Predicting Choice with Set-Dependent Aggregation

Nir Rosenfeld [1] **Kojin Oshiba** [1] **Yaron Singer** [1]

Abstract

Providing users with alternatives to choose from is an essential component of many online platforms, making the accurate prediction of choice vital to their success. A renewed interest in learning choice models has led to improved modeling power, but most current methods are either limited in the type of choice behavior they capture, cannot be applied to large-scale data, or both.

Here we propose a learning framework for predicting choice that is accurate, versatile, and theoretically grounded. Our key modeling point is that to account for how humans choose, predictive models must be expressive enough to accommodate complex choice patterns but structured enough to retain statistical efficiency. Building on recent results in economics, we derive a class of models that achieves this balance, and propose a neural implementation that allows for scalable end-to-end training. Experiments on three large choice datasets demonstrate the utility of our approach.

1. Introduction

One of the most prevalent activities of online users is *choosing*. In almost any online platform, users constantly face choices: what to purchase, who to follow, where to dine, what to watch, and even simply where to click. As the prominence of online services becomes ever more reliant on such choices, the accurate prediction of choice is quickly becoming vital to their success. The availability of large-scale choice data has spurred hopes of feasible individual-level prediction, and many recent works have been devoted to the modeling and prediction of choice (Benson et al., 2016; Ragain & Ugander, 2016; Kleinberg et al., 2017b;a; Mottini & Acuna-Agost, 2017; Shah & Wainwright, 2017; Negahban et al., 2018; Chierichetti et al., 2018; Overgoor et al., 2018).

In a typical choice scenario, a user is presented with a set of items $s = \{x^{(1)}, \dots, x^{(n)}\}$, $x^{(i)} \in \mathcal{X} = \mathbb{R}^d$, called the *choice set*, with $s \in \mathcal{S} \subseteq 2^{\mathcal{X}}$. From the alternatives in s, the user chooses an item $y \in s$; in economics this is known as the problem of *discrete choice* (Luce, 1959). We follow the standard machine learning setup and assume choice sets s and choices y are drawn i.i.d. from an unknown joint distribution D. Given a set of m examples $T = \{(s_i, y_i)\}_{i=1}^m$, our goal is to learn a choice predictor $h(s)$ that generalizes well to unseen sets, i.e., has low expected error w.r.t. D.

A natural way to predict choice is to learn an item score function $f(x)$ from a class of score functions $\mathcal{F}$ and use it to model the predicted probability of choosing x from s as $P_s(x) = e^{f(x)} / \sum_{x' \in s} e^{f(x')}$. If the learned $f \in \mathcal{F}$ scores chosen items higher than their alternatives, then $P_s(x)$ will provide useful predictions. While this approach seems appealing, it is in fact constrained by an undesired artifact known as the *Independence of Irrelevant Alternatives* (IIA):

Definition 1. *(Luce, 1959) P is said to satisfy **IIA** if for all $s \in \mathcal{S}$ and for any $a, b \in s$, it holds that*

$$P_{\{a,b\}}(a) / P_{\{a,b\}}(b) = P_s(a) / P_s(b)$$

IIA states that the likelihood of choosing a over b should not depend on what other alternatives are available. This means that the prediction rule $h_f(s) = \operatorname{argmax}_{x \in s} P_s(x)$ considers items with no regard of context, i.e., the alternatives available in s. IIA therefore imposes a rigid constraint on the types of choice behavior that can be expressed by the model—a fundamental limitation of this approach that cannot be mitigated simply by increasing the complexity of functions in $\mathcal{F}$ (e.g., adding layers or clever non-linearities).

From a practical point of view, this is discouraging, as there is ample empirical evidence that real choice data exhibits regular and consistent violations of IIA (see Rieskamp et al. (2006) for an extensive survey). This has led to a surge of interest in machine learning models that go beyond IIA (Oh & Shah, 2014; Osogami & Otsuka, 2014; Benson et al., 2016; Ragain & Ugander, 2016; Otsuka & Osogami, 2016; Ragain & Ugander, 2018; Chierichetti et al., 2018; Seshadri et al., 2019; Pfannschmidt et al., 2019).

The naïve way to avoid IIA is to directly model all possible subsets, but this is likely to make learning intractable (McFadden et al., 1977; Seshadri et al., 2019). A common approach for resolving this difficulty is to impose structure,

[1]School of Engineering and Applied Sciences, Harvard University. Correspondence to: Nir Rosenfeld <nirr@seas.harvard.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

typically in the form of a probabilistic choice model encoding certain inter-item dependencies. While allowing for violations of IIA, this approach presents several practical limitations. First, explicitly modeling the dependency structure restricts *how* IIA can be violated, which may not necessarily align with the choice patterns in the data. Second, many of these models are designed to satisfy certain choice axioms or asymptotic properties (e.g., consistency) rather than to predict accurately when trained on finite-sample data, limiting their practical effectiveness and making it hard to theoretically reason about their predictive capabilities. Finally, surprisingly few of these methods can be applied to large-scale online choice data, where the number of instances can be prohibitively large, choice sets rarely appear more than once, and items can be complex structured objects whose number is virtually unbounded.

To complement these works, and motivated by the growing need for choice models that are accurate and scalable, here we propose a framework for learning context-dependent choice models that are directly optimized for accuracy. Our framework is based on the idea of *set-dependent aggregation* (SDA), a principled approach in economics for modeling set-dependent choice (Ambrus & Rozen, 2015), that to the best of our knowledge has not yet been considered from a machine learning perspective. As we show, aggregation provide a means for aligning model complexity with the behavioral complexity of the underlying choice patterns, letting the data determine how and to what extent IIA should be relaxed.

The key challenge in designing a good choice model lies in properly balancing its expressivity and efficiency. On the one hand, the model must be flexible enough to capture the ways in which set-dependence is expressed in the data; on the other hand, it must be structured enough so that learning can be carried out efficiently. In this paper, we show that set-dependent aggregation achieves both. Our framework makes the following contributions:

- **Efficient and scalable discriminative training.** Our approach is geared towards predictive performance: it directly optimizes for accuracy, can be efficiently trained end-to-end, and scales well to realistically large and complex choice-prediction scenarios.
- **Behavioral inductive bias.** We propose a novel class of parametric aggregators, based on key principles from behavioral decision theory and implemented with a novel neural architecture. Empirically, our models are both more accurate *and* more compact than alternatives.
- **Rigorous error analysis.** Our theoretical results include bounds for two types of errors. For approximation error, our bound quantifies the capacity of aggregation to relax IIA as a function of model complexity. For estimation error, we give concise generalization bounds, thus establishing the learnability of aggregation.
- **Thorough empirical evaluation.** We conduct experiments on three large choice datasets—flight itineraries, hotel reservations, and news recommendations—demonstrating the utility of our approach. Our analysis gives insight as to how aggregation improves performance, complementing and supporting our theoretical results.

Overall, our work presents a practical and theoretically-grounded approach for predicting choice.

Paper organization. Our paper considers aggregation across multiple levels of granularity; we begin with a very general functional form, proceed by proposing a concrete structured model, and conclude with a practical neural implementation. After reviewing the related literature (Sec. 1.1), in Sec. 2 we introduce our model of set-dependent aggregation in gradually-increasing detail. The rest of the paper follows this structure: Our main theoretical results, given in Sec. 3, include an approximation error bound (Sec. 3.1) that applies to very general aggregators and estimation error bounds (Sec. 3.2) that apply to a large family of structured models. In Sec. 4 we present an experimental evaluation of our neurally-implemented aggregation model. We end with concluding remarks in Sec. 5.

1.1. Related material

IIA begins with Luce's Axiom of Choice (Luce, 1959), which for a certain noise distribution, induces the popular Multinomial Logit model (MNL) (McFadden et al., 1973; Train, 2009). Two common extensions, Nested MNL (McFadden, 1978) and Mixed MNL (McFadden & Train, 2000), relax IIA by grouping items via a tree structure or modeling a mixture population, respectively. These, however, impose restrictive assumptions on the nature of violations, require elaborate hand-coded auxiliary inputs, and are in many cases intractable. Although recent progress has alleviated some difficulties (Oh & Shah, 2014; Benson et al., 2016; Chierichetti et al., 2018), applying these models in practice remains difficult. Recent works have proposed other probabilistic models that deviate from IIA, by modeling pairwise utilities (Ragain & Ugander, 2018), k^{th}-order interactions (Ragain & Ugander, 2016; Seshadri et al., 2019), and general subset relations (Benson et al., 2018). These, however, do not optimize for predictive accuracy, and rarely apply to complex choice settings with many items and sparse choice sets. Others suggest discriminative solutions, but these include models that tend to be either overly-specific or overly-general (Mottini & Acuna-Agost, 2017; Pfannschmidt et al., 2019) and provide no guarantees as to how IIA is relaxed.

Our work draws on the literature of utility aggregation (or "multi-self" models), studied extensively in economics (Kalai et al., 2002; Fudenberg & Levine, 2006; Manzini & Mariotti, 2007; Green & Hojman, 2009; Ambrus & Rozen, 2015), psychology (Tversky, 1972; Shafir et al., 1993; Tversky & Simonson, 1993), and marketing (Kivetz et al., 2004). Whereas most models of this type focus on mathematical tractability or behavioral plausibility, our focus is on statistical and computational aspects relevant to machine learning. Our work is largely inspired by recent results in economics on the expressivity of aggregation (Ambrus & Rozen, 2015). While they give worst-case guarantees for a finite-item non-parametric setting under a realizability assumption, we consider the statistical aspects of learning parametric aggregators from data, in theory and in practice.

2. Proposed Model

At the core of our approach is the idea of *set-dependent aggregation* (SDA), where a collection of ℓ item-wise score functions $\boldsymbol{f} = (f_1, \ldots, f_\ell)$, $f_i \in \mathcal{F}$, are combined to produce a single set-wise score function $g_{\boldsymbol{f}}(x|s)$, which we will refer to as an *aggregator*. Aggregation in itself is an established concept in machine learning (e.g., bagging or boosting), but to express choice that deviates from IIA, aggregation must be done in a set-dependent manner.[1] We propose to consider aggregators of the form:

$$g_{\boldsymbol{f}}(x|s;\psi) = \sum_{i=1}^{\ell} \psi(f_i(x)|f_i(s)) \tag{1}$$

where $\psi : \mathbb{R} \times 2^{\mathbb{R}} \to \mathbb{R}$ is an *aggregation mechanism* remapping each score $f_i(x)$ conditional on the scores of all alternatives $f_i(s) = \{f_i(x')\}_{x' \in s}$. With a slight overload of notation, we can think of $\boldsymbol{f}$ as embedding items x in an ℓ-dimensional latent space via $\boldsymbol{f}(x) = \tilde{x} \in \mathbb{R}^\ell$, $\tilde{x}_i = f_i(x)$ (and in accordance $\tilde{s} = \{\tilde{x}\}_{x \in s}$ and $\tilde{s}_i = \{\tilde{x}_i\}_{x \in s}$). Eq. (1) therefore introduces set-dependence by applying to each dimension in the embedded space the set-dependent operator $\psi(\tilde{x}_i|\tilde{s}_i)$, independently and uniformly, and aggregating.

We can now define the class of functions we would like to learn, which we refer to as the *aggregator class*:

$$\mathcal{G} = \{g_{\boldsymbol{f}}(x|s;\psi) \,:\, \boldsymbol{f} \in \mathcal{F}^\ell, \psi \in \Psi\} \tag{2}$$

where Ψ is a (possibly parameterized) class of mechanisms. We will refer to ℓ as the *dimension* of aggregation, and to the item-wise score function class $\mathcal{F}$ as the *base class*. Given $\mathcal{G}$, our goal in learning will be to jointly learn score functions $\boldsymbol{f} \in \mathcal{F}^\ell$ and mechanism $\psi \in \Psi$ whose corresponding g provides good predictions via the decision rule:

$$\hat{y} = h_{g_{\boldsymbol{f}}}(s) = \operatorname*{argmax}_{x \in s} g_{\boldsymbol{f}}(x|s;\psi) \tag{3}$$

[1] Specifically, any method based on linear aggregation (i.e., $\sum_i w_i f_i(x)$), such as bagging or boosting, is inherently IIA.

2.1. Inductive Bias

The ability of functions in $\mathcal{G}$ to capture set-dependent choice relies on the expressive power of Ψ. Our first result in Sec. 3.1 demonstrates that even simple mechanisms can express intricate deviations from IIA. However, for practical purposes, it is important to add structure to g, which we do by encoding inductive bias into ψ. From Sec. 3.2 and thereafter we consider aggregation mechanisms of the form:

$$\psi(\tilde{x}_i|\tilde{s}_i) = w(\tilde{s}_i)\mu\left(\tilde{x}_i - r(\tilde{s}_i)\right) \tag{4}$$

Here, $w : 2^{\mathbb{R}} \to \mathbb{R}$ and $r : 2^{\mathbb{R}} \to \mathbb{R}$ are set functions (i.e., permutation invariant) and $\mu : \mathbb{R} \to \mathbb{R}$ is an asymmetric s-shaped function, all of which are shared across dimensions and learned. Together, these give our primary model:

$$g_{\boldsymbol{f}}(x|s;w,r,\mu) = \sum_{i=1}^{\ell} w(\tilde{s}_i)\mu\left(\tilde{x}_i - r(\tilde{s}_i)\right) \tag{5}$$

Our design choices follow from four key principles elicited from the literature on behavioral decision making, describing how users perceive value. First, users consider value across multiple (perhaps latent) dimensions (Tversky, 1972). These are captured in Eq. (5) by the different f_i. Second, within each dimension, value is relative, and is considered in relation to a set-dependent reference point (Tversky & Kahneman, 1991). In the model, reference points are determined by r, and the relative value in dimension i is $\tilde{x}_i - r(\tilde{s}_i)$. Third, the perception of losses and gains is a-symmetric and diminishing, resulting in loss aversion—the cornerstone of Prospect Theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). This is modeled using the s-shaped μ. Finally, the degree to which each valuation dimension contributes to the overall perceived value is also set-dependent (Tversky, 1969; Tversky & Simonson, 1993). In Eq. (5), the importance of each dimension is determined by w. Appendix A includes a simple illustration of these principles.

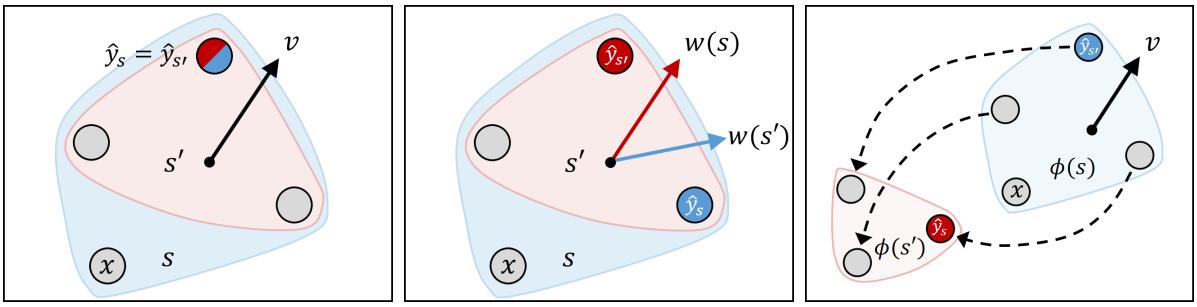
The above construction generalizes many choice models from economics, psychology, and marketing:

Claim 1. *The aggregation model in Eq. (5) subsumes the choice models of Tversky (1969); McFadden et al. (1973); McFadden (1978); Kaneko & Nakamura (1979); Kalai et al. (2002); Kivetz et al. (2004), and Orhun (2009).*

In the above models, w and r are simple and fixed set operations (e.g., max, min, average; details in Appendix A). In contrast, in Sec. 4 we will parameterize these elements using flexible neural components, that with end-to-end training provide a significant boost in accuracy.

2.2. Interpretation

One interpretation of Eq. (5) is that g is a linear model operating on learned feature representations, wherein *both*



(a) **No set dependence (IIA)**: $g(x|s) = \langle v, \boldsymbol{\phi}(x) \rangle$

(b) **Set-dependent weights**: $g(x|s) = \langle \boldsymbol{w}(s), \boldsymbol{\phi}(x) \rangle$

(c) **Set-dependent representations**: $g(x|s) = \langle v, \boldsymbol{\phi}(x|s) \rangle$

Figure 1: An illustration of how set-dependent aggregation can relax IIA. Points represent an embedding of items in $\mathbb{R}^\ell$ with $\ell = 2$. Two sets are shown: s and $s' = s \setminus \{x\}$ for some $x \in s$, with corresponding predictions $\hat{y}_s$ and $\hat{y}_{s'}$. (a) IIA dictates that removing x from s must not change the prediction, i.e., $\hat{y}_s = \hat{y}_{s'}$. This is the case for linear aggregation $g(x|s) = \langle v, \boldsymbol{\phi}(x) \rangle$, where neither weights $v \in \mathbb{R}^\ell$ nor representations $\boldsymbol{\phi}(x)$ depend on s. (b) When aggregation includes set-dependent weights $\boldsymbol{w}(s)$, removing an item can change the scoring direction. (c) When aggregation includes set-dependent representations $\boldsymbol{\phi}(x|s)$, removing an item can change the spatial position of items. Both forms of set-dependence allow s and s' to have different maximizing items, and hence different predictions (i.e., $\hat{y}_s \neq \hat{y}_{s'}$).

weight vector and representations vary with the choice set. To see this, denote $\phi(\tilde{x}_i|\tilde{s}_i) = \mu(\tilde{x}_i - r(\tilde{s}_i))$, and set:

$$\begin{aligned} \boldsymbol{\phi}(x|s) &= (\phi(\tilde{x}_1|\tilde{s}_1), \ldots, \phi(\tilde{x}_\ell|\tilde{s}_\ell)) \\ \boldsymbol{w}(s) &= (w(\tilde{s}_1), \ldots, w(\tilde{s}_\ell)) \end{aligned} \tag{6}$$

As both $\boldsymbol{\phi}$ and $\boldsymbol{w}$ map into $\mathbb{R}^\ell$, Eq. (5) can be rewritten as:

$$g_{\boldsymbol{f}}(x|s; w, r, \mu) = \langle \boldsymbol{w}(s), \boldsymbol{\phi}(x|s) \rangle \tag{7}$$

suggesting that predicted items $\hat{y} = h_g(s)$ are those whose (set-dependent) representation $\boldsymbol{\phi}(x|s)$ maximizes the inner product with the (set-dependent) weight vector $\boldsymbol{w}(s)$.

Eq. (7) reveals two means by which set-dependent aggregation can support violations of IIA. Consider the removal of an item $x \neq \hat{y}$ from some choice set s. A direct consequence of IIA is that if $\hat{y}$ is predicted for s, then it is also predicted for any subset of s containing $\hat{y}$, and in particular for $s' = s \setminus x$. Set-dependent aggregation can mitigate this constraint in two ways. First, removing an item can change the feature representation of all other items, i.e., $\boldsymbol{\phi}(x'|s) \neq \boldsymbol{\phi}(x'|s')$ for $x' \in s'$. In this way, supposing $\boldsymbol{w}$ is kept fixed, removing an item can reposition items in representation space, resulting in a possibly different argmax of the inner product with $\boldsymbol{w}$. Second, removing an item can change the linear separator, i.e., $\boldsymbol{w}(s) \neq \boldsymbol{w}(s')$. Hence, supposing $\boldsymbol{\phi}$ is kept fixed, the scoring direction can rotate and the prediction can change. Fig. 1 illustrates these effects.

3. Theoretical Analysis

The complexity of $\mathcal{G}$ is controlled by three elements: the base class $\mathcal{F}$, the mechanism class Ψ, and the dimension ℓ. In the next sections we consider how these effect learning.

The goal of learning is to find some $g \in \mathcal{G}$ that minimizes the *expected risk* over the data distribution D:

$$\varepsilon(g) = \mathbb{E}_{(s,y)\sim D}[\Delta(y, h_g(s))] \tag{8}$$

where $\Delta(y, \hat{y}) = \mathbb{1}_{\{y \neq \hat{y}\}}$ is the 0/1 loss. In practice, learning typically involves minimizing the *empirical risk* (or a proxy thereof) over the sample set $T = \{(s_i, y_i)\}_{i=1}^m$:

$$\hat{\varepsilon}(g) = \frac{1}{m} \sum_{i=1}^{m} \tilde{\Delta}(y_i, h_g(s_i)) \tag{9}$$

where $\tilde{\Delta}$ is a proxy for Δ, and possibly under some form of regularization. A good function class is therefore one that balances between being able to fit the data well *in principle* (i.e., having a low optimal $\varepsilon(g)$) and *in practice* (i.e., having a low optimal $\hat{\varepsilon}(g)$ and a guarantee on its distance from the corresponding $\varepsilon(g)$). This can be seen by decomposing the expected risk into two error types—*approximation error* and *estimation error*:

$$\varepsilon(g) = \underbrace{\varepsilon(g^*)}_{\text{approx.}} + \underbrace{\varepsilon(g) - \varepsilon(g^*)}_{\text{estimation}}, \qquad g^* = \operatorname*{argmin}_{g' \in \mathcal{G}} \varepsilon(g')$$

In this section we bound both types of errors. For approximation error, we show how the best achievable error is controlled by the capacity of functions in $\mathcal{G}$ to relax IIA. For estimation error, we give generalization bounds establishing the learnability of a large family of aggregators.

3.1. Approximation Error

Our goal in this section is to bound the approximation error $\varepsilon(g^*)$, and more specifically, to reason about how it decreases as the complexity of $\mathcal{G}$ grows. We focus on ℓ as

the main handle on complexity, allowing for fairly arbitrary $\mathcal{F}$ and targeting a broad class of "natural" aggregators (of the general form in Eq. (1)) whose mechanism ψ satisfies rudimentary properties from choice theory (see Ambrus & Rozen (2015)). These ensure that $g \in \mathcal{G}$ are consistent (if $\ell = 1$ then g_f predicts like f) and scale invariant (scaling all $f \in \boldsymbol{f}$ by a constant $\alpha \in \mathbb{R}$ does not change predictions).

Without clear structural assumptions on ψ, however, directly analyzing the optimal error is challenging. Our approach will be to express the error of g^* in terms of optimal (conditional) errors of the aggregated f_i. This is useful since optimal errors are established for many classes of item-wise functions. Specifically, the bound partitions the space of all sets $\mathcal{S}$ into $k+1$ regions $C_0, \ldots, C_k$, and within each C_i, bounds the error of g^* using the conditional error of the optimal f_i, denoted $\varepsilon(f_i|C_i) = \mathbb{E}_D[\Delta(y, h_{f_i}(s))|s \in C_i]$. The bound holds for any partition that is "appropriate"; see Definition 2 below.

Theorem 1. *Let $\mathcal{G}$ be a natural class of aggregators of dimension $\ell = 5k+1$ for some $k \geq 0$. Then for any appropriate partition $C_0, \ldots, C_k$ of $\mathcal{S}$,*

$$\min_{g \in \mathcal{G}} \varepsilon(g) \leq \sum_{i=0}^{k} p_i \min_{f_i \in \mathcal{F}} \varepsilon(f_i|C_i) \tag{10}$$

where $p_i = \mathbb{P}_D[s \in C_i]$.

Theorem 1 states that the optimal aggregator g^* is as good as the *locally best* f_i within each C_i (note that $\sum_{i=0}^{k} p_i = 1$). This sheds light on how the capacity of g to express violations of IIA can increase with ℓ: When $\ell = 1$, g acts like a single f, and so cannot express violations of IIA. If the data exhibits violations, then g will err with frequency that is at least the frequency of violations. As ℓ grows, g can account for violations by "carving out" regions of $\mathcal{S}$ to which different item score functions are applied locally. This can break the prediction-inheritance property of IIA: for $s' \subset s$, if $s' \in C_i$ but $s \in C_j$, then $\hat{y}' = h_g(s') = h_{f_i}(s')$ can differ from $\hat{y} = h_g(s) = h_{f_j}(s)$, even if $\hat{y} \in s'$ (whereas IIA would dictate a shared prediction, i.e., $\hat{y}' = \hat{y}$).

Interestingly, the partitions considered in Theorem 1 depend on the base class $\mathcal{F}$ as well.

Definition 2. *A partition $C_0, \ldots, C_k$ of $\mathcal{S}$ is **appropriate** if there exist $f_1', \ldots, f_k'$, $f_i' \in \mathcal{F}$, such that for $i = 1, \ldots, k$:*

$$C_i = \{s : f_i'(s) > 0\} \setminus C_{i+1} \cup \cdots \cup C_k$$

where $f'(s) > 0$ is true if $f'(x) > 0 \; \forall x \in s$, and all remaining choice sets are in $C_0 = \mathcal{S} \setminus C_1 \cup \cdots \cup C_k$.

Thus, functions in $\mathcal{F}$ are used not only to score items but also to partition choice sets, and what partitions are attainable depends on what regularities in C_i the f_i' can capture. This dual role is key to our proof, but requires that $\mathcal{G}$ be defined over a function class that is slightly more expressive than $\mathcal{F}$. We now give a short proof sketch, deferring further details and the full proof to Appendix B.

Proof sketch. The proof is constructive. For any $f_0, \ldots, f_k$ and $f_1', \ldots, f_k'$, we construct an aggregator g that agrees (i.e., induces the same item ranking) with f_i on C_i (as determined by f_i'). The first step is to construct for each $i = 1, \ldots, k$ a small module $g^{(i)}$, based on f_i and f_i', that agrees with f_i on all s for which $f_i'(s) > 0$, and is "indifferent" otherwise. The main lemma shows that such $g^{(i)}$ can be constructed by having it implement certain objects called *triple bases*—the main building block of Ambrus & Rozen (2015), whose key result is that triple bases exist for the type of aggregator classes we consider. The next step is to combine the modules into a single predictor $g = \sum_{i=0}^{k} \alpha_i g^{(i)}$, where $\alpha_i \in \mathbb{R}$ and $g^{(0)}$ includes only f_0. Our choice of α_i determines an order of precedence of the modules over choice sets, thus preventing collisions and ensuring that g agrees with $g^{(i)}$ on all $s \in C_i$ (and only on those). The final step is to conclude that the optimal g is at least as good as the locally-best f_i on any appropriate partition $C_0, \ldots, C_k$. □

Theorem 1 carries two practical implications. First, the dimension ℓ acts as a budget on the model's capacity to relax IIA. As a handle on complexity, it allows the practitioner to align the complexity of the model with the "behavioral" complexity of the observed choice patterns. The coverage-like nature of C_i implies that increasing ℓ has a diminishing returns property—an effect we observe empirically in the experiments in Sec. 4. The second implication concerns the dual role of the base class $\mathcal{F}$. In standard settings, the choice of $\mathcal{F}$ considers whether it is likely to include score functions capable of identifying chosen items. Our result reveals an additional consideration, unique to aggregation: to perform well, $\mathcal{F}$ must also include functions capable of targeting *choice sets*, an in particular, those violating IIA.

Theorem 1 quantifies how the approximation error decreases as the complexity of $\mathcal{G}$ increases. Our next results quantifies how this trades off with estimation error.

3.2. Estimation Error

In this section we establish the learnability of aggregators, and in particular, those presented in Sec. 2.1. Our results include generalization bounds for two large families of aggregators, showing how the error of learning with $\mathcal{G}$ decreases with the number of samples m and as a function of ℓ and properties of ψ and $\mathcal{F}$. The proofs make use of Rademacher generalization bounds, which can control the estimation error in many learning settings (see Shalev-Shwartz & Ben-David (2014)). Our main contribution is the analysis of the

Rademacher complexity of aggregators, which we derive by utilizing their compositional structure (see Appendix C).

To simplify the analysis, we focus on set operations and on a linear base class:

$$\mathcal{F}^{\rho}_{\text{lin}} = \{\boldsymbol{f}(x) = x^{\top}\Theta \,:\, \Theta \in \mathbb{R}^{d\times\ell},\ \|\Theta\|_{\rho} \le 1\}$$

where $\|\cdot\|_{\rho}$ is the induced ρ-norm. This suffices to cover all subsumed models from Claim 1.[2] For any function q, we denote its Lipschitz constant by λ_q when it is scalar-valued and by λ^{ρ}_q when it is vector valued and with respect to the ρ-norm. We also use $c = \max_{x\in\mathcal{X}} \|x\|_{\infty}$. Both theorems invoke the Rademacher-based generalization bound in (Bartlett & Mendelson, 2002), which assumes that $\tilde{\Delta}$ dominates Δ, is 1-Lipschitz, and onto $[0,1]$. However, since our main result here is the Rademacher characterizations of aggregator classes, other bounds can similarly be applied.

The first bound applies to aggregators whose mechanism relies on set-dependent weights.

Theorem 2. *Let $\mathcal{G}$ be a class of aggregators over $\mathcal{F}^{\infty}_{lin}$ of the form $g(x|s) = \langle \boldsymbol{w}(s), \boldsymbol{\phi}(x)\rangle$. Then for any D and any $\delta \in [0,1]$, it holds that for all $g \in \mathcal{G}$,*

$$\varepsilon(g) \le \hat{\varepsilon}(g) + 4c^2\lambda^{\rho}_{w}\sqrt{\frac{2\log 2d}{m}} + O\Big(\sqrt{\frac{\log(1/\delta)}{m}}\Big)$$

with probability of at least $1-\delta$.

The second bound applies to aggregators whose mechanism relies on set-dependent representations.

Theorem 3. *Let $\mathcal{G}$ be a class of aggregators over $\mathcal{F}^{1}_{lin}$ of the form $g(x|s) = \langle v, \boldsymbol{\phi}(x|s)\rangle$ where $v \in \mathbb{R}^{\ell}$, $\|v\|_1 \le 1$, and $\boldsymbol{\phi}(x|s)$ is defined as in Eq. (6). Then for any D and any $\delta \in [0,1]$, it holds that for all $g \in \mathcal{G}$,*

$$\varepsilon(g) \le \hat{\varepsilon}(g) + 4c\lambda_{\mu}(1+\lambda^{\rho}_{r})\sqrt{\frac{2\log 2d}{m}} + O\Big(\sqrt{\frac{\log(1/\delta)}{m}}\Big)$$

with probability of at least $1-\delta$.

In Theorems 2 and 3, the dimension ℓ enters only through the linear dependence on the set-operation Lipschitz constants. This suggests that the estimation error of aggregation scales well in ℓ, and in particular when compared to that of the base class (a special case for which $\ell = 1$ and the Lipschitz constants vanish). To the best of our knowledge, these are the first generalization bounds for choice models. Together with Theorem 1, they show how aggregation can effectively balance expressivity and statistical efficiency.

4. Experiments

We now present our experimental evaluation on real choice data. Our goal here is to show that aggregation performs well and at scale, and to support our results from Sec. 3.

[2] All except for one of the models of Kivetz et al. (2004).

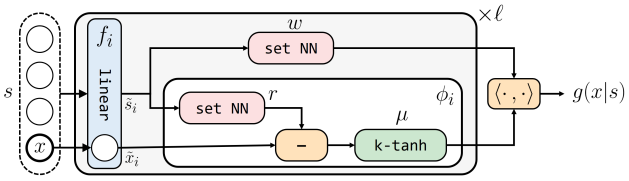


Figure 2: The proposed SDA architecture

Datasets. We evaluate our method on three large datasets: flight itineraries from Amadeus[3], hotel reservations from Expedia[4], and news recommendations from Outbrain[5]. Each dataset includes a collection of choice sets presented to users and their corresponding choices. We focus on examples where users chose (i.e., clicked on) exactly one item, which compose the vast majority of the data. Features describe items (e.g., price, quality, or category), context (e.g., query, date and time), and users (although for reasons of privacy, very little user information is available). Further details and dataset statistics are given in Appendix D.1.

Aggregation models. For our set-dependent aggregation approach (SDA), we use the model proposed in Sec. (2.1):

$$\textbf{SDA}: \qquad g_{\boldsymbol{f}}(x|s; w, r, \mu) = \sum_{i=1}^{\ell} w(\tilde{s}_i)\mu\left(\tilde{x}_i - r(\tilde{s}_i)\right)$$

where as before $\tilde{x}_i = f_i(x)$ and $\tilde{s}_i = \{\tilde{x}_i\}_{x\in s}$. We implement the above SDA model using a novel neural architecture (see Figure 2) in which the components w, r, and μ are parameterized as follows. Both set functions w and r are small permutation-invariant neural networks (Zaheer et al., 2017), each having 2 hidden layers of 16 units each with tanh activations and mean pooling. Motivated by the success of inner-product reference models (e.g., Vaswani et al. (2017)), we also experiment with a variant of SDA, referred to as SDA$_+$, having vector-valued score functions $F : \mathcal{X} \to \mathbb{R}^k$ and reference points $r : 2^{\mathbb{R}} \to \mathbb{R}^k$ that are compared using an inner product (see Appendix D.4 for details).

In both models, the loss aversion term μ is implemented as a 'kinked' tanh with slope parameter c:

$$\mu(z; c) = \tanh(z)(c\mathbb{1}_{\{z<0\}} + \mathbb{1}_{\{z\ge 0\}}), \quad c \ge 1$$

We set the base class $\mathcal{F}$ to include linear functions to allow for a clean differential comparison to other baselines. Our main results use $\ell = 24$, which strikes a good balance between performance and runtime (in general larger ℓ are better). For a comprehensive analysis of the contribution of each model component see the ablation study in Appendix D.6. Learned parameters include those in w, r, μ, and $\boldsymbol{f}$,

[3] See Mottini & Acuna-Agost (2017)

[4] www.kaggle.com/c/expedia-personalized-sort

[5] www.kaggle.com/c/outbrain-click-prediction

Table 1: Main results. Values are averaged over 10 random splits (standard errors in small font).

		Amadeus			Expedia			Outbrain		
		Top-1	Top-5	MRR	Top-1	Top-5	MRR	Top-1	Top-5	MRR
Ours	**SDA$_+$**	**45.42** ± 0.5	**93.37** ± 0.0	64.57 ± 0.3	**31.49** ± 0.2	**86.91** ± 0.2	**53.05** ± 0.1	**38.04** ± 0.3	**94.54** ± 0.1	**59.66** ± 0.2
	SDA	45.26 ± 0.4	93.23 ± 0.2	**64.68** ± 0.3	31.48 ± 0.2	86.73 ± 0.2	**53.05** ± 0.1	37.00 ± 0.3	94.66 ± 0.1	59.45 ± 0.2
IIA	MNL (McFadden et al., 1973)	38.42 ± 0.5	91.02 ± 0.3	59.53 ± 0.4	30.06 ± 0.2	86.34 ± 0.1	51.81 ± 0.2	37.74 ± 0.3	94.52 ± 0.2	59.41 ± 0.2
	SVMRank (Joachims, 2006)	40.27 ± 0.4	91.94 ± 0.3	60.91 ± 0.3	31.28 ± 0.2	86.24 ± 0.1	52.68 ± 0.2	37.68 ± 0.3	94.46 ± 0.1	59.65 ± 0.2
	RankNet (Burges et al., 2005)	37.44 ± 0.7	84.67 ± 1.7	54.51 ± 1.6	23.82 ± 0.5	81.85 ± 0.6	46.28 ± 0.5	35.32 ± 0.8	91.55 ± 0.7	57.56 ± 0.5
non-IIA	Mixed MNL (Train, 2009)	37.96 ± 0.3	90.40 ± 0.3	59.03 ± 0.3	27.28 ± 0.6	84.24 ± 0.3	50.21 ± 0.1	37.72 ± 0.3	94.42 ± 0.1	59.53 ± 0.2
	AdaRank (Xu & Li, 2007)	37.27 ± 0.4	72.34 ± 0.3	41.02 ± 0.3	26.70 ± 0.2	83.21 ± 0.2	48.74 ± 0.2	37.47 ± 0.3	94.40 ± 0.2	59.45 ± 0.2
	Deep Sets (Zaheer et al., 2017)	40.36 ± 0.5	91.92 ± 0.3	62.40 ± 0.5	29.87 ± 0.3	86.26 ± 0.2	51.66 ± 0.2	37.51 ± 0.3	94.30 ± 0.1	59.15 ± 0.2
basic	Price/Quality	36.44 ± 0.3	87.23 ± 0.2	43.44 ± 0.3	17.92 ± 0.1	77.67 ± 0.1	31.71 ± 0.2	24.17 ± 0.1	25.08 ± 0.1	31.49 ± 0.1
	Random	25.15 ± 0.5	32.87 ± 0.6	45.46 ± 0.4	14.13 ± 0.1	32.35 ± 0.2	27.65 ± 0.2	22.21 ± 0.1	23.10 ± 0.1	29.69 ± 0.2

which were trained to optimize the cross-entropy loss using Adam with step-wise exponential decay (see Appendix D.3).

Baselines. We consider baselines that can be applied to large-scale data, falling into one of three categories: (i) IIA models, (ii) non-IIA models, either with or without structural assumptions, and (iii) all aggregation models from Claim 1. For IIA methods, we use Multinomial Logit (MNL) (McFadden et al., 1973; Train, 2009), SVMRank (Joachims, 2006), and RankNet (Burges et al., 2005). These all consider a single item-wise score function, but differ in the way they are optimized. For non-IIA methods, we use a discrete Mixed MNL model (Train, 2009), AdaRank (Xu & Li, 2007), and Deep Sets (Zaheer et al., 2017). These differ in how they consider item dependencies: Mixed MNL relaxes IIA by modeling a mixture of user populations, ListNet captures set dependencies via the loss function, and Deep Sets attempts to universally approximate general set functions using a permutation-invariant architecture. We also add simple baselines based on price, quality, and random predictions. For details see Appendix D.2.

Setup. Results are based on averaging 10 random 50:25:25 train-validation-test splits. The methods we consider vary in their expressive power and computational requirements. Hence, for a fair comparison (and reasonable train times), each trial includes 10,000 randomly sampled examples. Performance measures include top-1 accuracy, top-5 accuracy, and mean reciprocal rank (MRR). For all methods we tuned regularization, dropout, and learning rate (when applicable) using Bayesian optimization. Default values were used for other hyper-parameters. For optimization we used Adam with step-wise exponential decay. Details in Appendix D.3.

4.1. Results

Our main results are presented in Table 1 which shows the performance of all methods for choice sets having at most 10 items (12 for Outbrain). As can be seen, SDA outperforms other baselines in all settings, with SDA$_+$ achieving slightly improved performance over SDA. These results are consistent across datasets and performance measures for other choice set sizes as well (see Appendix D.5).[6] While some methods perform adequately in some settings, they are generally inconsistent across datasets (e.g., Deep Sets), measures (e.g., SVMRank), or both (e.g., MNL).

Focusing on the Amadeus dataset, note that a simple baseline based on the cheapest price achieves a top-1 accuracy of 36.4. The performance of MNL, which is a special case of SDA wherein $\ell = 1$ and w, r, and μ are degenerate, improves performance by 2.0 accuracy points (5.4% increase). Introducing set-dependence using SDA improves performance by a considerable accuracy 8.8 points (24.2%), with SDA$_+$ improving further (9.0 points, 24.6%).

The above models present an increase in accuracy but also an increase in model complexity. Simply using more complex models, however, does not guarantee improvement: the optimal Deep Sets model has roughly 20-fold more parameters than SDA, but under-performs by a significant margin, indicating the importance of infusing models with appropriate inductive bias. Meanwhile, inductive bias alone does not suffice either. Figure 3 (left) compares SDA to the aggregation models from Claim 1, all of which are special cases of SDA having lightly or non-parameterized components. As can be seen, across all values of ℓ, the neural parameterization of SDA provides flexibility that is crucial for accuracy.

4.2. Analysis

Accuracy and model complexity. Figure 3 (right) presents the accuracy of SDA for increasing values of ℓ. Results show that accuracy steadily increases, but at a diminishing rate,

[6] All results are significant ($p < 10^{-5}$) under a Friedman test, and all comparisons to SDA are significant ($p < 0.007$) under a post-hoc pairwise signed-rank test (with Hochberg adjustment).

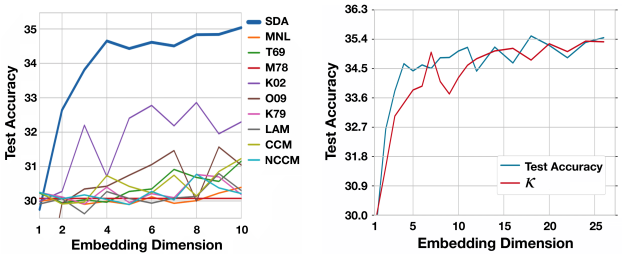


Figure 3: **Left**: Accuracy of SDA vs. the choice models it generalizes (Claim 1), demonstrating the benefit of replacing simple fixed operators with flexible neural components. **Right**: Accuracy (blue) and violation capacity κ (red) are highly correlated, suggesting that SDA improves by increasingly accounting for violations of IIA. Gains are diminishing, reaching 90% improvement by $\ell = 4$.

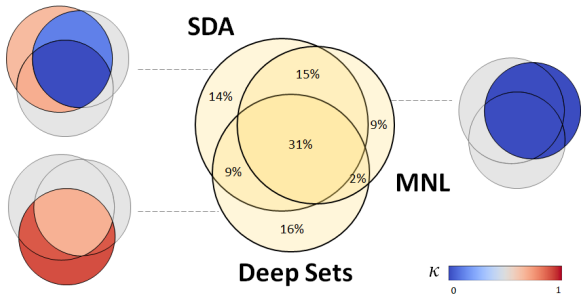


Figure 4: Diagram of overlap in correct predictions, colored by violation capacity (κ) per method per cross-section. MNL and Deep Sets represent two extremes: inability to express violations of IIA ($\kappa = 0$, dark blue), and over-flexible set-dependence ($\kappa = 1$, dark red), respectively. SDA strikes middle ground by properly allocating its violation "budget", focusing mostly on choice sets on which MNL errs (average κ, light red).

with roughly 90% of the gain achieved as early as $\ell = 4$. This aligns well with our conclusions from Sec. 3.1.

Theorem 1 implies that increasing ℓ can improve performance by allowing the model to better adapt to choice patterns deviating from IIA. To empirically verify this, we propose a measure of *violation capacity* that quantifies the capacity of a model to change its prediction when an item is removed from a choice set (here $\hat{y} = h_g(s)$):

$$\kappa(g) = \frac{1}{m} \sum_{s \in T} \frac{1}{|s| - 1} \sum_{x \in s \setminus \{\hat{y}\}} \mathbb{1}_{\{h_g(s \setminus \{x\}) \neq \hat{y}\}} \qquad (11)$$

Figure 3 (left) reveals a tight correlation between accuracy and violation capacity κ, suggesting that performance improves through the model's ability to increasingly relax IIA.

Targeted flexibility. One interpretation of the proof of Theorem 1 is that aggregators consist of one primary item-wise score function (f_0) whose decisions can be overruled by other score functions ($f_1, \ldots, f_k$) on certain regions of $\mathcal{S}$ (namely $C_1, \ldots, C_k$). Empirically, this implies that κ should vary across choice sets: high κ on sets in targeted regions, and low κ on all other sets.

To validate this, we compare SDA in detail to MNL and Deep Sets. The main diagram in Figure 4 presents the overlap in correct predictions across methods. For each method, the smaller diagrams present its violation capacity κ on each cross-section. As expected, MNL has $\kappa = 0$ in all sections. Deep Sets, meanwhile, shows extremely high κ values in most sections, suggesting it "overfits" in its set-dependent predictive flexibility. For SDA, we observe a mixed pattern: low κ on sections joint with MNL, and sufficiently (though not excessively) high κ on sections joint with Deep Sets. These results suggests that SDA targets not only the appropriate "amount" of relaxation of IIA, but also targets the appropriate choice sets for which IIA is relaxed.

5. Conclusions

In this work, we propose set-dependent aggregation as a framework for predicting human choice. There is a growing need for choice models that are accurate, scalable, and insightful. But the behavioral patterns of human choice are complex and intricate, requiring models to effectively balance expressivity and specificity. Our results suggest that aggregation achieves this balance, both in theory and in practice, and shed light on how this balance is achieved.

There are three main avenues in which our work can extend. First, our work focuses on a rudimentary choice task: choosing one item from a given set. There are, however, many other important choice tasks, of a sequential or combinatorial nature, each posing intriguing modeling challenges. Second, our theoretical results touch upon a connection between generalization and violation of IIA. We conjecture that this connection runs deep, and that a formal notion of "violation complexity" could be useful in characterizing the learnability of choice models in general. Third, our view of aggregation as a set-dependent linear model in a set-dependent latent space hints at the prospect of interpretability. Since our model is designed to capture in these latent representations the dimensions of perceived utility, an understanding of how aggregation relaxes IIA in latent space could contribute to our understanding of human choice at large. We leave these notions for future work.

Acknowledgments

This research was supported by NSF grant CAREER CCF 1452961, NSF CCF 1816874, BSF grant 2014389, NSF USICCS proposal 1540428, a Google Research award, and a Facebook research award.

References

Ambrus, A. and Rozen, K. Rationalising choice with multi-self models. *The Economic Journal*, 125(585):1136–1156, 2015.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Benson, A. R., Kumar, R., and Tomkins, A. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 963–973. International World Wide Web Conferences Steering Committee, 2016.

Benson, A. R., Kumar, R., and Tomkins, A. A discrete choice model for subset selection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 37–45. ACM, 2018.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96. ACM, 2005.

Chierichetti, F., Kumar, R., and Tomkins, A. Learning a mixture of two multinomial logits. In *International Conference on Machine Learning*, pp. 960–968, 2018.

Fudenberg, D. and Levine, D. K. A dual-self model of impulse control. *American economic review*, 96(5):1449–1476, 2006.

Green, J. and Hojman, D. Choice, rationality and welfare measurement. 2009.

Joachims, T. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 217–226. ACM, 2006.

Kahneman, D. and Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, March 1979.

Kalai, G., Rubinstein, A., and Spiegler, R. Rationalizing choice functions by multiple rationales. *Econometrica*, 70(6):2481–2488, 2002.

Kaneko, M. and Nakamura, K. The nash social welfare function. *Econometrica: Journal of the Econometric Society*, pp. 423–435, 1979.

Kivetz, R., Netzer, O., and Srinivasan, V. Alternative models for capturing the compromise effect. *Journal of marketing research*, 41(3):237–257, 2004.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2017a.

Kleinberg, J., Mullainathan, S., and Ugander, J. Comparison-based choices. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 127–144. ACM, 2017b.

Luce, R. D. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

Manzini, P. and Mariotti, M. Sequentially rationalizable choice. *American Economic Review*, 97(5):1824–1839, 2007.

McFadden, D. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.

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

McFadden, D., Tye, W. B., and Train, K. *An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model*. Institute of Transportation Studies, University of California, 1977.

McFadden, D. et al. Conditional logit analysis of qualitative choice behavior. 1973.

Mottini, A. and Acuna-Agost, R. Deep choice model using pointer networks for airline itinerary prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1575–1583. ACM, 2017.

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

Oh, S. and Shah, D. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, pp. 595–603, 2014.

Orhun, A. Y. Optimal product line design when consumers exhibit choice set-dependent preferences. *Marketing Science*, 28(5):868–886, 2009.

Osogami, T. and Otsuka, M. Restricted boltzmann machines modeling human choice. In *Advances in Neural Information Processing Systems*, pp. 73–81, 2014.

Otsuka, M. and Osogami, T. A deep choice model. In *AAAI*, pp. 850–856, 2016.

Overgoor, J., Benson, A. R., and Ugander, J. Choosing to grow a graph: Modeling network formation as discrete choice. *arXiv preprint arXiv:1811.05008*, 2018.

Pfannschmidt, K., Gupta, P., and Hüllermeier, E. Learning choice functions. *arXiv preprint arXiv:1901.10860*, 2019.

Ragain, S. and Ugander, J. Pairwise choice markov chains. In *Advances in Neural Information Processing Systems*, pp. 3198–3206, 2016.

Ragain, S. and Ugander, J. Choosing to rank. *arXiv preprint arXiv:1809.05139*, 2018.

Rieskamp, J., Busemeyer, J. R., and Mellers, B. A. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44 (3):631–661, 2006.

Seshadri, A., Peysakhovich, A., and Ugander, J. Discovering context effects from raw choice data. *arXiv preprint arXiv:1902.03266*, 2019.

Shafir, E., Simonson, I., and Tversky, A. Reason-based choice. *Cognition*, 49(1-2):11–36, 1993.

Shah, N. B. and Wainwright, M. J. Simple, robust and optimal ranking from pairwise comparisons. *Journal of machine learning research*, 18(199):1–199, 2017.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sridharan, K. Cornell cs6783 (machine learning theory), lecture notes: Rademacher complexity, 2014. URL: http://www.cs.cornell.edu/courses/cs6783/2014fa/lec7.pdf.

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

Tversky, A. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

Tversky, A. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

Tversky, A. and Kahneman, D. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4):1039–1061, 1991.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Tversky, A. and Simonson, I. Context-dependent preferences. *Management science*, 39(10):1179–1189, 1993.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Xu, J. and Li, H. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 391–398. ACM, 2007.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in Neural Information Processing Systems*, pp. 3391–3401, 2017.