value. The expected value of each romance and action movie are $p_1\mathbb{E}[\text{Exp}(\lambda)]$ and $p_2\mathbb{E}[\text{Exp}(\lambda)]$ respectively, so $S_{n,n}$ contains only romance movies.

$S_{n,1}$ maximizes the expected value of the best recommended movie. The expected value of the best item among a_1 romance movies and a_2 action movies (where $a_1 + a_2 = n$) is

$$p_1 \cdot \mathbb{E}[\text{max of } a_1 \text{ draws from } \text{Exp}(\lambda)] + p_2 \cdot \mathbb{E}[\text{max of } a_2 \text{ draws from } \text{Exp}(\lambda)] \tag{1}$$

$$\approx \frac{p_1}{\lambda} \log(a_1) + \frac{p_2}{\lambda} \log(a_2), \tag{2}$$

which is maximized when $a_1 = p_1 n$ and $a_2 = p_2 n$ (as shown using Lagrange multipliers). Therefore, $S_{n,1}$ has proportional representation, recovering Steck's standard of calibration.

1.3 Intuition for our overarching result

In the model we propose, a user's value depends only on the recommended items from the user's preferred type. Therefore, conditional on a user preferring type t, the user's expected value from a set of recommendations with a_t items of type t is given by $h_t(a_t)$ for some function h_t. Then the user's expected value from a set of recommendations with a_t items of type t for all $t \in [m]$ is of the form

$$\sum_{t=1}^{m} p_t h_t(a_t). \tag{3}$$

The key idea of our work is that when accounting for capacity constraints, the function h_t reflects diminishing returns: recommending additional items from a type becomes less and less valuable when we care only about the value of the best recommended items—when a user can only watch one movie, recommending three options for romance movies might be significantly better than two, because the platform doesn't know which exact romance movie the user may like; however, recommending 20 such movies is barely better than 19. Thus, when maximizing (3), it becomes preferable to recommend items from other types. Meanwhile, without modeling user capacity constraints, there are not necessarily diminishing returns since the user can use all of the additional items recommended.

The particular shape of the diminishing returns regulates the amount of resulting diversity. In Example 1, we had that $h_t(x) \propto \log x$, in which case (3) is maximized when $a_t \propto p_t$, yielding proportional representation.[1] Our technical work thus involves analyzing the functions h_t in different settings (this reduces to analyzing order statistics of different distributions). Roughly speaking, heavier-tailed conditional item values imply larger marginal returns, resulting in less diversity.

1.4 Related work

As we have noted, there is a wide literature devoted to developing methods to navigate the accuracy-diversity trade-off [1–23]. Such work is supported by empirical evidence suggesting that a combination of these two metrics is preferred by users [26–29]. Of particular interest to us, however, is work that focuses on objectives that *implicitly* correspond to diversity. In the context of web search, optimizing for the probability that the user is shown a satisfactory search result has been associated with diversification [33, 34], since effective search results must account for different intents of queries ("pandas" can refer to an animal or a Python package). More recent work has adapted this objective to the recommender system setting as a metric that unifies accuracy and diversity [35]. Our work, by explicitly characterizing optimal amounts of diversity, identifies capacity constraints as the underlying reason for why such approaches result in diversity. We discuss additional related work in Appendix A.

2 Evaluating diversity

We now formalize our approach to evaluating diversity. For a set S of items, we define

$$r_t(S) := \frac{\#\text{ of items in } S \text{ of type } t}{|S|}, \tag{4}$$

the **representation** of type t in S. Intuitively, a set of recommendations is diverse if all types are well represented. We now define an interpretable family of representations that interpolates between maximum diversity and maximum homogeneity, and which arises naturally in many of our results.

Definition 1 (γ-homogeneity)**.** A set S is **γ-homogeneous** if for all $t \in [m]$,

$$r_t(S) = \frac{p_t^{\gamma}}{\sum_{i=1}^{m} p_i^{\gamma}}. \tag{5}$$

[1] The mathematical result in this example also appears in the context of resource allocation (e.g., [31]) and betting (e.g., §22.2 in [32]) in the presence of logarithmic utility.

γ-homogeneity captures several intuitive notions of diversity, using $p_1, \cdots, p_m$ as a benchmark:

- When $\gamma = 0$, $r_t(S) = \frac{1}{m}$. There is "equal representation."
- When $\gamma = 1$, $r_t(S) = p_t$. There is "proportional representation," where an item type is represented in proportion to its likelihood.
- When $\gamma = \infty$, $r_t(S) = 1$ for $t = \arg\max_{i \in [m]} p_i$ and $r_t(S) = 0$ otherwise. There is "complete homogeneity," where only the highest-likelihood item type is represented.

A smaller γ corresponds to more diversity, with $\gamma \leq 1$ indicating *at least proportional* representation. In practice, it is challenging to show that individual sets are γ-homogeneous; for one, since sets have an integer number of items from each type, it is typically impossible to obtain the exact ratios in (5). Instead, we will give primarily asymptotic results, showing that as n grows large, the optimal set $S_{n,k}$ approaches γ-homogeneity. Formally, we define γ-homogeneity over sequences of sets:

Definition 2 (γ-homogeneity for set sequences)**.** A sequence of sets $\{S_n\}_{n=1}^{\infty}$ is **γ-homogeneous** if for all $t \in [m]$,

$$\lim_{n \to \infty} r_t(S_n) = \frac{p_t^{\gamma}}{\sum_{i=1}^{m} p_i^{\gamma}}. \tag{6}$$

One perhaps surprising aspect of our results is that γ-homogeneity is sufficient to characterize diversity in a large class of settings, as opposed to requiring more complicated functions of proportions p_t.

3 Main results

We now state our main results, which consider several settings reflecting different assumptions about the conditional item values $X_i^{(t)}$. In Section 3.1, we assume that $X_i^{(t)} \overset{\text{iid}}{\sim} \mathcal{D}$ are drawn from a shared distribution, which implies that the recommender has little information about the value of specific items. In Section 3.2, $X_i^{(t)}$ are Bernoulli random variables with success probabilities differing depending on i and t, meaning that the recommender has information about which items are more likely to satisfy a user.

In each setting, we analyze the diversity of optimal sets $S_{n,k}$ for when k is fixed (i.e., accounting for capacity constraints) and $k = n$ (i.e., not accounting for capacity constraints). We sketch our proof strategy in Section 4 and defer full proofs to Appendix D.

3.1 i.i.d. conditional item values

Consider the setting in which $X_i^{(t)} \overset{\text{iid}}{\sim} \mathcal{D}$, i.e., conditional item values are drawn from a shared distribution. This implies that the values of items behave similarly across types and within types, and the platform cannot easily distinguish between the items in a type. From the perspective of a hiring platform, there may be many candidates with similar backgrounds (e.g., education or work history), none of whom can be distinguished from another by the platform. Conditional on a recruiter preferring this background, candidate values can be modeled as coming from a shared distribution.

We show that for a fixed k, as n grows large, the diversity of $S_{n,k}$ theoretically varies between equal representation, proportional representation, and near-complete homogeneity depending on the tail-behavior of $\mathcal{D}$. In particular, distributions that are bounded or have exponential tails induce at least proportional representation. Meanwhile, $S_{n,n}$ is completely homogeneous.

Theorem 1. *Suppose $X_i^{(t)} \overset{\text{iid}}{\sim} \mathcal{D}$ where $\mathcal{D}$ has finite mean. Then the following statements hold.*

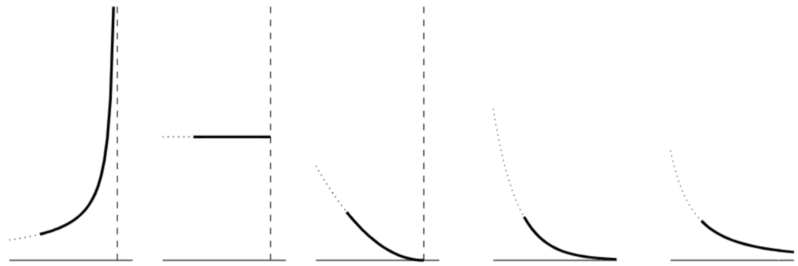
(i) ***[Finite Discrete]*** *If $\mathcal{D}$ is a finite discrete distribution, $\{S_{n,k}\}_{n=1}^{\infty}$ is 0-homogeneous.*

(ii) ***[Bounded]*** *If $\mathcal{D}$ has support bounded from above by M with pdf $f_{\mathcal{D}}$ satisfying*

$$\lim_{x \to M} \frac{f_{\mathcal{D}}(x)}{(M-x)^{\beta-1}} = c \tag{7}$$

for some $\beta, c > 0$, then $\{S_{n,k}\}_{n=1}^{\infty}$ is $\frac{\beta}{\beta+1}$-homogeneous.
(This pdf class contains beta distributions, including the uniform distribution.)

Figure 1: Theorem 1 summary. For $X_i^{(t)} \overset{\text{iid}}{\sim} \mathcal{D}$, distributions $\mathcal{D}$ with heavier tails induce less diversity.

	bounded Thm. 1(ii)			**exp. tail** Thm. 1(iii)	**heavy tail** Thm. 1(iv)
example $\mathcal{D}$		Beta$(\cdot, \beta)$		Exp(λ)	Pareto(α)
	$0 < \beta < 1$	$\beta = 1$	$\beta > 1$	$\lambda > 0$	$\alpha > 1$
graph of pdf					
$\{S_{n,k}\}_{n=1}^{\infty}$ γ-homog. for $\gamma \in$	$(0, 1/2)$	$1/2$	$(1/2, 1)$	1 (i.e., proportional)	$(1, \infty)$
		$\longleftarrow$ more diverse		less diverse $\longrightarrow$	

(iii) ***[Exponential tail]*** *If $\mathcal{D} = \text{Exp}(\lambda)$ for $\lambda > 0$, then $\{S_{n,k}\}_{n=1}^{\infty}$ is 1-homogeneous.*

(iv) ***[Heavy tail]*** *If $\mathcal{D} = \text{Pareto}(\alpha)$ for $\alpha > 1$, then $\{S_{n,k}\}_{n=1}^{\infty}$ is $\frac{\alpha}{\alpha-1}$-homogeneous.*

Additionally,

(v) $S_{n,n}$ contains only items of type $t = \arg\max_{t \in [m]} p_t$.

As Figure 1 illustrates, the theorem shows how for fixed k, the diversity of optimal solutions depends on the tail behavior of $\mathcal{D}$. In fact, we can obtain γ-homogeneity for any γ:

Corollary 1. *For any $\gamma \geq 0$, there exists $\mathcal{D}$ such that when $X_i^{(t)} \overset{\text{iid}}{\sim} \mathcal{D}$ and k is fixed, $\{S_{n,k}\}_{n=1}^{\infty}$ is γ-homogeneous.*

Intuitively, heavy-tailed distributions (part (iv)) induce less diverse recommendations since the marginal returns of recommending more items from the same type remains high: drawing more samples from a heavy-tailed distribution produces ever-increasing item values. This contrasts with bounded distributions like the uniform distribution (part (ii)), where once an item has close to the maximum value, additional draws of that type will not further improve the utility significantly.

A result for finite n and larger k. One limitation of our main results is that they are asymptotic ($n \to \infty$) and are restricted to fixed capacity constraints k. Stronger results can be obtained by considering specific distributions. For example, the result below characterizes for any n, k the representation of each type when conditional item values are uniformly distributed on $[0, 1]$.

Proposition 2. *When $X_i^{(t)} \overset{\text{iid}}{\sim} U([0, 1])$,*

$$\left| r_t(S_{n,k}) - \frac{\sqrt{p_t}}{\sum_{i=1}^{m} \sqrt{p_i}} \right| \leq \frac{m+1}{n}. \tag{8}$$

for all $k \leq \frac{\sqrt{p_m}}{\sum_{i=1}^{m} \sqrt{p_i}} n - m - 1$.

Therefore, for any n, $S_{n,k}$ is approximately $\frac{1}{2}$-homogeneous. In addition, for any k that is smaller than a constant fraction of n, the diversity $S_{n,k}$ does not depend on k. Thus, even for small sets of recommendations and large capacity constraints, diversity is optimal in this setting. We further give simulated results for small n and large k in Appendix B corroborating our theoretical results.

3.2 Heterogeneous Bernoulli conditional item values

Now consider when $X_i^{(t)}$ are independent random variables drawn from $\text{Ber}(q_i^{(t)})$, reflecting a model in which items have binary values (i.e., a user is either satisfied or not satisfied by an item). In this section, we allow $q_i^{(t)}$ to differ across i and t, implying that the recommender has knowledge about which items are more likely to be successful conditional on a user's preferred type. Specifically, in Section 3.2.1 we allow $q_i^{(t)}$ to vary across i and in Section 3.2.2 we allow $q_i^{(t)}$ to vary across t.

Our results will focus on $S_{n,1}$ and $S_{n,n}$, which both have natural interpretations in this setting:

- $S_{n,1}$ maximizes the probability that the user will be satisfied by at least one recommended item.
- $S_{n,n}$ maximizes the the expected number of recommended movies the user will be satisfied by, which is equivalent to the standard metric of accuracy.

Before proceeding, we note that the basic case $q_i^{(t)} = q$ for all i, t is handled as a direct corollary of Theorem 1(i).

Corollary 3 (Conditional item values are i.i.d. Bernoulli). *When $X_i^{(t)} \stackrel{\text{iid}}{\sim} \text{Ber}(q)$ for $q > 0$, then $S_{n,1}$ is 0-homogeneous.*

Therefore, if the success probability is the same for all items, optimal solutions are 0-homogeneous (each item is equally represented) for large n, even as the likelihoods p_t vary across type.

3.2.1 Decaying success probabilities

We now consider a setting in which among items of the same type, the recommender knows that some items have higher success probability. This maps onto settings where the recommender knows which items of a type are most likely to be satisfactory, e.g., some action movies are more commonly liked than are others. Thus, we assume that the recommender has access to items with decaying success probabilities.

Theorem 2 (Decaying success probabilities). *Suppose that $X_i^{(t)} \stackrel{\text{iid}}{\sim} \text{Ber}(q_i^{(t)})$ are i.i.d. Bernoulli random variables such that $q_i^{(t)} = c(i+d)^{-\alpha}$ for all $i \geq 1$ and some $\alpha, c, d \geq 0$. Then the following statements hold.*

(i) $\{S_{n,1}\}_{n=1}^{\infty}$ *is 0-homogeneous for* $\alpha < 1$.

(ii) $\{S_{n,1}\}_{n=1}^{\infty}$ *is* $\frac{1}{1+c}$*-homogeneous for* $\alpha = 1$.

(iii) $\{S_{n,1}\}_{n=1}^{\infty}$ *is* $\frac{1}{\alpha}$*-homogeneous for* $\alpha > 1$.

Additionally,

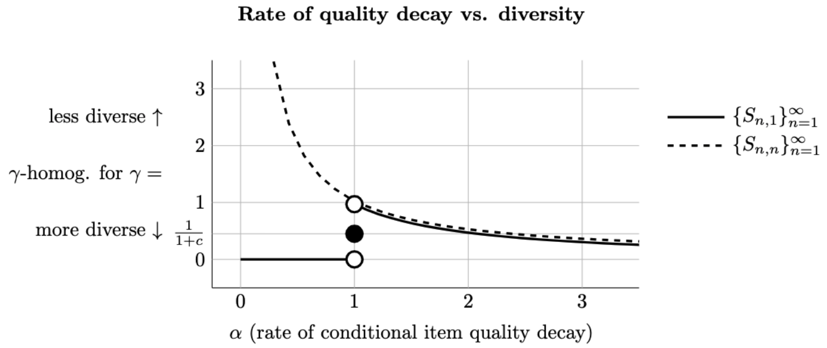
(iv) $\{S_{n,n}\}_{n=1}^{\infty}$ *is* $\frac{1}{\alpha}$*-homogeneous for* $\alpha \geq 0$.

When the success probabilities of items have moderate decay ($\alpha < 1$), then 0-homogeneity is maintained in the case $k = 1$ (note that $\alpha = 0$ recovers Corollary 3). Moreover, for all rates of decay, optimal recommendations reflect at least proportional diversity for large n and $k = 1$.

Theorem 2 also reveals surprising *non-monotonic* behavior. In particular, there is a discontinuity at $\alpha = 1$, where homogeneity suddenly increases, but then decreases as α continues to increase. At $\alpha = 1$, the optimal amount of diversity when $\alpha = 1$ can range between 0 and 1 depending on c.

When $k = n$, a larger rate of decay induces more diverse recommendations. Intuitively, when there is a larger rate of decay, the recommender has fewer high-quality options of a given type and is more incentivized to recommend high-quality options of other types. Note that for moderate rates of decay ($\alpha < 1$), $S_{n,n}$ remains less than proportionally diverse for large n, unlike $S_{n,1}$.

Figure 2: Theorem 2 considers items decay in quality within each type. For moderate rates of decay ($\alpha < 1$), $S_{n,k}$ is completely diverse for large n while $S_{n,n}$ is less than proportionally diverse.



3.2.2 Varying success probability across types

We now consider a setting in which the success probability of an item varies across types. This can arise when a users are more picky for some types of items, or when the recommender has more information about items from one type than another.

Theorem 3 (Varying success probability across types)**.** *Suppose that for each fixed t, $X_i^{(t)} \overset{\text{iid}}{\sim} \text{Ber}(q_t)$ are i.i.d. Bernoulli random variables. Then*

$$\lim_{n\to\infty} r_t(S_{n,1}) \propto \frac{1}{\log \frac{1}{1-q_t}} \tag{9}$$

while $S_{n,n}$ contains only items of type $t = \arg\max_{t\in[m]} p_t q_t$.

The surprising high-level takeaway from Theorem 3 is that, for large n, a *smaller* success probability q_t results in *more* representation of type t. The less likely an item of a given type is satisfactory, the more that type is recommended. Moreover, note that the amount of representation in this setting is independent of the popularities $p_1, \cdots, p_m$.

This paradox is illustrated in grocery stores, where more space is devoted to ice cream than milk, despite milk being much more popular than ice cream. Here, p_1 (the popularity of milk) is higher than p_2 (the popularity of ice cream). However, q_1 (the likelihood a given milk product satisfies a shopper looking for milk) is also higher than q_2 (the likelihood a given ice cream product satisfies a shopper looking for ice cream), since people tend to have more specific tastes for ice cream. Thus, since q_2 is smaller than q_1 and the grocery store should "recommend" many more ice creams than milks, explaining why more space is devoted to ice cream. Intuitively, while more shoppers want milk, these consumers can be satisfied with a small selection of milk; thus, it is more beneficial to devote more space to ice cream, for which shoppers have more specific tastes.

4 Sketch of Proofs

We now sketch the proof of Theorem 1(ii). (The general strategy applies to the rest of Theorem 1, as well as Theorem 2 and Theorem 3.) First, recall the general setup in Theorem 1, where the conditional item values $X_i^{(t)}$ are drawn i.i.d. from a shared distribution $\mathcal{D}$. In other words, a user prefers type $t \in [m]$ with probability p_t such that the user prefers exactly one type of item, and conditioned on the user preferring type t, the value of an item of that type is drawn i.i.d. from $\mathcal{D}$. We are interested in analyzing, depending on $\mathcal{D}$, the composition of $S_{n,k}$, the set of n items maximizing the expected sum of the k highest value items in the set. If $S_{n,k}$ contains $a_t^{(n)}$ items of type t, we need to analyze

$$\lim_{n\to\infty} r_t(S_{n,k}) = \lim_{n\to\infty} \frac{a_t^{(n)}}{n}. \tag{10}$$

We first provide an expression for the expected sum of the k highest value items in a set S with a_t items of type t. The following definition will be useful.

Definition 3. Define $\mu_{\mathcal{D}}(i, a)$ to be the expected value of the i-th order statistic[2] of a random variables drawn i.i.d. from $\mathcal{D}$. (Thus, $\mu_{\mathcal{D}}(1, a)$ is the expected minimum of a i.i.d. draws from $\mathcal{D}$ and $\mu_{\mathcal{D}}(a, a)$ is the expected maximum.)

Then conditioned on the user preferring type t, the expected sum of the k highest value of items in S is equal to $h(a_t) := \sum_{i=1}^{\min\{k,a_t\}} \mu_{\mathcal{D}}(a_t - i + 1, a_t)$, which follows from the linearity of expectation. Therefore, the expected sum of the k highest value items in S is $\sum_{t=1}^{m} p_t h(a_t)$. Define A_n to be the set of tuples of non-negative integers whose entries sum to n. Then $(a_1^{(n)}, a_2^{(n)}, \cdots, a_m^{(n)}) = \arg\max_{(a_1,\cdots,a_m)\in A_n} \sum_{t=1}^{m} p_t h(a_t)$.

We can then determine the limit in (38) given asymptotic information about h. In Lemma D.1 in the appendix, we develop general technical machinery for this task. Below, we state Lemma D.1(ii), which can be used to prove Theorem 1(ii).

Lemma D.1(ii). If h is monotonically increasing and there exist constants $A, B > 0$ and $\sigma < 0$ such that $\lim_{a\to\infty} \frac{A-h(a)}{Ba^{\sigma}} = 1$, then $\lim_{n\to\infty} \frac{a_t^{(n)}}{n} = \frac{p_t^{\frac{1}{1-\sigma}}}{\sum_{i=1}^{m} p_i^{\frac{1}{1-\sigma}}}$.

Then, considering $\mathcal{D}$ as in Theorem 1(ii), we can prove the necessary asymptotic result about h:

Lemma 1. *If $\mathcal{D}$ has support bounded from above by M with pdf $f_{\mathcal{D}}$ such that $\lim_{x\to M} \frac{f_{\mathcal{D}}(x)}{(M-x)^{\beta-1}} = c$ for some $\beta, c > 0$, then $\lim_{a\to\infty} \frac{Mk-h(a)}{Ba^{-\frac{1}{\beta}}} = 1$.*

Combining Lemma 1 with Lemma D.1(ii), with $\sigma = -\frac{1}{\beta}$, we show that for $\mathcal{D}$ as in Theorem 1(ii),

$$\lim_{n\to\infty} \frac{a_t^{(n)}}{n} = \frac{p_t^{\frac{\beta}{\beta+1}}}{\sum_{i=1}^{m} p_i^{\frac{\beta}{\beta+1}}}. \tag{11}$$

5 Conclusion

We introduce and analyze a stylized model that reconciles the apparent accuracy-diversity trade-off in recommendations. We characterize the diversity of optimal sets both when "optimality" captures and does not capture user capacity constraints. Broadly speaking, we show that the former naturally induces diversity while the latter results in homogeneity. Therefore, the apparent accuracy-diversity trade-off is partially due to traditional accuracy metrics not accounting for capacity constraints.

Limitations and future work. A particular strength of our model is that we were able to derive precise and interpretable characterizations of diversity in many settings. One limitation of our work is that many of our results are asymptotic (i.e., $n \to \infty$). We expect that it is possible to obtain further results in our model for finite n. We gave Proposition 2 as one such example, and outline additional possible directions in Appendix B.

The purposefully stylized nature of our model—in which users prefer only one type of item and items fall under only one type—allows for a particularly interpretable evaluation of diversity where we compare the representation of each type in comparison to the likelihood that the user prefers that type. This simple model is expressive enough to admit a wide range of results. Still, it would be interesting to generalize our findings in more complex models of users and items. In Appendix C, we discuss these possibilities further,

[2] The i-th order statistic of n random variables is the i-th smallest of the n realized values.

providing some basic results both when users can prefer multiple types of items and when users and items are represented by embeddings.

Finally, there are possible model characteristics beyond capacity constraints that implicitly reward diversity. As discussed in Section 1.3, diversity in our model is a consequence of diminishing returns when recommending additional items of a type. Diminishing returns may arise from other assumptions, such as decaying user attention (e.g., [18]). We point those interested in considering these alternate assumptions to Lemma D.1 in the appendix, in which we abstract the technical relationship between diminishing returns and diversity.

Broader impacts. Recommender systems have broad societal consequences, particularly in high-stakes settings like employment. Our work does not model many additional reasons for diversity in these settings, including those based on fairness and equity. We emphasize again that our results are intended to convey *minimal* assumptions that induce diversity and illustrate the tendency of certain objectives to produce homogeneous recommendations. Designers of recommender systems, and researchers in the area, should take a broader view when making decisions regarding diversity.

References

[1] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-based systems*, 123:154–162, 2017.

[2] Gediminas Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24:896–911, 2012.

[3] Shaina Raza and Chen Ding. Deep neural network to tradeoff between accuracy and diversity in a news recommender system. *2021 IEEE International Conference on Big Data (Big Data)*, pages 5246–5256, 2021.

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

[5] Georgios Alexandridis, Georgios Siolas, and Andreas Stafylopatis. Accuracy versus novelty and diversity in recommender systems: A nonuniform random walk approach. In *Recommendation and Search in Social Networks*, 2015.

[6] Zhipeng Hou and Jing Liu. A two-phase evolutionary algorithm for solving the accuracy-diversity dilemma in recommendation. *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020.

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

[8] Bibek Paudel, Thilo Haas, and Abraham Bernstein. Fewer flops at the top: Accuracy, diversity, and regularization in two-class collaborative filtering. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017.

[9] Xiaoyun He. Does utilizing online social relations improve the diversity of personalized recommendations? *Int. J. Strateg. Decis. Sci.*, 13:1–15, 2022.

[10] Ignacio Fernández-Tobías, Paolo Tomeo, Iván Cantador, T. D. Noia, and Eugenio Di Sciascio. 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.

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

[12] Maurizio Ferrari Dacrema. 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.

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

[14] Farzad Eskandanian and Bamshad Mobasher. 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.

[15] Shengqi Wu, Huaizhen Kou, Chao Lv, Wanli Huang, Lianyong Qi, and Hongya Wang. Service recommendation with high accuracy and diversity. *Wirel. Commun. Mob. Comput.*, 2020:8822992:1–8822992:10, 2020.

[16] Emanuel Lacić, Dominik Kowald, Markus Reiter-Haas, Valentin Slawicek, and E. Lex. Beyond accuracy optimization: On the value of item embeddings for student job recommendations. *ArXiv*, abs/1711.07762, 2017.

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

[18] Jon M. Kleinberg, Emily Ryu, and Éva Tardos. Calibrated recommendations for users with decaying attention. *ArXiv*, abs/2302.03239, 2023.

[19] Sinan Seymen, Himan Abdollahpouri, and Edward C. Malthouse. A constrained optimization approach for calibrated recommendations. *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021.

[20] Himan Abdollahpouri, Zahra Nazari, Alex Gain, Clay Gibson, Maria Dimakopoulou, J. Anderton, Benjamin Carterette, Mounia Lalmas, and Tony Jebara. Calibrated recommendations as a minimum-cost flow problem. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023.

[21] Mi Zhang and Neil J. Hurley. Avoiding monotony: improving the diversity of recommendation lists. In *ACM Conference on Recommender Systems*, 2008.

[22] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. Optimal greedy diversity for recommendation. In *International Joint Conference on Artificial Intelligence*, 2015.

[23] Tao Zhou, Zoltan Kuscsik, Jianguo Liu, Matú Medo, Joseph R. Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107:4511 – 4515, 2008.

[24] Ivan Medvedev, Taylor Gordong, and Haotian Wu. Powered by ai: Instagram's explore recommender system. 2019.

[25] David Goldberg. Diversity in search. 2014.

[26] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, WWW '20, page 2155–2165, New York, NY, USA, 2020. Association for Computing Machinery.

[27] João Sá, Vanessa Queiroz Marinho, Ana Rita Magalhães, Tiago Lacerda, and Diogo Gonçalves. 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.

[28] Jae kyeong Kim, Il Young Choi, and Qinglong Li. Customer satisfaction of recommender system: Examining accuracy and diversity in several types of recommendation approaches. *Sustainability*, 13:6165, 2021.

[29] Sung-Hyuk Park and Sang Pil Han. From accuracy to diversity in product recommendations: Relationship between diversity and customer retention. *International Journal of Electronic Commerce*, 18:51 – 72, 2013.

[30] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 154–162, 2018.

[31] Mo Ghorbanzadeh, Ahmed Abdel-Hadi, and Thomas Charles Clancy. A utility proportional fairness radio resource block allocation in cellular networks. *2015 International Conference on Computing, Networking and Communications (ICNC)*, pages 499–504, 2014.

[32] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.

[33] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, page 5–14, New York, NY, USA, 2009. Association for Computing Machinery.

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

[35] Javier Parapar and Filip Radlinski. Towards unified metrics for accuracy and diversity for recommender systems. *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021.

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

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

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

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

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

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

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

[43] James Mario Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment optimization under variants of the nested logit model. *Oper. Res.*, 62:250–273, 2014.

[44] Srikanth Jagabathula. Assortment optimization under general choice. *Available at SSRN 2512831*, 2014.

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

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

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

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

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

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

[51] Jaime Carbonell and Jade Goldstein. 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*, pages 335–336, 1998.

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

[53] William Brown and Arpit Agarwal. Diversified recommendations for agents with adaptive preferences. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

[55] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021.

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