

A Random Attention Model^{*}

Matias D. Cattaneo[†] Xinwei Ma[‡] Yusufcan Masatlioglu[§] Elchin Suleymanov[¶]

November 27, 2017

Abstract

We introduce a Random Attention Model (RAM) allowing for a large class of stochastic consideration maps in the context of an otherwise canonical limited attention model for decision theory. This model relies on a new restriction on the stochastic consideration map, termed *Monotonic Attention*, which is intuitive and nests many recent contributions in the literature, is implied by the classical Luce and Random Utility Models, and covers many other models of choice behavior. We develop revealed preference theory within RAM and obtain precise testable implications for observed choice probabilities. Using these results, we show that a set (possibly a singleton) of strict preference orderings compatible with RAM are identifiable from the decision maker's choice probabilities, and establish a representation of this identified set of unobserved preferences as a collection of inequality constraints on her choice probabilities. Given this nonparametric identification result, we develop uniformly valid inference methods for the (partially) identifiable preferences. We showcase the performance of our proposed econometric methods using simulations, and we also provide general-purpose software implementation of our estimation and inference results in the R software package `ramchoice`.

Keywords: revealed preference, limited attention models, random utility models, nonparametric identification, partial identification.

^{*}We thank Andres Santos and Joerg Stoye for very helpful comments and suggestions. Financial support from the National Science Foundation through grant SES-1628883 is gratefully acknowledged.

[†]Department of Economics and Department of Statistics, University of Michigan.

[‡]Department of Economics, University of Michigan.

[§]Department of Economics, University of Maryland.

[¶]Department of Economics, University of Michigan.

1 Introduction

Revealed preference theory not only is a cornerstone of modern economics, but also has important theoretical, methodological and policy implications for many social and behavioral sciences. This theory aims to identify the preferences of decision making units (e.g., households or firms) from their observed choices (e.g., buying a house or hiring a worker). In its classical formulation, revealed preference theory assumes that the decision maker selects the best available option after full consideration of all possible alternatives presented to her. This assumption leads to specific testable implications based on observed choice probabilities but, unfortunately, empirical testing of classical revealed preference theory shows that it is not always compatible with observed choice behavior.¹ For example, [Reutskaja, Nagel, Camerer, and Rangel \(2011\)](#) provides interesting experimental evidence against classical revealed preference theory under full attention using eye tracking and choice data.

Motivated by these findings, and the fact that certain theoretically important and empirically relevant choice patterns can not be explained using classical revealed preference theory based on full attention, scholars have proposed other economic models of choice behavior. An increasingly popular alternative is the limited attention model ([Masatlioglu, Nakajima, and Ozbay, 2012](#); [Lleras, Masatlioglu, Nakajima, and Ozbay, 2017](#); [Dean, Kibris, and Masatlioglu, 2017](#)), where decision makers are assumed to select the best available option after having decided to consider only a subset of all possible alternatives. This framework takes the consideration set formation, sometimes called attention rule, as unobservable and hence an intrinsic feature of the decision maker. Nonetheless, it is possible to develop a fruitful theory of revealed preference within this framework, employing only mild and intuitive nonparametric restrictions on how the decision maker decides to focus attention on specific subsets of all possible alternatives presented to her.

Until very recently, limited attention models have been deterministic, a feature that diminished their empirical applicability: Testable implications via revealed preference have been relying on the assumption that the decision maker will always pay attention to the same subset of options every time that she is confronted to the same set of available alternatives. Put differently, the framework assumes a single, representative decision maker and rules out (unobserved) heterogeneity in the

¹See, e.g., [Hauser and Wernerfelt \(1990\)](#), [Goeree \(2008\)](#), [van Nierop et al. \(2010\)](#) and [Honka et al. \(2017\)](#).

population. This is obviously restrictive, and can lead to predictions that are inconsistent with observed choice behavior. Aware of this fact, a few scholars have improved limited attention models by allowing for stochastic attention maps (Horan, 2013; Manzini and Mariotti, 2014; Aguiar, 2015; Brady and Rehbeck, 2016), a feature that allows the decision maker to pay attention to different subsets with some non-zero probability given the same set of alternatives to choose from. All available results in this literature first parametrize the consideration set formation (i.e., commit to particular a parametric attention rule) and then study the revealed preference implications of these (parametric) models.

As opposed to earlier approaches, in this paper we introduce a Random Attention Model (RAM) where we abstain from any particular (stochastic) consideration set formation, and instead consider a large class of nonparametric random attention rules. Our model imposes one intuitive condition, termed *Monotonic Attention*, which is satisfied by many possible stochastic consideration set formations. Given that consideration sets are unobservable, this feature is crucial for applicability of our revealed preference results, as our finding and empirical implications will remain valid for many different, particular attention rules that could be operating in the background. In other words, our revealed preference results will be derived from nonparametric restrictions on the consideration set formation and hence will be fully robust to misspecification.

Our key identifying assumption, *Monotonic Attention*, restricts the possibly stochastic attention formation process in a very intuitive way: Each consideration set competes for the decision maker's attention, and hence the probability of paying attention to a particular subset is assumed to not decrease when the total number of possible consideration sets decreases. We show that this single nonparametric assumption is general enough to nest most (if not all) previously proposed deterministic and random limited attention models, in addition to being satisfied in many other cases of interest (including the Luce model or classical random utility models). Furthermore, under our proposed monotonic attention assumption, we are able to develop a theory of revealed preference, obtain specific testable implications, and (partially) identify the underlying preferences of the decision maker by investigating her observed choice probabilities. Our revealed preference results are applicable to wide range of consideration set formations, including the parametric ones currently available literature because we show below that our generic *Monotonic Attention* property holds in those cases as well. Based on these theoretical findings, we also develop econometric results for

identification, estimation and inference of the decision maker’s preferences, as well as specification testing of RAM. To be specific, we show that RAM implies that the set of partially identified preference orderings (possibly a singleton) containing the decision maker’s true preference is equivalent to a set of inequalities restrictions on the choice probabilities (one for each preference ordering in the identified set). This result allows us to employ the identifiable choice probabilities to (i) develop a model specification test (i.e., test whether there exists a non-empty set of preference orderings compatible with RAM), (ii) conduct hypothesis testing on specific preference orderings (i.e., test whether the inequality constraints on the choice probabilities are satisfied), and (iii) develop confidence sets containing the decision maker’s true preference with pre-specified coverage (i.e., via test inversion), among other possibilities. Our econometric methods rely on ideas and results from the literature on partially identified models and moment inequality testing: see [Canay and Shaikh \(2017\)](#) and [Ho and Rosen \(2017\)](#) for a recent review and further references, and [Romano, Shaikh, and Wolf \(2014\)](#), [Chernozhukov, Chetverikov, and Kato \(2014\)](#), [Chernozhukov, Newey, and Santos \(2015\)](#), [Fang and Santos \(2015\)](#), and references therein, for a sample of recent contributions related to the estimation and inference approaches proposed herein.

Our work contributes to both economic theory and econometric literatures. We discuss in detail the connections and distinctions between this paper and the economic theory literature in Section 6, after we introduce our proposed RAM, give several decision theory examples, and present results concerning revealed preference theory. In particular, we show how RAM nests and/or connects to the recent work by [Manzini and Mariotti \(2014\)](#), [Brady and Rehbeck \(2016\)](#), [Gul, Natenzon, and Pesendorfer \(2014\)](#), [Echenique, Saito, and Tserenjigmid \(2014\)](#), [Echenique and Saito \(2017\)](#), [Fudenberg, Iijima, and Strzalecki \(2015\)](#) and [Aguiar, Boccoardi, and Dean \(2016\)](#), among others.

This paper is also related to a rich econometric literature on nonparametric identification, estimation and inference both in the specific context of Random Utility Models (RUMs) and more generally. See [Matzkin \(2013\)](#) for a review and further references on nonparametric identification, [Hausman and Newey \(2017\)](#) for a recent review and further references on nonparametric welfare analysis, and [Blundell, Kristensen, and Matzkin \(2014\)](#), [Manski \(2014\)](#), [Kitamura and Stoye \(2016\)](#), [Deb, Kitamura, Quah, and Stoye \(2017\)](#) and [Kawaguchi \(2017\)](#) for a sample of recent contributions and further references on nonparametric analysis under RUM. As mentioned above, a key feature of RAM distinguishing our main results from previous work is that our proposed Monotonic Attention

condition on consideration formation maps nests previous models as special cases and also covers many new models of choice behavior. In particular, RAM is more general than RUM, which is important because numerous studies in psychology, finance and marketing have shown that decision makers have limited attention when making choices: they only compare (and choose from) a subset of all available options. Whenever decision makers do not pay full attention to all options, implications from revealed preference theory under RUM do no longer hold in general, implying that empirical testing of substantive hypothesis as well as policy recommendations based on RUM will be invalid. On the other hand, our results may remain valid because RAM is a strict, non-trivial generalization of RUM.

The rest of the paper proceeds as follows. Section 2 presents the basic setup, Section 3 introduces our key monotonicity assumption on the decision maker’s stochastic consideration map, Section 4 discusses in detail our random attention model, including the main revealed preference results, and Section 5 shows that monotonic attention rules can be interpreted as a linear combination of deterministic attention filters. Given these theoretical results, Section 6 discusses in detail the connections between our work and the related economic theory literature. Section 7 presents our main econometrics methods, including nonparametric (partial) identification, estimation and inference results, while Section 8 summarizes the findings from a simulation study analyzing their performance in finite samples. Finally, Section 9 concludes with a discussion of directions for future research. To improve the exposition, Appendix A includes more examples, extensions and other methodological results, Appendix B reports omitted proofs, and a companion online supplemental appendix includes the complete set of simulation results. Finally, we also provide the general-purpose software package `ramchoice` for R, which implements our estimation and inference methods: see <https://cran.r-project.org/package=ramchoice> for details.

2 Setup

We designate a finite set X to act as the universal set of all mutually exclusive alternatives. This set is thus viewed as the grand alternative space, and is kept fixed throughout. A typical element of X is denoted by a and its cardinality is $|X| = K$. We let \mathcal{X} denote the set of all nonempty subsets of X . Each member of \mathcal{X} defines a choice problem.

Definition 1 (Choice Rule). A choice rule is a map $\pi : X \times \mathcal{X} \rightarrow [0, 1]$ such that for all $S \in \mathcal{X}$, $\pi(a|S) \geq 0$ for all $a \in S$, $\pi(a|S) = 0$ for all $a \notin S$, and $\sum_{a \in S} \pi(a|S) = 1$.

Thus, $\pi(a|S)$ is the probability that alternative a is chosen from the choice problem S . Our formulation allows both stochastic and deterministic choice rules. If $\pi(a|S)$ is either 0 or 1, then choices are deterministic. For simplicity in the exposition, we assume that all choice problems are potentially observable throughout the main paper, but this assumption is relaxed in Appendix A.2 to account for cases where only data on a subcollection of choice problems is available.

The key ingredient in our model is probabilistic consideration sets. Given a choice problem S , each non-empty subset of S could be a consideration set with certain probability. We impose that each frequency is between 0 and 1 and that the total frequency adds up to 1. Formally,

Definition 2 (Attention Rule). An attention rule is a map $\mu : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ such that for all $S \in \mathcal{X}$, $\mu(T|S) \geq 0$ for all $T \subset S$, $\mu(T|S) = 0$ for all $T \not\subset S$, and $\sum_{T \subset S} \mu(T|S) = 1$.

Thus, $\mu(T|S)$ represents the probability of paying attention to the consideration set $T \subset S$ when the choice problem is S . This formulation captures both deterministic and stochastic attention rules. For example, $\mu(S|S) = 1$ represents an agent with full attention. Given our approach, we can always extract the probability of paying attention to a specific alternative: For a given $a \in S$, $\sum_{a \in T \subset S} \mu(T|S)$ is the probability of paying attention to x in the choice problem S . The probabilities on consideration sets allow us derive the attention probabilities on alternatives uniquely.

3 Monotonic Attention

We consider a choice model where a decision maker picks the maximal alternative with respect to her preference among the alternatives she pays attention to. Our ultimate goal is to elicit her preference from observed choice behavior without requiring any information on consideration sets. Of course, this is impossible without any restrictions on her (possibly random) attention rule. For example, a decision maker’s choice can always be rationalized by assuming she only pays attention to singleton sets. Because the consumer never considers two alternatives together, one cannot infer her preference at all.

We propose a property (i.e., an identifying restriction) on how stochastic consideration sets change

as choice problems change, as opposed to explicitly modeling how the choice problem determines the consideration set. We argue below that this nonparametric property is indeed satisfied by many problems of interest and mimics heuristics people use in real life (see examples below and in Appendix A.1). This approach makes it possible to apply our method to elicit preference without relying on a particular formation mechanism of consideration sets.

Assumption 1. (Monotonic Attention) For any $a \in S - T$, $\mu(T|S) \leq \mu(T|S - a)$.

Monotonic μ captures the idea that each consideration set competes for consumers' attention: The probability of a particular consideration does not shrink when the number of possible consideration sets decreases. Removing an alternative that does not belong to the consideration set T results in less competition for T , hence the probability of T being the consideration set in the new choice problem is weakly higher. Our assumption is similar to the regularity condition proposed by [Suppes and Luce \(1965\)](#). The key difference is that their regularity condition is defined on choice probabilities, while our assumption is defined on attention probabilities.

To demonstrate the richness of the framework and motivate the analysis to follow, we discuss six leading examples of families of monotonic attention rules, that is, attention rules satisfying Assumption 1. We offer several more examples in the Appendix A.1. The first upcoming example is deterministic (i.e., $\mu(T|S)$ is either 0 or 1), but the others are all stochastic.

1. (ATTENTION FILTER; [Masatlioglu, Nakajima, and Ozbay, 2012](#)) A large class of deterministic attention rules, leading to consideration sets that do not change if an item not attracting attention is made unavailable (Attention Filter), was introduced by [Masatlioglu et al. \(2012\)](#). A classical example in this class is when a decision maker considers all the items appearing in the first page of search results and overlooks the rest. Formally, let $\Gamma(S)$ be the consideration set when the choice problem is S , and hence $\Gamma(S) \subset S$. Then, Γ is an Attention Filter if $a \notin \Gamma(S)$, then $\Gamma(S - a) = \Gamma(S)$. In our framework, this class corresponds to the case $\mu(T|S) = 1$ if $T = \Gamma(S)$, and 0 otherwise.
2. (RANDOM ATTENTION FILTERS) Consider a decision maker whose attention is deterministic but she utilizes different deterministic attention filters on different occasions. For example, it is well-known that search behavior on distinct platforms (mobile, tablet, and desktop) is

drastically different (e.g., the same search engine produces different first page lists depending on the platform, or different platforms utilize different search algorithms). In such cases, while the consideration set comes from a (deterministic) attention filter for each platform, the resulting consideration set is random. Formally, if a decision maker utilizes each attention filter Γ_j with probability ψ_j then the attention rule can be written as

$$\mu(T|S) = \sum_j \mathbb{1}(\Gamma_j(S) = T) \cdot \psi_j.$$

3. (INDEPENDENT CONSIDERATION; [Manzini and Mariotti, 2014](#)) Consider a decision maker who pays attention to each alternative a with a fixed probability $\gamma(a) \in (0, 1)$. γ represents the degree of brand awareness for a product, or the willingness of an agent to seriously evaluate a political candidate. Given that, the frequency of each set being the consideration set can be expressed as follow: for all $T \subset S$,

$$\mu(T|S) = \frac{1}{\beta_S} \prod_{a \in T} \gamma(a) \prod_{a \in S-T} (1 - \gamma(a)),$$

where $\beta_S = 1 - \prod_{a \in S} (1 - \gamma(a))$, which represents the probability that the decision maker considers no alternative in S , is used to adjust each probability so that they sum up to 1.

4. (LOGIT ATTENTION; [Brady and Rehbeck, 2016](#)) Consider a decision maker who assigns a positive weight for each non-empty subset of X . Psychologically w_T is a strength associated with the subset T . The probability of considering T in S can be written as follows:

$$\mu(T|S) = \frac{w_T}{\sum_{T' \subset S} w_{T'}}.$$

Even though there is no structure on weights in the general version of this model, there are two interesting special cases where weights solely depend on the size of the set. These are $w_T = |T|$ and $w_T = \frac{1}{|T|}$, which are conceptually completely different. In the latter, the decision maker tends to have smaller consideration sets while the larger consideration sets are more likely in the former.

5. (DOGIT ATTENTION) This example is a generalization of Logit Attention, and is based on the idea of the Dogit model (Gaundry and Dagenais, 1979). A decision maker is captive to a particular consideration set with certain probability, to the extent that she pays attention to that consideration set regardless of the weights of other possible consideration sets. Formally, let

$$\mu(T|S) = \frac{1}{1 + \sum_{T' \subset S} \theta_{T'}} \frac{w_T}{\sum_{T' \subset S} w_{T'}} + \frac{\theta_T}{1 + \sum_{T' \subset S} \theta_{T'}},$$

where $\theta_T \geq 0$ represents the degree of captivity (impulsivity) of T . “Captivity parameter” reflects the attachment of a decision maker to a certain consideration set. Since w_T are non-negative, the second term, which independent of w_T , is the smallest lower bound for $\mu(T|S)$. The larger θ_T is, the more likely the decision maker is captive to T and pays attention to it. When $\theta_T = 0$ for all T , this model becomes Logit Attention. This formulation is able to distinguish between impulsive and deliberate attention behavior.

6. (ELIMINATION BY ASPECTS) Consider a decision maker who intentionally or unintentionally focuses on a certain “aspect” of alternatives, and then refuses or ignores those alternatives that do not possess that aspect. This model is similar in spirit to Tversky (1972). Let $\{j, k, \ell, \dots\}$ be the set of aspects. Let ω_j represent the probability that aspect j “draws attention to itself.” It reflects the salience and/or importance of aspect j . All alternatives without that aspect do not receive attention. Let B_j be the set of alternatives that possess aspect j . We assume that each alternative must belong to at least one B_j with $\omega_j > 0$. If aspect j is the salient aspect, the consideration set is $B_j \cap S$ when S is the set of feasible alternatives. The total probability of T being the consideration set is the sum of ω_j such that $T = B_j \cap S$. When there is no alternative in S possessing the salient aspect, a new aspect will be drawn. Formally, the probability of T being consideration set under S is given by

$$\mu(T|S) = \sum_{B_j \cap S = T} \frac{\omega_j}{\sum_{B_k \cap S \neq \emptyset} \omega_k}.$$

These six examples give a sample of different limited attention models of interest in economics, psychology, marketing, and many other disciplines. While these examples are quite distinct from each other, all of them are monotonic attention rules. In other words, they illustrate that Assump-

tion 1 is very general and potentially useful. As a consequence, our revealed preference characterization will be applicable to a wide range of choice rules without committing to a particular attention mechanism, which is not observable in practice and hence untestable. Furthermore, as illustrated by the examples above (and those in Appendix A.1), our upcoming characterization, identification, estimation and inference results nest important previous contributions in the literature as special cases.

4 A Random Attention Model

We are ready to introduce our random attention model based on Assumption 1. We assume the decision maker has a strict preference ordering \succ on X . To be precise, we assume the preference ordering is an *asymmetric*, *transitive* and *complete* binary relation: A binary relation \succ on a set X is (i) asymmetric, if for all $x, y \in X$, $x \succ y$ implies $y \not\succeq x$; (ii) transitive, if for all $x, y, z \in X$, $x \succ y$ and $y \succ z$ imply $x \succ z$; and (iii) complete, if for all $x \neq y \in X$, either $x \succ y$ or $y \succ x$ is true. Consequently, the decision maker always picks the maximal alternative with respect to her preference among the alternatives she pays attention to. Formally,

Definition 3. A choice rule π has a *random attention representation* if there exists a preference ordering \succ over X and a monotonic attention rule μ (Assumption 1) such that

$$\pi(a|S) = \sum_{T \subset S} \mathbb{1}(a \text{ is } \succ\text{-best in } T) \cdot \mu(T|S)$$

for all $a \in S$ and $S \in \mathcal{X}$. In this case, we say π is represented by (\succ, μ) . We may also say that \succ represents π , which means that there exists some monotonic attention rule μ such that (\succ, μ) represents π . We also say π is a *Random Attention Model* (RAM).

While our framework is designed to model stochastic choices, our model captures deterministic choices as well. In classical choice theory, a decision maker chooses the best alternative according to her preferences with probability 1, hence choice is deterministic. In our framework, this case is captured by a monotone attention rule with $\mu(S|S) = 1$. Figure 1 gives a graphical representation of RAM.

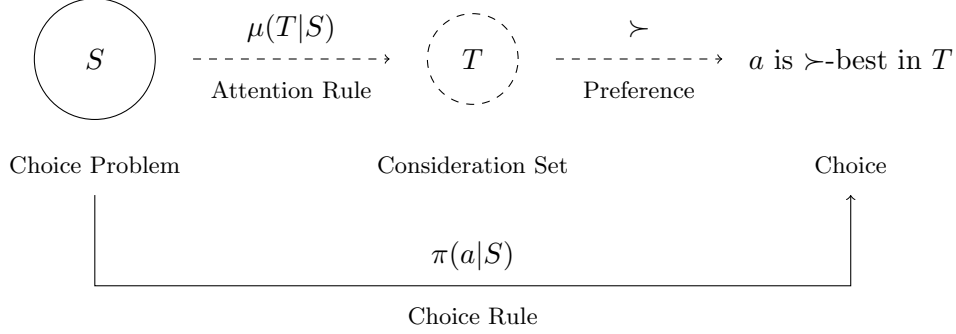


Figure 1. Illustration of a RAM model. *Observable*: choice problem and choice (solid line). *Unobservable*: attention rule, consideration set and preference (dashed line).

We now derive some implications for our random attention model. They can be used to test the model in terms of observed choice rules/probabilities. In the literature, there is a principle called *regularity* (see [Suppes and Luce, 1965](#)), according to which adding an new alternative should only decrease the probability of choosing one of the existing alternatives. However, empirical findings suggest otherwise; see [Rieskamp, Busemeyer, and Mellers \(2006\)](#) for a review of empirical evidence on violations of regularity and alternative theories explaining these violations. Importantly, our model allows regularity violations.

Remark 1 (Regularity Violation). In RAM, removing an alternative from the feasible set can decrease the likelihood that an existing alternative is selected. That is, it is possible to have $\pi(a|S) > \pi(a|S - b)$. ┘

The next example illustrates that adding an alternative to the feasible set can increase the likelihood that an existing alternative is selected. This never happens either in Luce model or in any Random Utility Model (RUM). In RAM, the addition of an alternative changes the choice set and therefore decision maker’s attention, which could increase the probability of an existing alternative being chosen.

Example 1 (RUM Violation). Let $X = \{a, b, c\}$ with $a \succ b \succ c$ and consider the following

monotonic attention rule μ . Each row corresponds to a different choice problem.

$\mu(T S)$	$T = \{a, b, c\}$	$\{a, b\}$	$\{a, c\}$	$\{b, c\}$	$\{a\}$	$\{b\}$	$\{c\}$
$S = \{a, b, c\}$	2/3	0	0	1/6	0	0	1/6
$\{a, b\}$		1/2			0	1/2	
$\{a, c\}$			1/2		0		1/2
$\{b, c\}$				1/2		0	1/2

Then $\pi(a|\{a, b, c\}) = 2/3 > 1/2 = \pi(a|\{a, b\}) = \pi(a|\{a, c\})$. \perp

This example shows that RAM can explain choice patterns which cannot be explained by the classical RUM. Given that the model allows regularity violations, one might think that the model has very limited empirical implications (i.e., it is “too” general to have empirical content). However, the next example illustrates that it is easy to find a choice rule π which lies outside RAM with only three alternatives.

Example 2 (RAM Violation). The following choice rule π is not compatible with our random attention model. Each column corresponds to a different choice problem.

$\pi(\cdot S)$	$S = \{a, b, c\}$	$\{a, b\}$	$\{a, c\}$	$\{b, c\}$
a	1/3	1	0	
b	1/3	0		1
c	1/3		1	0

The choice behavior is symmetric among all the alternatives. We now illustrate that π is not a RAM. Since the choice is symmetric, without loss of generality, assume $a \succ b \succ c$. Then $\mu(\{c\}|\{a, b, c\})$ must be 1/3. Assumption 1 implies that $\mu(\{c\}|\{b, c\})$ must be greater than 1/3. This yields a contradiction since $\pi(c|\{b, c\}) = 0$. \perp

Given that RAM has non-trivial empirical content, it is natural to investigate to what extent Assumption 1 can be used to elicit (unobserved) strict preference orderings given (observed) choices of decision makers.

4.1 Revealed Preference

In general, a choice rule can have multiple RAM representations with different preference orderings and different attention rules. When multiple representations are possible, we say that a is revealed to be preferred to b if and only if a is preferred to b in all possible RAM representations. This is a very conservative approach as it makes sure we never make false claims about the preference of the decision maker. The same approach is also used in [Masatlioglu, Nakajima, and Ozbay \(2012\)](#).

Definition 4 (Revealed Preference). Let (\succ_j, μ_j) be all random attention representations of π . Then we say a is *revealed to be preferred to* b if $a \succ_j b$ for all j .

We now show how revealed preference theory can still be developed successfully in our RAM framework. If there is only one representation (\succ, μ) , then the revealed preference will be equal to \succ . If one wants to know whether a is revealed to be preferred to b , it would appear necessary to check whether every (\succ_j, μ_j) represents her choice. However, this is not practical, especially when there are many alternatives. Instead, we shall now provide a handy method to obtain the revealed preference completely. We first provide an observation about no revelation.

Lemma 1. Suppose for all $a, b \in S$ and $S \in \mathcal{X}$, $\pi(a|S) \leq \pi(a|S - b)$. Then π has a random attention representation. Moreover, there is no preference revelation.

Proof. Given any $S \in \mathcal{X}$, let $\mu(\{a\}|S) = \pi(a|S)$ for all $a \in S$ and $\mu(T|S) = 0$ for all T with $|T| > 1$. The hypothesis in Lemma 1 guarantees that μ is monotonic. Pick any arbitrary preference relation \succ on X . It is obvious that (\succ, μ) represents π . Since \succ is arbitrary, there is no preference revelation in the sense of Definition 4. ■

One implication of Lemma 1 is that RAM is strictly more general than RUM. That's because any RUM satisfies regularity and by Lemma 1 any choice rule which satisfies regularity has a RAM representation. Moreover, Lemma 1 suggests that if there is no regularity violation in choice behavior, there is no preference revelation. The next lemma shows the “opposite”: Whenever there is a regularity violation, we can learn about underlying preference.

Lemma 2. Let π be a RAM. If $\pi(a|S) > \pi(a|S - b)$, then a is revealed to be preferred to b .

Proof. Pick any arbitrary random attention representation (\succ, μ) of π , and assume $b \succ a$. We will show this leads to a contradiction. Note that if $b \in T \subset S$, then a cannot be \succ -best in T . Now

$$\begin{aligned}
\pi(a|S) &= \sum_{T \subset S} \mathbf{1}(a \text{ is } \succ\text{-best in } T) \cdot \mu(T|S) \\
&= \sum_{T \subset S-b} \mathbf{1}(a \text{ is } \succ\text{-best in } T) \cdot \mu(T|S) \\
&\leq \sum_{T \subset S-b} \mathbf{1}(a \text{ is } \succ\text{-best in } T) \cdot \mu(T|S-b) \\
&= \pi(a|S-b),
\end{aligned}$$

where the first and the last equality follow from the fact that (\succ, μ) represents π , the second equality follows from the assumption that $b \succ a$, and the inequality follows from the fact that μ is monotonic. Since we have a contradiction to our hypothesis that $\pi(a|S) > \pi(a|S-b)$, $b \succ a$ is not possible. Hence, in all RAM representations of π , $a \succ b$ must hold. ■

Lemma 2 allows us to define the following binary relation. For any distinct a and b , define:

$$aPb, \text{ if there exists } S \in \mathcal{X} \text{ including } a \text{ and } b \text{ such that } \pi(a|S) > \pi(a|S-b).$$

By Lemma 2, if aPb then a is revealed to be preferred to b . In other words, this condition is sufficient to reveal preference. In addition, since the underlying preference is transitive, we also conclude that she prefers a to c if aPb and bPc for some b , even when aPc is not directly revealed from her choices. Therefore, the transitive closure of P , denoted by P_R , must also be part of her revealed preference. One may wonder whether some revealed preference is overlooked by P_R . The following theorem, which is our first main result, shows that P_R includes all preference information, given the observed choice probabilities and under Assumption 1.

Theorem 1 (Revealed Preference). Let π be a RAM. Then a is revealed to be preferred to b if and only if $aP_R b$.

Proof. The “if” part of the Theorem follows from Lemma 2. To prove the “only if” part, we show that given any preference \succ that includes P_R , there exists a monotonic attention rule μ such that (\succ, μ) represents π . The details of the construction can be found in the proof of Theorem 2 given below. ■

Our revealed preference result includes the one given in [Masatlioglu, Nakajima, and Ozbay \(2012\)](#) for non-random attention filters. In their model a is revealed to be preferred to b if there is a choice problem such that a is chosen and b is available, but it is no longer chosen when b is removed from the choice problem. This means we have $1 = \pi(a|S) > \pi(a|S - b) = 0$. Given [Theorem 1](#), this reveals that a is better than b . On the other hand, generalizing to non-deterministic attention rules allows for a broader class of empirical and theoretical settings to be analyzed, hence our revealed preference result ([Theorem 1](#)) is strictly richer than those obtained in previous work. For example, in a deterministic world with three alternatives, there is no data revealing the entire preference. On the other hand, we illustrate that it is possible to reveal the entire preference in RAM with only three alternatives. This discussion makes the connection between deterministic and probabilistic choice in terms of revealed preference clear.

Example 3 (Full Revelation). Consider the following stochastic choice with three alternatives:

$\pi(\cdot S)$	$S = \{a, b, c\}$	$\{a, b\}$	$\{a, c\}$	$\{b, c\}$
a	1/2	1	0	
b	1/2	0		1
c	0		1	0

We first verify that π has a random attention representation. To this end, pick a preference \succ with $a \succ b \succ c$ and consider the following attention rule: $\mu(\{a, b, c\}|\{a, b, c\}) = \mu(\{b, c\}|\{a, b, c\}) = 1/2$, $\mu(\{a, b\}|\{a, b\}) = \mu(\{a, c\}|\{a, c\}) = \mu(\{b, c\}|\{b, c\}) = 1$, and $\mu(T|S) = 0$ for all other T and S . It is clear that μ is monotonic and (\succ, μ) represents π , hence the choice rule is a RAM.

Now we show that in all possible representations of π , $a \succ b \succ c$ must hold. By [Lemma 2](#), $\pi(a|\{a, b, c\}) > \pi(a|\{a, c\})$ implies that a is revealed to be preferred to b . Similarly, $\pi(b|\{a, b, c\}) > \pi(b|\{a, b\})$ implies b is revealed to be preferred to c . Hence there is unique revealed preference $a \succ b \succ c$. ┘

4.2 A Characterization

[Theorem 1](#) characterizes revealed preference. However, it is not applicable unless the observed choice behavior has a random attention representation, which motivates the following question:

How can we test whether a choice rule is consistent with RAM? It turns out that RAM can be simply characterized by only one behavioral postulate of choice.

Our characterization is based on a similar idea due to [Houthakker \(1950\)](#). Choices reveal information about preferences. If these revelations are consistent in the sense that there is no cyclical preference revelations, the choice behavior has a RAM representation.

Theorem 2 (Characterization). A choice rule π has a random attention representation if and only if \mathbf{P}_R has no cycle.

The idea of the proof is as follows. One direction of the statement follows directly from [Lemma 2](#). For the other direction, we need to construct a preference and a monotonic attention rule representing the choice rule. Given that \mathbf{P}_R has no cycle, there exists a preference relation \succ including \mathbf{P}_R . Indeed, we illustrate that any such completion of \mathbf{P}_R represents π by an appropriately chosen μ . Therefore, the construction of μ depends on a particular completion of \mathbf{P}_R . We then illustrate that the constructed μ satisfies [Assumption 1](#). At the last step, we show that (\succ, μ) represents π .

Proof. Suppose π has a random attention representation (\succ, μ) . Then [Lemma 2](#) implies that \succ must include \mathbf{P}_R so \mathbf{P}_R must be acyclic.

Now suppose that \mathbf{P}_R has no cycle. Pick any preference \succ that includes \mathbf{P}_R and enumerate all alternatives with respect to \succ : $a_{1,\succ} \succ a_{2,\succ} \succ \dots \succ a_{K,\succ}$. To specify μ , for each k , we define $L_{k,\succ} = \{a_{k,\succ}, \dots, a_{K,\succ}\}$. We define μ as follows.

$$\mu(T|S) = \begin{cases} \pi(a_{k,\succ}|S) & \text{if } \exists k \text{ s.t. } T = L_{k,\succ} \cap S \\ 0 & \text{otherwise} \end{cases}$$

There may be multiple subsets of S where $a_{k,\succ}$ is the unique \succ -best element. However, our definition of μ guarantees that among those, $L_{k,\succ} \cap S$ is the only consideration set with non-zero frequency. Since $\mu(L_{k,\succ} \cap S|S) = \pi(a_{k,\succ}|S)$, the representation holds trivially. So all we need to show is that μ is monotonic.

Suppose there exists $a \in S$ and $T \subset S - a$ such that $\mu(T|S) > \mu(T|S - a)$. Since $\mu(T|S) > 0$, there exists $a_{k,\succ}$ such that T is equal to $L_{k,\succ} \cap S$ and $a_{k,\succ} \in S$. Notice that $a \notin T$, and hence

$T = L_{k,\succ} \cap (S - a)$. By definition, $\mu(T|S - a) = \pi(a_{k,\succ}|S - a)$ and $\mu(T|S) = \pi(a_{k,\succ}|S)$. Since $\mu(T|S) > \mu(T|S - a)$, we have $\pi(a_{k,\succ}|S) > \pi(a_{k,\succ}|S - a)$, which by definition implies that $a_{k,\succ} \mathbf{P}_R a$. But then since \succ includes \mathbf{P}_R , we have $a \in L_{k,\succ} \cap S = T$ which is a contradiction. ■

Recall that Example 2 is outside of our model. Theorem 2 implies that \mathbf{P}_R must have a cycle. Indeed, we have $a \mathbf{P} b$ due to the regularity violation $\pi(a|\{a, b, c\}) > \pi(a|\{a, c\})$. Similarly, we have $b \mathbf{P} c$ by $\pi(b|\{a, b, c\}) > \pi(b|\{a, b\})$ and $c \mathbf{P} a$ by $(\pi(c|\{a, b, c\}) > \pi(c|\{b, c\}))$. Since \mathbf{P}_R must have a cycle, Example 2 must be outside of our model. Therefore, Theorem 2 provides a very simple test for RAM.

5 Random Attention Filter

We focus on random attention filters, which are one of the motivating examples of monotonic attention rules. A random attention filter is a linear combination of deterministic attention filters. For example, the same individual might be utilizing different platforms during her search. Each platform yields a different deterministic attention filter, and the usage frequency of each platform is equal to the weight of that attention filter. Random attention filters also give a different interpretation of our model.

The set of all random attention filters is a strict subset of monotonic attention rules. This is not surprising given that the class of monotonic attention rules is very large. What is (arguably) surprising is the following fact that we are able to show: If (π, \succ, μ) is a RAM with μ being a monotonic attention rule, then there exists a random attention filter μ' such that (π, \succ, μ') is still a RAM. This turns out to be very useful for econometric identification, estimation and inference. Before presenting this result, however, observe that μ and μ' do not need to be the same, which means there exists monotonic attention rules that cannot be written as a linear combination of deterministic attention filters.

Example 4. Let $X = \{a_1, a_2, a_3, a_4\}$. Consider a monotonic attention rule μ such that (i) $\mu(T|S)$ is either 0 or 0.5, (ii) $\mu(T|S) = 0$ if $|T| > 1$, and (iii) if $\mu(\{a_j\}|S) = 0$ and $k < j$ then $\mu(\{a_k\}|S) = 0$. Then we must have $\mu(\{a_3\}|\{a_1, a_2, a_3, a_4\}) = \mu(\{a_4\}|\{a_1, a_2, a_3, a_4\}) = 0.5$. We now show that μ is not a random attention filter.

Suppose μ can be written a linear combination of attention filters. Then $\mu(\{a_3\}|\{a_1, a_2, a_3, a_4\}) = \mu(\{a_4\}|\{a_1, a_2, a_3, a_4\}) = 0.5$ implies that only attention filters for which $\Gamma(\{a_1, a_2, a_3, a_4\}) = \{a_3\}$ or $\Gamma(\{a_1, a_2, a_3, a_4\}) = \{a_4\}$ must be assigned positive probability. On the other hand, $\mu(\{a_2\}|\{a_1, a_2, a_3\}) = 0.5 = \mu(\{a_2\}|\{a_1, a_2, a_4\}) = 0.5$ imply that for all Γ which are assigned positive probability $\Gamma(\{a_1, a_2, a_3\}) = \{a_2\}$ whenever $\Gamma(\{a_1, a_2, a_3, a_4\}) = \{a_4\}$ and $\Gamma(\{a_1, a_2, a_4\}) = \{a_2\}$ whenever $\Gamma(\{a_1, a_2, a_3, a_4\}) = \{a_3\}$. To see this, notice that attention filter property implies $\Gamma(\{a_1, a_2, a_3\}) = \{a_3\}$ for all Γ with $\Gamma(\{a_1, a_2, a_3, a_4\}) = \{a_3\}$ and $\Gamma(\{a_1, a_2, a_4\}) = \{a_4\}$ for all Γ with $\Gamma(\{a_1, a_2, a_3, a_4\}) = \{a_4\}$. But then it must be $\Gamma(\{a_1, a_2\}) = \{a_2\}$ for all Γ which are assigned positive probability or that $\mu(\{a_2\}|\{a_1, a_2\}) = 1$, a contradiction. \perp

We now show that if we restrict our attention to a certain type of monotonic attention rules, then we can show that within that class every monotonic attention rule is a random attention filter. Before showing the result we define this subclass of triangular attention rules and justify why focusing on this subclass is sufficient and without loss of generality. First, we need additional notation. Given a strict preference ordering \succ over X , reorder the alternatives so that $a_{1,\succ} \succ \dots \succ a_{K,\succ}$. Then, we define $L_{k,\succ}$ to be the lower counter set of $a_{k,\succ}$ including itself, namely the set $\{a_{k,\succ}, \dots, a_{K,\succ}\}$.

Definition 5 (Triangular Attention Rule). A attention rule μ is triangular with respect to \succ if for any $S \in \mathcal{X}$ and $T \subset S$, $\mu(T|S) > 0$ only if $T = S \cap L_{k,\succ}$ for some $k \in \{1, \dots, K\}$.

Remark 2. In the proof of Theorem 2, the construction of the monotonic attention rule is based on triangular attention rules. This implies that a choice rule which can be represented by a monotonic attention rule can also be represented by a monotonic triangular attention rule. Formally, if π has a random attention representation, (\succ, μ) , then (\succ, μ') also represents π where μ' is monotonic and triangular with respect to \succ . Hence, it is without loss of generality to focus on monotonic triangular attention rules. \perp

Let $\mathcal{MT}(\succ)$ denote the set of all attention rules which are both monotonic and triangular with respect to \succ , and $\mathcal{AF}(\succ)$ denote all deterministic attention filters which are triangular with respect to \succ . We are now ready to state the main result of this section.

Theorem 3 (Random Attention Filter). For any $\mu \in \mathcal{MT}(\succ)$, there exists a probability law ψ on $\mathcal{AF}(\succ)$ such that for any $S \in \mathcal{X}$ and $T \subset S$

$$\mu(T|S) = \sum_{\Gamma \in \mathcal{AF}(\succ)} \mathbf{1}(\Gamma(S) = T) \cdot \psi(\Gamma).$$

The proof of this results is long and hence left to Appendix B, but here we provide a sketch of it. First, $\mathcal{MT}(\succ)$ is a compact and convex set, and thus the above theorem can alternatively be stated as follows: The set of extreme points of $\mathcal{MT}(\succ)$ is $\mathcal{AF}(\succ)$. (An attention rule $\mu \in \mathcal{MT}(\succ)$ is an extreme point of $\mathcal{MT}(\succ)$ if it cannot be written as a nondegenerate convex combination of any $\mu', \mu'' \in \mathcal{MT}(\succ)$.) Then, Minkowski's Theorem guarantees that every element of $\mathcal{MT}(\succ)$ lies in the convex hull of $\mathcal{AF}(\succ)$.

Obviously, every element of $\mathcal{AF}(\succ)$ is an extreme point of $\mathcal{MT}(\succ)$. We then show that non-deterministic triangular attention rules cannot be extreme points, i.e. given any $\mu \in \mathcal{MT}(\succ) - \mathcal{AF}(\succ)$ we can construct $\mu', \mu'' \in \mathcal{MT}(\succ)$ such that $\mu = \frac{1}{2}\mu' + \frac{1}{2}\mu''$. The key step is to show that both μ' and μ'' that we construct are monotonic. After this step, we have shown that no $\mu \in \mathcal{MT}(\succ) - \mathcal{AF}(\succ)$ can be an extreme point, thus concluding the proof.

Remark 3. If π has a random attention representation, (\succ, μ) then (\succ, μ') also represents π where μ' is a random attention filter. Even though monotonic attention rules are much larger compared to random attention filters, there is no additional benefit in terms of explanatory power. \lrcorner

Theorem 3 shows that π has a random attention representation if and only if π has a random attention filter representation. In Section 7, we leverage this result to develop econometric methods for identification, estimation and inference of the strict preference ordering \succ on X given observed (or estimable) choice probabilities, as well as for specification testing. Before doing so we discuss more precisely our results in light of the most closely related literature on (random) limited attention and choice theory.

6 Related Literature

Manzini and Mariotti (2014) and Brady and Rehbeck (2016) are the two closest related papers to our work. Both provide parametric random attention models, which are described in the third and

fourth examples in Section 3, respectively. Since their attention rules are monotonic, these models are two interesting special cases of our random attention model. To provide an accurate comparison, we need to introduce an outside/default option, which is required by both models. Thus, we first extend RAM to accommodate for an outside option and then offer a detailed comparison between our work and both of these papers.

Let $a^* \notin X$ be the default option. In the model with the default option, we will allow an empty consideration set. Hence, now $\mu(\cdot|S)$ is defined over all subsets of S including the empty set. The default option is always available and can be interpreted as choosing nothing whenever the consideration set is empty. Let $X^* = X \cup \{a^*\}$ and $S^* = S \cup \{a^*\}$ for all $S \in \mathcal{X}$. We require that the choice rule satisfies $\sum_{a \in S^*} \pi(a|S) = 1$ and $\pi(a|S) \geq 0$ for all $a \in S^*$. We say that a choice rule π has a random attention representation with a default option if there exist a preference ordering \succ on X and a monotonic attention rule μ such that for each $a \in S$, $\pi(a|S) = \sum_{T \subset S} \mathbb{1}(a \text{ is } \succ\text{-best in } T) \cdot \mu(T|S)$. Thus $\pi(a^*|S) = \mu(\emptyset|S)$.

An implication of Assumption 1 is that for all S , $\mu(\emptyset|S) \leq \mu(\emptyset|S - a)$. In terms of choice probabilities, this implies that default option satisfies regularity. In fact, it is easy to see that regularity on default and acyclicity of P are necessary and sufficient for π to have a RAM representation with a default option.

Remark 4. A choice rule π has a RAM representation with a default option if and only if it satisfies acyclicity of P and regularity on default. \square

In [Manzini and Mariotti \(2014\)](#), a choice rule has the representation

$$\pi(a|S) = \gamma(a) \prod_{b \in S: b \succ a} (1 - \gamma(b))$$

where $\gamma(a)$ is the fixed consideration probability of alternative a . See [Horan \(2013\)](#) for an axiomatic characterization of this model when there is no default option. It follows that this model is a special case of RAM with a default option if we set

$$\mu(T|S) = \prod_{a \in T} \gamma(a) \prod_{b \in S-T} (1 - \gamma(b)).$$

In [Brady and Rehbeck \(2016\)](#), a choice rule has the representation

$$\pi(a|S) = \frac{\sum_{T \subset S} \mathbf{1}(a \in T \text{ is } \succ\text{-maximal in } T) \cdot w(T)}{\sum_{T \subset S} w(T)}.$$

where $w(T)$ is the weight of consideration set T . This model is also a special case of RAM with an outside option if we set

$$\mu(T|S) = \frac{w(T)}{\sum_{T' \subset S} w(T')}.$$

Another closely related paper on random consideration sets is [Aguiar \(2015\)](#). In fact, acyclicity of P appears as an axiom in his representation. Hence, his model is also a special case of RAM.

We also compare our model with other random choice models even though they are not in the framework of random consideration sets. [Gul, Natenzon, and Pesendorfer \(2014\)](#) consider the following model. There is a collection of attributes, denoted by Z . Each attribute $z \in Z$ has a weight $v(z)$ and each alternative $a \in X$ has intensity $i(a, z)$ in attribute z . The decision maker first randomly picks an attribute using Luce rule given the weights of all attributes. Then the decision maker picks an alternative using Luce rule given the intensities of all alternatives in attribute z . If an alternative has a high intensity in more valuable attributes, then it is more likely to be chosen. This model reduces to Luce rule if no pair of alternatives share common attributes. [Gul, Natenzon, and Pesendorfer \(2014\)](#) show that any attribute rule is a random utility model. Moreover, any random utility model can be approximated by an attribute rule. Since RUM is a subset of RAM, any choice behavior that can be explained by an attribute rule can also be explained by RAM.

Another recent paper on stochastic choices is [Echenique, Saito, and Tserenjigmid \(2014\)](#). Their model (PALM) attributes all violations of Luce's IIA to perception and uses that to reveal perception priority of alternatives. It is easy to construct examples which can be explained by RAM but not PALM. For example, consider any stochastic choice data in which the outside option is never chosen. When the outside option is never chosen, PALM is exactly Luce rule, and hence Luce's IIA axiom holds. However, RAM can accommodate violations of Luce's IIA even in the absence of an outside option. Since PALM is exactly Luce rule when there is no outside option, to be able to make comparisons we consider RAM with an outside option. Recall that RAM with an outside option satisfies regularity on default. However, PALM does not necessarily satisfy this property.

Hence, RAM with an outside option does not nest PALM.

Echenique and Saito (2017) consider a general Luce model (GLM) in which the decision maker uses Luce rule to choose among alternatives in her consideration set instead of the whole choice set. Ahumada and Ulku (2017) also provide a characterization of GLM. In contrast to RAM, consideration sets are deterministic in GLM. Moreover, RAM assumes ordinal utility while in GLM utility has a cardinal meaning. Despite these differences, both models can be used to explain the same choice behavior. For example, Luce model is a subset of both RAM and GLM. We now show that the models are distinct. Recall Example 2 which cannot be explained by RAM. This example can be explained by GLM by assuming equal utility weights for all alternatives and that whenever an alternative is not chosen that is because the alternative does not belong to the consideration set. Now consider a choice rule in which all alternatives are always chosen with positive probability but Luce’s IIA is not satisfied. We can construct such an example in which acyclicity of P holds and hence the choice rule has a RAM representation. However, this types of choice behavior cannot be explained by GLM as GLM reduces to Luce rule when all alternatives are always chosen with positive probability. Echenique and Saito (2017) also consider two special cases of GLM: (i) In two-stage Luce model, consideration sets are induced by an asymmetric, transitive binary relation, (ii) In threshold Luce model, an alternative belongs to the consideration set only if its utility is not too small compared to the utility of the maximal element in the choice set. Threshold Luce model is a subset of two-stage Luce model. Even though RAM and two-stage Luce model are distinct, threshold Luce model is a subset of RAM. To see that two-stage Luce model and RAM are distinct first notice that as a subcase of GLM, two-stage Luce model cannot nest RAM. On the other hand, consider a choice rule with $\pi(a|\{a, b, c, d\}) > \pi(a|\{a, c, d\})$ and $\pi(b|\{a, b, c, d\}) > \pi(b|\{b, c, d\})$. This choice rule violates acyclicity of P , and hence a RAM representation does not exist. However, these observations are compatible with two-stage Luce model. Finally, to see that RAM nests threshold Luce model, notice that in threshold Luce model $\pi(a|S) > \pi(a|S - b)$ implies that $u(b) > u(a)$. Hence, any choice rule consistent with threshold Luce model must satisfy acyclicity of P .

Fudenberg, Iijima, and Strzalecki (2015) consider a model (Additive Perturbed Utility-APU) in which agents deliberately randomize as making deterministic choices can be costly. In their model, choices always satisfy regularity. Since any choice rule that satisfies regularity has a RAM

representation (Lemma 1), RAM includes APU.

Aguiar, Boccardi, and Dean (2016) consider a satisficing model in which the decision maker searches till she finds an alternative above the satisficing utility level. If there is no alternative above satisficing utility level, the decision maker picks the best available alternative. They show that the general model they consider has no behavioral content. They focus on two special cases: (i) Full Support Satisficing Model (FSSM) in which in any menu each alternative has a positive probability of being searched first, and (ii) Fixed Distribution Satisficing Model (FDSM). They show that FDSM is a subset of RUM and hence it is also a subset of RAM. There exist choice rules that can be explained by FSSM but not RAM. In fact, FSSM has no restrictions if all alternatives are always chosen with positive probability. On the other hand, suppose $\pi(a|\{a, b, c\}) = \pi(a|\{a, b\}) = \pi(a|\{a, c\}) = 1$ and $\pi(b|\{b, c\}) = 1/2$. FSSM cannot explain this choice behavior even though regularity is satisfied. Hence, FSSM and RAM are distinct and FDSM is a subset of both.

7 Econometric Methods

Theorem 1 shows that if the choice probability π is a RAM then preference revelation is possible. Theorem 2 gives a falsification result, based on which a specification test can be designed. Theorem 3 gives a generic representation result of monotonic attention rules as random triangular attention filters. This latter result turns out to be useful for our proposed identification, estimation and inference methods, which are developed in this section.

7.1 Nonparametric Identification

We first define the set of partially identified preferences, denoted by Θ_π . This mirrors Definition 3, with the only difference that now we assume the choice rule is observed/identifiable from data.

Definition 6 (Compatible Preferences). Let π be the underlying choice rule/data generating process. A preference \succ is compatible with π , denoted by $\succ \in \Theta_\pi$, if there exists some monotonic attention rule μ such that (π, \succ, μ) is a RAM.

When π is known, it is possible to employ Theorem 1 directly to construct Θ_π . For example, consider the specific preference ordering $a \succ b$, which can be checked by the following procedure.

First, check whether $\pi(b|S) \leq \pi(b|S - a)$ is violated for some S . If so, then we know the preference ordering is not compatible with RAM and hence does not belong to Θ_π (Lemma 2). On the other hand, if the preference ordering is not rejected in the first step, we need to check along “longer chains” (Theorem 1). That is, whether $\pi(b|S) \leq \pi(b|S - d)$ and $\pi(d|T) \leq \pi(d|T - a)$ are simultaneously violated for some S, T and d . If so, the preference ordering is rejected (i.e., incompatible with RAM), while if not then a chain of length three needs to be considered. This process goes on for longer chains until either at some step we are able to reject the preference ordering, or all possibilities are exhausted. This algorithm, albeit feasible, can be hard to implement in practice, even when the choice probabilities are known. The fact that π has to be estimated makes the problem even more complicated, since it becomes a sequential multiple hypothesis testing problem.

Theorem 2 shows that π is RAM if and only if no cycles can be deduced when applying Theorem 1. Equivalently, π is RAM if and only if Θ_π is nonempty. This provides one way to falsify the model. Conceptually, existence of cycles can be tested by employing Theorem 1 and checking whether Θ_π is empty, but such methodology suffers from the same problem mentioned earlier: It is too complicated to implement, and using estimated choice probabilities requires adjustments for sequential and multiple hypotheses testing.

One of the main purposes of this section is to provide an equivalent form of identification, which (i) is simple to implement, and (ii) remains statistically valid even when applied using estimated choice rules. For the ease of exposition, we rewrite the choice rule π as a long vector $\boldsymbol{\pi}$, whose elements are simply the probability of each alternative $a \in X$ being chosen from a choice problem $S \in \mathcal{X}$. The order of elements in $\boldsymbol{\pi}$ does not matter, and rearrangements simply correspond to column permutations of \mathbf{R}_\succ in the following theorem.

Theorem 4 (Inequality Constraints). Given any preference \succ , there exists a unique matrix \mathbf{R}_\succ such that $\succ \in \Theta_\pi$ if and only if $\mathbf{R}_\succ \boldsymbol{\pi} \leq \mathbf{0}$.

This theorem states that in order to decide whether a preference \succ is compatible with the (identifiable) choice rule π , it suffices to check a collection of inequality constraints. In particular, it is no longer necessary to consider the sequential and multiple testing problems mentioned earlier. Moreover, given the large econometric literature on moment inequality testing, many techniques

can be adapted when Theorem 4 is applied to estimated choice rules, $\hat{\pi}$, as we discuss further below.

Proof. Recall that (π, \succ) has a RAM representation if and only if there exists a monotonic attention rule $\tilde{\mu}$ such that π is induced by $\tilde{\mu}$ and \succ . As discussed earlier, it is without loss of generality to focus on triangular attention rules, which implies $\succ \in \Theta_\pi$ if and only if there exists a monotonic triangular attention rule μ which induces π . (See the discussion before Theorem 3.) The constraint matrix \mathbf{R}_\succ is constructed to take the product form: $\mathbf{R}\mathbf{C}_\succ$, where the first matrix, \mathbf{R} , consists of constraints on the attention rules, and the second matrix, \mathbf{C}_\succ , maps the choice rule back to a triangular attention rule. First consider \mathbf{R} . The only restrictions imposed on attention rules are from the monotonicity assumption 1. Again we represent a generic attention rule μ as a long vector $\boldsymbol{\mu}$, then \mathbf{R} is easily constructed as a collection of inequalities taking the form: $\mu(T|S) - \mu(T|S-a) \leq 0$, for all $a \in S-T$. And note that \mathbf{R} does not depend on any preference.

Next we consider \mathbf{C}_\succ . We recall Definition 5, and given the choice rule π and some preference \succ , the only possible triangular attention rule that can be constructed is the following:

$$\mu(T|S) = \sum_{k: a_{k,\succ} \in S} \mathbf{1}(T = S \cap L_{k,\succ}) \cdot \pi(a_{k,\succ}|S),$$

which defines the mapping \mathbf{C}_\succ . The mapping depends on the preference/hypothesis, since the triangular attention rule does.

Along the construction, both \mathbf{R} and \mathbf{C}_\succ are unique, hence showing \mathbf{R}_\succ is uniquely determined by the preference \succ . ■

We combine \mathbf{R} and \mathbf{C}_\succ in the proof of Theorem 4, and illustrate the form of the final constraint matrix \mathbf{R}_\succ . Fix some preference \succ and let μ be the triangular attention rule shown in the previous proof. Monotonicity (i.e., \mathbf{R}) requires that $\mu(T|S) - \mu(T|S - a_k) \leq 0$. This is trivially satisfied if T is not a lower contour set in S since $\mu(T|S) = 0$. Assume $T = S \cap L_{\ell,\succ} \neq \emptyset$ is a lower contour set in S (without loss of generality let $a_{\ell,\succ} \in S$). Then it will also be a lower contour set in $S - a_k$. Moreover, $a_k \in S - T$ implies that $a_k \succ a_{\ell,\succ}$, and hence the monotonicity assumption translates into $\pi(a_{\ell,\succ}|S) - \pi(a_{\ell,\succ}|S - a_k) \leq 0$, for all $a_k \succ a_{\ell,\succ}$ in S . We have the following algorithm constructing \mathbf{R}_\succ *directly*, albeit the idea does come from the two auxiliary matrices \mathbf{R} and \mathbf{C}_\succ .

Algorithm 1 Construction of \mathbf{R}_\succ .

Ensure: Ordering of elements in π is set.

Require: Set a preference \succ .

```
 $\mathbf{R}_\succ \leftarrow$  empty matrix
for  $S$  in  $\mathcal{X}$  do
  for  $a$  in  $S$  do
    for  $b \prec a$  in  $S$  do
       $\mathbf{R}_\succ \leftarrow$  add row corresponding to  $\pi(b|S) - \pi(b|S - a) \leq 0$ .
    end for
  end for
end for
```

The only input needed in the previous algorithm is the preference \succ , which we are interested in testing against, and the precise ordering in the vector π stacking the identifiable choice rule π . Before closing this subsection, we compute the number of constraints (i.e. rows) in \mathbf{R}_\succ . Following the previous algorithm, it is

$$\#\text{row}(\mathbf{R}_\succ) = \sum_{S \in \mathcal{X}} \sum_{a, b \in S} \mathbb{1}(b \prec a) = \sum_{S \in \mathcal{X}, |S| \geq 2} \binom{|S|}{2} = \sum_{k=2}^{|\mathcal{X}|} \binom{|\mathcal{X}|}{k} \binom{k}{2}.$$

Not surprisingly, the number of constraints increases very fast with the size of the grand set $|\mathcal{X}|$.

7.2 Null Hypotheses and Test Statistics

Given the identification result in Theorem 4, we can replace the unobserved but identifiable choice rule with its estimate to conduct estimation and inference of the (partially identifiable) preferences. We can also conduct specification testing by evaluating whether the identified set Θ_π is empty. To proceed, we assume the following data structure.

Assumption 2 (DGP). The data is a random sample of choice problems Y_i and corresponding choices y_i , $\{(y_i, Y_i) : y_i \in Y_i, 1 \leq i \leq N\}$, generated by the underlying choice rule $\mathbb{P}[y_i = a | Y_i = S] = \pi(a|S)$, with $\mathbb{P}[Y_i = S] \geq \underline{p} > 0$ for all $S \in \mathcal{X}$.

We only assume the data is generated from some choice rule π . We allow for the possibility that it is not RAM, since our identification result permits falsifying the RAM representation: π has a RAM representation if and only if Θ_π is not empty according to Theorem 4. In addition, we only assume that the a choice problem Y_i and the corresponding selection $y_i \in Y_i$ are observed for each

unit, while the underlying (possibly random) consideration set for the i^{th} decision maker remains unobserved (i.e., the set T in Definition 2 and Figure 1).

The estimated choice rule is denoted by $\hat{\pi}$,

$$\hat{\pi}(a|S) = \frac{\sum_{1 \leq i \leq N} \mathbf{1}(y_i = a, Y_i = S)}{\sum_{1 \leq i \leq N} \mathbf{1}(Y_i = S)}, \quad a \in S, \quad S \in \mathcal{X}.$$

For convenience, we represent $\hat{\pi}(\cdot|S)$ by the vector $\hat{\boldsymbol{\pi}}_S$, and its population counterpart by $\boldsymbol{\pi}_S$. The choice rules are stacked into a long vector, denoted by $\hat{\boldsymbol{\pi}}$, with the population counterpart $\boldsymbol{\pi}$.

We discuss how test statistics are constructed for different null hypotheses. In particular, we consider three cases most useful for empirical analysis: (i) $H_0 : \succ \in \Theta_\pi$ for testing against a single preference; (ii) $H_0 : \mathcal{P} \subset \Theta_\pi$ for testing against a collection of preferences; and (iii) $H_0 : \mathcal{P} \cap \Theta_\pi \neq \emptyset$, for testing model compatibility/correct specification.

We consider Studentized test statistics, and hence we introduce some additional notation. Let $\sigma_{\pi, \succ}$ be the standard deviation of $\mathbf{R}_\succ \hat{\boldsymbol{\pi}}$, and $\hat{\sigma}_\succ$ be its plug-in estimate. That is,

$$\sigma_{\pi, \succ} = \sqrt{\text{diag}(\mathbf{R}_\succ \boldsymbol{\Omega}_\pi \mathbf{R}_\succ')} \quad \text{and} \quad \hat{\sigma}_\succ = \sqrt{\text{diag}(\mathbf{R}_\succ \hat{\boldsymbol{\Omega}} \mathbf{R}_\succ')},$$

where $\text{diag}(\cdot)$ denotes the operator that extracts the diagonal elements of a square matrix or constructs a diagonal matrix when applied to a vector. Here $\boldsymbol{\Omega}_\pi$ is block diagonal, with blocks given by $\frac{1}{\mathbb{P}[Y_i=S]} \boldsymbol{\Omega}_{\pi, S}$, and $\boldsymbol{\Omega}_{\pi, S} = \text{diag}(\boldsymbol{\pi}_S) - \boldsymbol{\pi}_S \boldsymbol{\pi}_S'$. The estimator $\hat{\boldsymbol{\Omega}}$ is simply constructed by plugging in the estimated choice rule. For two matrices with the same dimension, \oslash denotes elementwise division (i.e, Hadamard division).

Testing Against a Single Preference – $H_0 : \succ \in \Theta_\pi$

This null hypothesis is useful if the researcher believes a certain preference represents the underlying data generating process. It also serves as the basis for constructing a confidence set for Θ_π , or for ranking preferences according to their (im)plausibility in repeated sampling (for example, via employing associated p-values).

Given a specific preference, the test statistic is constructed as the *maximum of scaled sample*

moments:

$$\mathcal{T}(\succ) = \sqrt{N} \cdot \max \left\{ (\mathbf{R}_\succ \hat{\boldsymbol{\pi}}) \odot \hat{\boldsymbol{\sigma}}_\succ, 0 \right\},$$

that is, the test statistic is the largest element of the vector $\sqrt{N}(\mathbf{R}_\succ \hat{\boldsymbol{\pi}}) \odot \hat{\boldsymbol{\sigma}}_\succ$ if it is positive, or zero otherwise. The reasoning behind such construction is straightforward: if the preference is compatible with the underlying choice rule, then in the population we have $\mathbf{R}_\succ \boldsymbol{\pi} \leq \mathbf{0}$.

Other test statistics have been proposed for testing moment inequalities, and usually the specific choice depends on the context. When many moment inequalities can be potentially violated simultaneously, it is usually preferred to use a statistic based on truncated Euclidean norm. In our problem, however, we expect only a few moment inequalities be violated, and therefore we prefer to employ $\mathcal{T}(\succ)$. Having said this, the large sample approximation results given in the appendix can be adapted to handle other test statistics commonly encountered in the literature on moment inequalities.

Testing Against a Collection of Preferences – $H_0 : \mathcal{P} \subset \Theta_\pi$

We consider this testing problem because it naturally generalizes the previous one, and because it can be easily accommodated by our methodological results. Despite being more general, this approach may have limited use in practice: even if the null hypothesis is rejected, it is unclear how such information should be incorporated into data analysis and decision theoretic modeling. Further, it may suffer from low power in finite samples because of testing against many preferences, especially when only one or two preferences in the collection are incompatible with the underlying choice rule.

Let \mathcal{P} be a collection of preferences, then the test statistic is constructed as

$$\mathcal{T}(\mathcal{P}_\forall) = \max_{\succ \in \mathcal{P}} \mathcal{T}(\succ) = \max_{\succ \in \mathcal{P}} \left[\sqrt{N} \cdot \max \left\{ (\mathbf{R}_\succ \hat{\boldsymbol{\pi}}) \odot \hat{\boldsymbol{\sigma}}_\succ, 0 \right\} \right].$$

We use \mathcal{P}_\forall to emphasize that the null hypothesis is “all preferences in \mathcal{P} are compatible with the underlying choice rule”.

Testing Model Compatibility – $H_0 : \mathcal{P} \cap \Theta_\pi \neq \emptyset$

Given a collection of preferences, an empirically relevant question is whether *any* of them is compatible with the data generating process – a basic model specification question. That is, whether the null hypothesis $H_0 : \mathcal{P} \cap \Theta_\pi \neq \emptyset$ should be rejected. If the null hypothesis is rejected, then certain features shared by the collection of preferences is incompatible with the underlying decision theory (up to Type I error). This idea has been considered previously in [Bugni et al. \(2015\)](#) and references therein.

For a concrete example, consider the question: Is $a \succ b$ compatible with the data generating process? As long as there are more than 2 alternatives in the grand set, a question like this can be accommodated by setting $\mathcal{P} = \{\succ : a \succ b\}$. When this null hypothesis is rejected, then there is evidence in favor of a not preferred to b (up to Type I error). Of course with more preferences included in the collection, it becomes more difficult to reject the null hypothesis. Another example is to set \mathcal{P} to be the collection of all possible preferences. Then, the null hypothesis becomes $H_0 : \Theta_\pi \neq \emptyset$, a specification test for RAM. Rejection in this case implies that at least one of the underlying assumptions is violated, and the data generating process cannot be represented by RAM (up to Type I error).

The test statistic for this case is constructed as follows:

$$\mathcal{T}(\mathcal{P}_\exists) = \sqrt{N} \cdot \min_{\succ \in \mathcal{P}} \mathcal{T}(\succ) = \min_{\succ \in \mathcal{P}} \left[\sqrt{N} \cdot \max \left\{ (\mathbf{R}_\succ \hat{\boldsymbol{\pi}}) \odot \hat{\boldsymbol{\sigma}}_\succ, 0 \right\} \right].$$

We use \mathcal{P}_\exists to emphasize that the null hypothesis is “at least one preference in \mathcal{P} is compatible with the data generating process”.

7.3 Critical Values and Implementation

We describe how critical values leading to uniformly valid testing procedures can be constructed based on simulating from multivariate normal distributions. The literature on moment inequalities includes several other approaches, which we discuss briefly in [Appendix A.3](#).

To illustrate the intuition behind our construction, consider the test statistic $\mathcal{T}(\succ)$:

$$\mathcal{T}(\succ) = \max \left\{ \left(\mathbf{R}_\succ \sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + \sqrt{N} \mathbf{R}_\succ \boldsymbol{\pi} \right) \odot \hat{\boldsymbol{\sigma}}_\succ, 0 \right\}.$$

By the central limit theorem, the first component $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ is approximately distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\pi)$. The second component, $\mathbf{R}_>\boldsymbol{\pi}$, although unknown, is bounded above by zero under the null hypothesis. Motivated by these observations, we propose approximating the distribution of $\mathcal{T}(>)$ by simulation as follows:

$$\mathcal{T}^*(>) = \sqrt{N} \cdot \max \left\{ \left(\mathbf{R}_>\mathbf{z}^* + \kappa_N \min\{\mathbf{R}_>\hat{\boldsymbol{\pi}}, 0\} \right) \oslash \hat{\boldsymbol{\sigma}}_>, 0 \right\}.$$

Here \mathbf{z}^* is a random vector simulated from the distribution $\mathcal{N}(\mathbf{0}, \hat{\boldsymbol{\Omega}}/N)$, and κ_N is a non-negative sequence that shrinks to zero slowly. One extreme choice is $\kappa_N = 0$, so that the upper bound $\mathbf{0}$ is used to replace the unknown $\mathbf{R}_>\boldsymbol{\pi}$. Such a choice also delivers uniformly valid inference in large samples, and is usually referred to as “critical value based on the least favorable model” (see Appendix A.3). However, for practical purposes it is better to be less conservative and thus κ_N is chosen to be a positive slowly vanishing sequence. In our implementation we employ $\kappa_N = 1/\sqrt{\ln N}$, which turns out to work well in the simulations described in Section 8.

In practice, M simulations are conducted to obtain the simulated statistics $\{\mathcal{T}_m^*(>) : 1 \leq m \leq M\}$. Then, given some $\alpha \in (0, 1)$, the critical value is constructed as

$$c_\alpha(>) = \inf \left\{ t : \frac{1}{M} \sum_{m=1}^M \mathbb{1} \left(\mathcal{T}_m^*(>) \leq t \right) \geq 1 - \alpha \right\},$$

and the null hypothesis $H_0 : > \in \Theta_\pi$ is rejected if and only if $\mathcal{T}(>) > c_\alpha(>)$. Alternatively, one can compute the p-value as

$$\text{pVal}(>) = \frac{1}{M} \sum_{m=1}^M \mathbb{1} \left(\mathcal{T}_m^*(>) > \mathcal{T}(>) \right).$$

For the other two hypotheses, critical values are constructed similarly:

$$c_\alpha(\mathcal{P}_\forall) = \inf \left\{ t : \frac{1}{M} \sum_{m=1}^M \mathbb{1} \left(\mathcal{T}^*(\mathcal{P}_\forall) \leq t \right) \geq 1 - \alpha \right\}, \quad \mathcal{T}^*(\mathcal{P}_\forall) = \max_{> \in \mathcal{P}} \mathcal{T}^*(>),$$

and

$$c_\alpha(\mathcal{P}_\exists) = \inf \left\{ t : \frac{1}{M} \sum_{m=1}^M \mathbb{1} \left(\mathcal{T}^*(\mathcal{P}_\exists) \leq t \right) \geq 1 - \alpha \right\}, \quad \mathcal{T}^*(\mathcal{P}_\exists) = \min_{> \in \mathcal{P}} \mathcal{T}^*(>).$$

To justify the critical values we proposed, it is important to address uniformity issues because in finite samples the moment inequalities could be close to binding. A testing procedure is (asymptotically) uniform among a class of data generating processes, if the asymptotic size does not exceed the nominal level across this class. The following theorem shows that conducting inference using the critical values proposed above is uniformly valid.

Theorem 5 (Uniformly Valid Testing). Assume Assumption 2 holds.

1. Let Π be a class of choice rules, and \succ a preference, such that: (i) for each $\pi \in \Pi$, $\succ \in \Theta_\pi$; and (ii) $\inf_{\pi \in \Pi} \min(\sigma_{\pi, \succ}) > 0$. Then,

$$\limsup_{N \rightarrow \infty} \sup_{\pi \in \Pi} \mathbb{P}[\mathcal{T}(\succ) > c_\alpha(\succ)] \leq \alpha.$$

2. Let Π be a class of choice rules and \mathcal{P} a class of preferences, such that: (i) for each $\pi \in \Pi$, $\mathcal{P} \subset \Theta_\pi$; and (ii) $\inf_{\pi \in \Pi, \succ \in \mathcal{P}} \min(\sigma_{\pi, \succ}) > 0$. Then,

$$\limsup_{N \rightarrow \infty} \sup_{\pi \in \Pi} \mathbb{P}[\mathcal{T}(\mathcal{P}_\forall) > c_\alpha(\mathcal{P}_\forall)] \leq \alpha.$$

3. Let Π be a class of choice rules and \mathcal{P} a class of preferences, such that: (i) for each $\pi \in \Pi$, $\mathcal{P} \cap \Theta_\pi \neq \emptyset$; and (ii) $\inf_{\pi \in \Pi, \succ \in \mathcal{P}} \min(\sigma_{\pi, \succ}) > 0$. Then,

$$\limsup_{N \rightarrow \infty} \sup_{\pi \in \Pi} \mathbb{P}[\mathcal{T}(\mathcal{P}_\exists) > c_\alpha(\mathcal{P}_\exists)] \leq \alpha.$$

The proof is postponed to Appendix B.2. The only requirement is that each moment condition is nondegenerate so that the normalized statistics are well-defined in large samples, but no restrictions on correlations among moment conditions are imposed.

Given the uniformly valid testing procedure presented above, we obtain confidence sets for the (partially) identified preferences by test inversion. Therefore, using Theorem 5, we obtain the following confidence interval

$$\mathcal{C}(\alpha) = \left\{ \succ : \mathcal{T}(\succ) \leq c_\alpha(\succ) \right\},$$

whose asymptotic uniform coverage rate is at least $1 - \alpha$. Our inference methods offer uniformly valid confidence intervals for each member of the partially identified set with prespecified probability,

which is the preferred approach in the partially identified literature (Imbens and Manski, 2004).

7.4 Connection to J-test Inference Approach

The testing methods we proposed are connected to a class of inference strategies based on projection residuals, sometimes also known as J-tests. To describe this alternative inference approach, recall that any preference induces a surjective map from attention rules to choice rules (Definition 3), denoted by $\tilde{\mathbf{C}}_{\succ}$. Then, by our definition of partially identified preferences, $\succ \in \Theta_{\pi}$ if and only if π belongs to $\{\tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu} : \mathbf{R}\boldsymbol{\mu} \leq 0\}$, where \mathbf{R} represents the monotonicity assumption imposed on attention rules. Therefore, inference can be based on the following fact:

$$\succ \in \Theta_{\pi} \quad \text{if and only if} \quad \inf_{\boldsymbol{\mu}: \mathbf{R}\boldsymbol{\mu} \leq 0} (\boldsymbol{\pi} - \tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu})' \mathbf{W} (\boldsymbol{\pi} - \tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu}) = 0,$$

where \mathbf{W} is some positive definite weighting matrix. Hence a preference is compatible with π is equivalent to a zero residual from projecting $\boldsymbol{\pi}$ to the corresponding set $\{\tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu} : \mathbf{R}\boldsymbol{\mu} \leq 0\}$. This strategy is used by Kitamura and Stoye (2016) in random utility models. Chernozhukov et al. (2015) and Fang and Santos (2015) consider such problems in a more general context to test whether a collection of conditional moment conditions have solution subject to (in)equality constraints on the parameter, and develop uniformly valid inference based on coupling arguments.

To see the connection of J-tests to moment inequality testing, observe that if the mapping defined by $\tilde{\mathbf{C}}_{\succ}$ is invertible, then the above reduces to

$$\succ \in \Theta_{\pi} \quad \text{if and only if} \quad \mathbf{R}\tilde{\mathbf{C}}_{\succ}^{-1}\boldsymbol{\pi} \leq 0.$$

Such reduction may not be feasible analytically, or it may be numerically prohibitive. With a careful inspection of our problem, we showed that it is without loss of generality to focus on triangular attention rules, which can be uniquely constructed from preferences and choice rules. That is, we showed how the inversion $\tilde{\mathbf{C}}_{\succ}^{-1}$ is constructed, which is denoted by \mathbf{C}_{\succ} in Theorem 4 and its proof. Moreover, we provide an algorithm which constructs the constraint matrix directly, avoiding the detour of forming it as the product of two matrices.

Compared to directly testing inequality constraints as we propose, employing a J-test inference

procedure has two potential drawbacks. First, constructing the projection residual requires numerical minimization, which can be computationally costly especially when the dimension of $\boldsymbol{\mu}$ is nontrivial. Second, the J-test statistic is essentially the (weighted) Euclidean norm of the projection residual, which may suffer from low power if only a few inequalities are violated.

8 Simulation Evidence

This section gives a summary of a simulation study conducted to assess the finite sample properties of our proposed econometric methods. We consider a class of Logit attention rules indexed by ς :

$$\mu_\varsigma(T|S) = \frac{w_{T,\varsigma}}{\sum_{T' \subset S} w_{T',\varsigma}}, \quad w_{T,\varsigma} = |T|^\varsigma,$$

where $|T|$ is the cardinality of T . Thus the decision maker pays more attention to larger sets if $\varsigma > 0$, and pays more attention to smaller sets if $\varsigma < 0$. When ς is very small (negative and large in absolute magnitude), the decision maker almost always pays attention to singleton sets, hence nothing will be learned about the underlying preference from the choice data.

Other details on the data generating process used in the simulation study are as follows. First, the grand set X consists of five alternatives, a_1, a_2, a_3, a_4 and a_5 . Without loss of generality, assume the underlying preference is $a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5$. Second, the data consists of choice problems of size four and five. That is, only $X, X - a_1, X - a_2, X - a_3, X - a_4, X - a_5$ will realize in data, according to probability: $\mathbb{P}[Y_i = X] = 1/2$, and $\mathbb{P}[Y_i = X - a_k] = 1/10$ for $1 \leq k \leq 5$. Third, given a specific realization of Y_i , a consideration set is generated from the logit attention model with $\varsigma \in \{0, 1, 2\}$, after which the choice y_i is determined by the aforementioned preference. Finally, the observed data is a random sample $\{(y_i, Y_i) : 1 \leq i \leq N\}$, where the sample size N can be 2,500, 5,000, 7,500 or 10,000.

Before proceeding, we list the set of identified preferences (i.e., the set of preferences compatible with the data generating process) for different ς . The following can be obtained from our Lemma 2 with straightforward calculations.

ς	Θ_ς
0	all possible preferences
1	$\{\succ: a_3 \succ a_4 \succ a_5\}$
2	$\{\succ: a_2 \succ a_3 \succ a_4 \succ a_5\}$

We test against five preferences:

$$\begin{aligned}
H_{0,1} &: a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5, & \text{compatible with } \varsigma = 0, 1, 2 \\
H_{0,2} &: a_2 \succ a_3 \succ a_4 \succ a_5 \succ a_1, & \varsigma = 0, 1, 2 \\
H_{0,3} &: a_3 \succ a_4 \succ a_5 \succ a_2 \succ a_1, & \varsigma = 0, 1 \\
H_{0,4} &: a_4 \succ a_5 \succ a_3 \succ a_2 \succ a_1, & \varsigma = 0 \\
H_{0,5} &: a_5 \succ a_4 \succ a_3 \succ a_2 \succ a_1, & \varsigma = 0
\end{aligned}$$

Both $H_{0,1}$ and $H_{0,2}$ are compatible with the data generating process for all the three ς values we consider. $H_{0,3}$ is compatible with $\varsigma = 0$ or 1. Finally $H_{0,4}$ and $H_{0,5}$ are only compatible with $\varsigma = 0$. Therefore, our simulation has 3 (different ς) \times 4 (different N) \times 5 (different null hypotheses) = 60 designs. For each design, 2,000 simulation repetitions are used, and the five null hypotheses are tested using our proposed method at the 5% nominal level.

Simulation results are summarized in Table 1. We first focus on $\varsigma = 0$ (row 1–5 in Table 1). In this case, all the five preferences are compatible with the data generating process, and we can see that the rejection probabilities do not exceed the 0.05 nominal level, except when the sample size is small. In fact, the empirical rejection rate can be far below 0.05. This illustrates a generic feature of any (reasonable) procedure for testing moment inequalities – to maintain uniform asymptotic size control, empirical size is below the nominal level when the inequalities are far from binding. Consider $H_{0,1}$. The population moment inequalities are $(-0.275, -0.275, -0.138, -0.138, -0.069, -0.069, -0.069, -0.034, -0.034, -0.034, -0.034, -0.034)$, of which most are far from binding, hence the empirical rejection probabilities (row 1, column 1–4 in Table 1) are almost zero. On the other hand, hypothesis $H_{0,5}$ is very different from the preference we use in simulation (but it is still compatible with the DGP of $\varsigma = 0$), hence one should expect that the inequality constraints are close to binding, which is indeed the case: $(-0.017, -0.017, -0.017, -0.017, -0.017, -0.009,$

Table 1. Empirical Rejection Probabilities

	Feasible				Infeasible			
	$N = 2500$	5000	7500	10000	$N = 2500$	5000	7500	10000
$\varsigma = 0$								
$H_{0,1}$	0.001	0.000	0.000	0.000	0.048	0.002	0.001	0.000
$H_{0,2}$	0.018	0.010	0.006	0.004	0.057	0.054	0.047	0.052
$H_{0,3}$	0.030	0.024	0.019	0.014	0.060	0.058	0.058	0.062
$H_{0,4}$	0.045	0.035	0.028	0.018	0.069	0.066	0.064	0.062
$H_{0,5}$	0.054	0.046	0.032	0.022	0.076	0.072	0.065	0.064
$\varsigma = 1$								
$H_{0,1}$	0.002	0.000	0.000	0.000	0.070	0.024	0.005	0.000
$H_{0,2}$	0.008	0.002	0.002	0.000	0.056	0.043	0.054	0.046
$H_{0,3}$	0.041	0.053	0.044	0.046	0.054	0.065	0.055	0.060
$H_{0,4} \times$	0.095	0.118	0.118	0.145	0.102	0.127	0.126	0.150
$H_{0,5} \times$	0.153	0.191	0.196	0.224	0.154	0.200	0.200	0.228
$\varsigma = 2$								
$H_{0,1}$	0.122	0.008	0.001	0.000	0.174	0.065	0.043	0.026
$H_{0,2}$	0.080	0.007	0.000	0.000	0.164	0.060	0.046	0.046
$H_{0,3} \times$	0.116	0.113	0.138	0.154	0.154	0.148	0.173	0.174
$H_{0,4} \times$	0.222	0.284	0.376	0.440	0.244	0.310	0.399	0.462
$H_{0,5} \times$	0.338	0.450	0.568	0.650	0.354	0.462	0.582	0.668

Shown in the table are empirical rejection probabilities testing the five null hypothesis through 2,000 simulations, with nominal size 0.05. Three DGPs indexed by $\varsigma \in \{0, 1, 2\}$ are described in the text. For each simulation repetition, four sample sizes are considered $N \in \{2,500, 5,000, 7,500, 10,000\}$. Details on the five hypotheses are given in the text, and a hypothesis is labeled with \times if it should be rejected (i.e. the corresponding preference is not compatible with the DGP). **Feasible**: testing based on the proposed critical value. **Infeasible**: testing based on the (infeasible) oracle critical value. This serves as the benchmark since it has the highest power while maintaining size uniformly.

$-0.009, -0.009, -0.004, -0.004, -0.002, -0.002$), and the rejection probabilities are much closer to the 0.05 nominal level (row 5, column 1–4 in Table 1).

Overall, our proposed testing procedure controls size very well. This can be seen from other cases: row 6–8, 11 and 12 in Table 1. For any inference procedure with desirable size property, one concern would be its power. Consider $\varsigma = 1$. Here two hypotheses/preferences are *not* compatible with the DGP, $H_{0,4}$ and $H_{0,5}$. Population moment inequalities for $H_{0,4}$ are $(-0.025, -0.025, -0.025, -0.025, -0.025, -0.019, 0.000, 0.000, 0.000, 0.006, 0.006, 0.006)$, and are only marginally violated. Even in this difficult case, our procedure still delivers nontrivial power: 0.1 with only 2,500 sample

size (row 9 in Table 1).

Higher power is obtained when more inequality constraints are violated. Consider the hypothesis $H_{0,5}$ when $\varsigma = 2$. The population inequalities are $(-0.033, -0.033, -0.033, -0.033, -0.033, 0.008, 0.008, 0.008, 0.008, 0.012, 0.012, 0.012)$. Still, the constraints are only marginally violated, since the largest one is merely 0.012, yet our procedure delivers nontrivial power: the rejection probability exceeds 0.3 for sample size 2,500, and can be as high as 0.65 for sample size 10,000.

In the online Supplemental Appendix, we report additional results for (i) all possible 120 preferences, and (ii) inference using other critical values from some other methods in the moment inequality literature.

9 Conclusion

We introduced a limited attention model allowing for a general class of monotonic stochastic consideration rules, which we called a Random Attention Model (RAM). We showed that this model nests several important recent contributions in both economic theory and econometrics, in addition to other classical results from choice and decision theory. Using our RAM framework, we obtained a testable theory of revealed preferences and developed partial identification results for the decision maker's unobserved strict preference ordering. Our results include a precise constructive characterization of the identified set for preferences as well as uniformly valid inference methods based on this characterization. We illustrated good finite sample performance of our methods in a simulation experiment. Finally, we provide the general-purpose R software package `ramchoice`, which allows other researchers to easily employ our econometric methods in empirical applications.

We regard these results as a first step in a research program, since there are many open questions and extensions within our proposed RAM framework worth investigating. For example, from an economic theory perspective, two natural lines of inquiry are (i) to develop a theory of revealed attention allowing to learn about random attention maps from observed choice behavior, and (ii) to expand the domain of the choice problem so that the effect of framing (e.g., presentation or advertising) on random attention map can be studied. Similarly, from an econometric perspective, two other natural extensions of our work include (i) allowing for observed covariates to enter RAM and hence the use of other (conditional) choice probability estimates, and (ii) learning about

unobserved heterogeneity in order to develop counterfactual policy analysis (Theorem 3 gives a first result along this line). Research along these lines is underway.

A Appendix A: Extensions and Other Material

This appendix collects extensions of our work and other related material not included in the main text to improve the exposition.

A.1 Other Examples of RAM

Here we provide more examples of random consideration sets that satisfy our key monotonicity assumption (Assumption 1).

1. (FULL ATTENTION) The decision maker considers everything with probability one: $\mu(T|S) = \mathbb{1}(T = S)$.
2. (TOP N; Salant and Rubinstein, 2008) The decision maker faces a list of alternatives created by some ordering. She pays attention to the first N elements among available alternatives (e.g., first page of Google search results). If the number of available alternatives is less than N , she pays attention to the entire set. Formally, let $S(k, R)$ denote the set of first k elements in S according ordering R provided that $k \leq |S|$. ($S(|S|, R)$ is equal to S .) In our framework: $\mu(T|S) = \mathbb{1}(T = S(\min\{|S|, N\}, R))$.
3. (SATISFICING CONSIDERATION SET) The decision maker observes alternatives sequentially from a pre-determined list. The order of alternatives is unknown to the decision maker in the beginning of the search and uncovers them during the search process. The decision maker stops searching upon finding a satisfactory alternative (Simon, 1955). If there is no such alternative, she searches the entire budget set. Formally, given the list L , $RS_L(S)$ denotes the range of search (the consideration set) when the budget set is S . In our framework: $\mu(T|S) = \mathbb{1}(T = RS_L(S))$.
4. (AT MOST k ALTERNATIVES) The decision maker considers at most k alternatives for any decision problem. If there are more alternatives than k , she considers only subsets including exactly k alternatives with equal probability. If there are less alternatives than k , she considers everything. In our framework:

$$\mu(T|S) = \begin{cases} 1 & \text{if } |S| \leq k \text{ and } T = S \\ \binom{|S|}{k}^{-1} & \text{if } |S| > k \text{ and } |T| = k \\ 0 & \text{otherwise} \end{cases}$$

5. (UNIFORM CONSIDERATION) The decision maker considers any subset of the feasible set with equal probabilities. That is, for all $T \subset S$, $\mu(T|S) = 1/(2^{|S|} - 1)$.
6. (FIXED CORRELATED CONSIDERATION; Barberà and Grodal, 2011; Aguiar, 2017) The decision maker pays attention to each alternative with a fixed probability but the consideration of alternatives is potentially correlated. Formally, let ω be a probability distribution over \mathcal{X} . Then each alternative $a \in X$ is considered with a fixed probability $\sum_{T \in \mathcal{X}} \mathbb{1}(a \in T) \cdot \omega(T)$ for all $S \ni a$. In our framework:

$$\mu(T|S) = \sum_{T' \in \mathcal{X}} \mathbb{1}(T' \cap S = T) \cdot \omega(T').$$

7. (ORDERED LOGIT ATTENTION) This is another generalization of the logit attention example. The decision maker ranks all subsets in terms of their attention priority, and she only considers subsets which are maximal with respect to that ordering. When there are several best subsets, the decision maker considers each of them with certain frequency as in the Logit Attention example. Thus, consideration sets are constructed in the spirit of standard maximization paradigm. Formally, let \succeq be a complete

and transitive priority order over subsets \mathcal{X} . $S \succeq T$ reads as “ S has a higher attention priority than T ”. The case when S and T have the same attention priority is denoted by $S \bowtie T$. Formally,

$$\mu(T|S) = \begin{cases} \frac{w_T}{\sum_{T' \subset S} \mathbf{1}(T' \bowtie T) \cdot w_{T'}} & \text{if } T \text{ is } \succeq \text{-best in } S \\ 0 & \text{otherwise} \end{cases}$$

8. (ELIMINATION BY ASPECTS, GENERAL) The example is similar to the one given in the main text, except that when the decision maker picks an irrelevant aspect, she selects a subset at random drawn from the uniform distribution. Formally,

$$\mu(T|S) = \sum_{B_i \cap S = T} \omega(i) + \frac{1}{2^{|S|} - 1} \sum_{B_k \cap S = \emptyset} \omega(k).$$

The probability that the decision maker selects an irrelevant aspect is $\sum_{k: B_k \cap S = \emptyset} \omega(k)$. In this case, T is randomly chosen, which is reflected by the number $\frac{1}{2^{|S|} - 1}$. We can generate similar examples. For instance, if the initial screening is not successful (choosing an irrelevant aspect), the decision maker may consider all the alternatives. Formally,

$$\mu(T|S) = \begin{cases} \sum_{B_i \cap S = T} \omega(i) & \text{if } T \neq S \\ \sum_{B_k \cap S = S, \emptyset} \omega(k) & \text{if } T = S \end{cases}$$

9. (STOCHASTIC SATISFICING) Suppose the satisficer faces multiple lists. The probability that the decision maker faces list L is denoted by $p(L)$. As opposed to Example 3, consideration sets are stochastic. Formally,

$$\mu(T|S) = \sum_{L: T = RS_L(S)} p(L)$$

where $RS_L(S)$ is the range of search when the budget set is S and L is the list.

10. (1/ N RULE) The decision maker utilizes N different sources of recommendations with equal probabilities. Given a fixed source s , she considers only top k_s alternatives according to ordering R_s that source s is using, which could be different from her preference. For example, she utilizes either Google or Yahoo with equal probability. Once she decides which one to look at, she pays attention to only the products appearing in the first page of the corresponding search result. Formally,

$$\mu(T|S) = \frac{|\{s \mid T = S(\min\{k_s, |S|\}, R_s)\}|}{N}$$

where $S(k, R)$ denotes the first k alternatives in S according to R .

11. (RANDOM PRODUCT NETWORK; Masatlioglu and Suleymanov, 2017) Consider a decision maker faced with a product network \mathcal{N} . If $(a, b) \in \mathcal{N}$, then the decision maker who considers a is recommended alternative b (or alternatively, b is linked to a). The decision maker starts search from a random starting point. Given a realized starting point in the product network, the decision maker considers all alternatives which are directly or indirectly linked to that starting point. Formally, let η be a probability distribution on X . Then

$$\mu(T|S) = \sum_{a \in S} \mathbf{1}(T = N_a(S)) \cdot \frac{\eta(a)}{\sum_{b \in S} \eta(b)}$$

where $N_a(S)$ denotes all alternatives which are directly or indirectly linked to a in S .

12. (PATH DEPENDENT CONSIDERATION; Suleymanov, 2017) This example is similar to the one above, except that now the decision maker starts searching from a fixed starting point $a^* \notin X$ (the default option) and takes a random path on a network. Let $X^* = X \cup \{a^*\}$, and \mathcal{P}_{a^*} stands for all possible paths in X^* with the initial node a^* . When the choice set is X , the decision maker takes the path

$\rho \in \mathcal{P}_{a^*}$ with probability $\gamma(\rho)$. Given $\rho \in \mathcal{P}_{a^*}$ and $S \in \mathcal{X}$, $\rho_S \in \mathcal{P}_{a^*}$ is the subpath of ρ in S with the same initial node a^* . For any ρ , let $V(\rho)$ be the vertices of the path ρ excluding the initial node. Then

$$\mu(T|S) = \sum_{\rho \in \mathcal{P}_{a^*}} \mathbb{1}(T = V(\rho_S)) \cdot \gamma(\rho).$$

This model is a subset of RAM with a default option.

A.2 Limited Data

We discuss how our results can be adapted to handle limited data, that is, settings where not all possible choice probabilities or choice problems are observed.

A.2.1 Theory and Identification

To investigate the implications of random attention models with limited data, assume that an outside observer does not have access to choices from all menus. In other words, suppose we observe choices from the set of menus \mathcal{S} . Let π_{obs} denote the observed choice behavior. We say that π_{obs} is consistent with the random attention model if there exists π defined on the entire domain \mathcal{X} such that $\pi(a|S) = \pi_{\text{obs}}(a|S)$ for all $S \in \mathcal{S}$ and π is a RAM. We call such π an extension of π_{obs} . As [De Clippel and Rozen \(2014\)](#) point out, it is possible that π_{obs} satisfies the acyclicity of \mathbf{P} even though it is inconsistent with RAM. Here we provide an example with stochastic choices in which the same problem occurs.

Example A.1. Let $\mathcal{S} = \{\{a, b, c, d\}, \{b, c, d\}, \{a, c\}\}$. Consider the following choice rule.

$\pi_{\text{obs}}(\cdot S)$	$S = \{a, b, c, d\}$	$\{b, c, d\}$	$\{a, c\}$
a	1/4		1/5
b	1/4	1/5	
c	1/4	3/5	4/5
d	1/4	1/5	

In this example, we observe choices only from 3 menus instead of 15 potential menus. We show that these observations are sufficient to conclude that π_{obs} is not consistent with RAM. Suppose there exists a RAM π that extends π_{obs} . Then the observation $\pi(a|\{a, b, c, d\}) > \pi(a|\{a, c\})$ tells us that either aPb or aPd . To see this, notice that at least one of $\pi(a|\{a, b, c, d\}) > \pi(a|\{a, c, d\})$ and $\pi(a|\{a, c, d\}) > \pi(a|\{a, c\})$ must hold or we get a contradiction. On the other hand, from $\pi(b|\{a, b, c, d\}) > \pi(b|\{b, c, d\})$ and $\pi(d|\{a, b, c, d\}) > \pi(d|\{b, c, d\})$ we learn that dPa and bPa . Hence, even though acyclicity of \mathbf{P} is satisfied on the limited domain, there does not exist an extension of π_{obs} that is RAM. \perp

From the example above we can note the following: if $\pi_{\text{obs}}(a|S) > \pi_{\text{obs}}(a|S - A)$, then it must be that aPb for some $b \in A$. Now suppose $\pi_{\text{obs}}(a|S) + \pi_{\text{obs}}(b|S') > 1$ and $\{a, b\} \subset S \cap S'$ where $a \neq b$. Then if π is an extension of π_{obs} it has to be the case that either $\pi(a|S) > \pi(a|S \cap S')$ or $\pi(b|S) > \pi(b|S \cap S')$. Hence, either aPc for some $c \in S - S'$ or bPd for some $d \in S' - S$. This is exactly the probabilistic analog of the condition in [De Clippel and Rozen \(2014\)](#). However, this condition is not enough in probabilistic domain. The next example illustrates this point.

Example A.2. Consider the following choice rule.

$\pi_{\text{obs}}(\cdot S)$	$S = \{a, b, c, d\}$	$\{a, b, c, e\}$	$\{a, b, d\}$	$\{a, c, d\}$	$\{b, c, e\}$
a	1/4	2/3	1/2	1/2	
b	1/4	1/6	1/2		5/6
c	1/4	0		1/2	1/12
d	1/4		0	0	
e		1/6			1/12

First, $\pi_{\text{obs}}(d|\{a, b, d\}) < \pi_{\text{obs}}(d|\{a, b, c, d\})$ and $\pi_{\text{obs}}(d|\{a, c, d\}) < \pi_{\text{obs}}(d|\{a, b, c, d\})$ imply that dPc and dPb . Furthermore, $\pi_{\text{obs}}(e|\{b, c, e\}) < \pi_{\text{obs}}(e|\{a, b, c, e\})$ implies that ePa . Now consider the set $\{a, b, c\}$ and

notice that $\pi_{\text{obs}}(a|\{a, b, c, e\}) + \pi_{\text{obs}}(b|\{a, b, c, d\}) + \pi_{\text{obs}}(c|\{a, b, c, d\}) > 1$. Thus if we had observations on the menu $\{a, b, c\}$ one of the following would have been true: (i) the probability that a is chosen decreases when e is removed from the menu $\{a, b, c, e\}$, (ii) the probability that b is chosen decreases when d is removed from the menu $\{a, b, c, d\}$, or (iii) the probability that c is chosen decreases when d is removed from the menu $\{a, b, c, d\}$. Hence, one of the following must be true: aPe , bPd , or cPd . Since we have a contradiction in all cases, π_{obs} is inconsistent with RAM. \perp

We generalize the intuition from this example. Suppose there exists a collection of pairs $(a_i, S_i)_{i=1}^m$ such that $\{a_1, \dots, a_m\} \subset \bigcap_{i=1}^m S_i$ and $\sum_{i=1}^m \pi_{\text{obs}}(a_i, S_i) > 1$ where a_i are all distinct. Now in the menu $\bigcap_{i=1}^m S_i$ the probability that a_i is chosen must decrease for at least one $i \in \{1, \dots, m\}$. From here we can conclude that $a_i Pb_i$ for some $b_i \in S_i - (S_1 \cap \dots \cap S_{i-1} \cap S_{i+1} \cap \dots \cap S_m)$ for some $i \in \{1, \dots, m\}$. Hence, the existence of an acyclic P that satisfies this condition is necessary for π_{obs} to be consistent with RAM. Theorem A.1 shows that it is also sufficient.

Theorem A.1. A choice rule π_{obs} is consistent with RAM if and only if there exists an acyclic binary relation P on X which satisfies the following: for any collection $(a_i, S_i)_{i=1}^m$ with distinct a_i such that $\{a_1, \dots, a_m\} \subset \bigcap_{i=1}^m S_i$ and $\sum_{i=1}^m \pi_{\text{obs}}(a_i, S_i) > 1$, $a_i Pb_i$ for some $b_i \in S_i - (S_1 \cap \dots \cap S_{i-1} \cap S_{i+1} \cap \dots \cap S_m)$ for some $i \in \{1, \dots, m\}$.

Proof. Let \succ be a transitive completion of P . We reorder the alternatives in X such that $a_{1,\succ} \succ \dots \succ a_{K,\succ}$. We define μ as follows. For any $S \in \mathcal{S}$,

$$\mu(T|S) = \begin{cases} \pi_{\text{obs}}(a_{k,\succ}|S) & \text{if } \exists k \text{ s.t. } T = L_{k,\succ} \cap S \\ 0 & \text{otherwise} \end{cases}$$

where $L_{k,\succ} = \{a_{k,\succ}, \dots, a_{K,\succ}\}$. For any $S \in \mathcal{X} - \mathcal{S}$, if there is $S' \in \mathcal{S}$ with $S' \supset S$, then

$$\mu(T|S) = \max_{S' \in \mathcal{S}: S' \supset S} \mu(T|S') \quad \text{if } T \subsetneq S$$

and $\mu(S|S) = 1 - \sum_{T \subsetneq S} \mu(T|S)$. Finally, for $S \in \mathcal{X} - \mathcal{S}$, if there is no $S' \in \mathcal{S}$ with $S' \supset S$, then $\mu(S|S) = 1$ and $\mu(T|S) = 0$ for all $T \subsetneq S$.

It is easily seen that (\succ, μ) represents π_{obs} . We first need to show that $\mu(\cdot|S)$ is a probability distribution. The only case we need to check is when $S \in \mathcal{X} - \mathcal{S}$ and there exists $S' \supset S$ with $S' \in \mathcal{S}$. We need to show that $\mu(S|S) \geq 0$ or that $\sum_{T \subsetneq S} \mu(T|S) \leq 1$. Suppose $\sum_{T \subsetneq S} \mu(T|S) > 1$. By definition, for each $T \subsetneq S$ such that $\mu(T|S) > 0$, there exists a pair (a_{k_T}, S_T) such that $S_T \in \mathcal{S}$ with $S_T \supset S$ and $\mu(T|S) = \mu(T|S_T) = \pi_{\text{obs}}(a_{k_T}|S_T)$ where $T = L_{a_{k_T},\succ} \cap S_T$. Then, $\sum_{T \subsetneq S} \pi_{\text{obs}}(a_{k_T}|S_T) > 1$. Notice that since μ is triangular, a_{k_T} are distinct. By definition of P , there exist $T \subsetneq S$ and an alternative b_{k_T} in $S_T - S$ such that $a_{k_T} Pb_{k_T}$. But this is a contradiction as $b_{k_T} \notin T$ and $T = L_{a_{k_T},\succ} \cap S_T$.

We now need to show that μ defined as above is monotonic. We have a few cases to consider.

Case 1: $S, S - b \in \mathcal{S}$. Suppose $\mu(T|S) > \mu(T|S - b)$ where $b \notin T$. Since $S \in \mathcal{S}$ and $\mu(T|S) > 0$ it must be that $T = L_{k,\succ} \cap S$ for some k and $\mu(T|S) = \pi_{\text{obs}}(a_{k,\succ}|S)$. Since $b \notin T$ and $T = L_{k,\succ} \cap S$ we must have $b \succ a_{k,\succ}$. Therefore, it must be the case that $T = L_{k,\succ} \cap (S - b)$. By definition, $\mu(T|S - b) = \pi_{\text{obs}}(a_{k,\succ}|S - b)$. But then we have $\pi_{\text{obs}}(a_{k,\succ}|S) > \pi_{\text{obs}}(a_{k,\succ}|S - b)$ which implies that $a_{k,\succ} Pb$, a contradiction.

Case 2: $S \in \mathcal{S}, S - b \notin \mathcal{S}$. Let T with $b \notin T$ given. Since $S \in \mathcal{S}$ it must be that either $\mu(T|S) = 0$ in which case monotonicity is trivial or $T = L_{k,\succ} \cap S$ for some k and $\mu(T|S) = \pi_{\text{obs}}(a_{k,\succ}|S)$. First, suppose $T \subsetneq S - b$. Now $S \supset S - b$, and hence by definition, $\mu(T|S - b) = \max_{S' \in \mathcal{S}: S' \supset S - b} \mu(T|S') \geq \mu(T|S)$. This establishes that the claim holds for all $T \subsetneq S - b$. Notice that if $\mu(T|S - b) = \mu(T|S)$ for all $T \subsetneq S - b$, then $\mu(S - b|S - b) \geq \mu(S - b|S)$ also follows.

Suppose $\mu(S - b|S) > \mu(S - b|S - b)$. Then since $\mu(S - b|S) > 0$ and μ is triangular we must have that b is \succ maximal in S . Thus $\mu(S|S) = \pi_{\text{obs}}(b|S)$. Furthermore, by the argument in the previous paragraph, there exists $T \subsetneq S - b$ such that $\mu(T|S - b) > \mu(T|S)$. Suppose there exists only one such T . A similar argument will work if there is more than one such T . By definition, there exists $S_T \in \mathcal{S}$ with $S_T \supset S - b$ such that $\mu(T|S_T) > \mu(T|S)$. Thus there exists a_{k_T} such that $\mu(T|S - b) = \mu(T|S_T) = \pi_{\text{obs}}(a_{k_T}|S_T)$ and $T = L_{a_{k_T},\succ} \cap S_T = L_{a_{k_T},\succ} \cap (S - b)$ where the second equality follows from the fact that $T \subset S - b$.

Notice that $a_{k_T} \in S - b$ and since b is \succ maximal in S we have $b \succ a_{k_T}$. Hence we have $T = L_{a_{k_T}, \succ} \cap S$ which by definition implies $\mu(T|S) = \pi_{\text{obs}}(a_{k_T}|S) < \pi_{\text{obs}}(a_{k_T}|S_T)$. Now $\mu(S - b|S) > \mu(S - b|S - b)$ implies that $\pi_{\text{obs}}(b|S) + \pi_{\text{obs}}(a_{k_T}|S) < \pi_{\text{obs}}(a_{k_T}|S_T)$. Consider the collection $\{(a_{k_T}, S_T)\} \cup \{(a_i, S) | a_i \in S \text{ and } a_i \neq b, a_i \neq a_{k_T}\}$. Since $\pi_{\text{obs}}(a_{k_T}|S_T) + (1 - \pi_{\text{obs}}(a_{k_T}|S) - \pi_{\text{obs}}(b|S)) > 1$, the observed choice probabilities summed over this collection adds up to greater than one. By definition of \mathbb{P} , either there exists an alternative in $a_i \in S - b$ such that $a_i \mathbb{P} b$ or there exists an alternative $c \in S_T - (S - b)$ such that $a_{k_T} \mathbb{P} c$. The first case leads to a contradiction since b is \succ maximal in S . The second case leads to a contradiction since $T = L_{a_{k_T}, \succ} \cap S_T = L_{a_{k_T}, \succ} \cap (S - b) \not\ni c$.

Case 3: $S \notin \mathcal{S}$, $S - b \in \mathcal{S}$. If there exists no $S' \supset S$ such that $S' \in \mathcal{S}$, then monotonicity property is trivial as $\mu(S|S) = 1$. Hence, suppose there is $S' \supset S$ such that $S' \in \mathcal{S}$ and there exists $T \subset S - b$ such that $\mu(T|S) > \mu(T|S - b)$. Let a_{k_T} and S_T be such that $S_T \supset S$ and $\mu(T|S) = \pi_{\text{obs}}(a_{k_T}|S_T)$ where $T = L_{a_{k_T}, \succ} \cap S_T$. Also notice that since $a_{k_T} \in T \subset S - b$, $a_{k_T} \neq b$. By definition of μ , it must be that $\pi_{\text{obs}}(a_{k_T}|S_T) > \pi_{\text{obs}}(a_{k_T}|S - b)$. Now consider the collection $\{(a_{k_T}, S_T)\} \cup \{(a_i, S - b) | a_i \in S - b, a_i \neq a_{k_T}\}$. If we add the choice probabilities over this collection, they will add up to greater than one. Hence, by definition of \mathbb{P} , there exists $c \in S_T - (S - b)$ such that $a_{k_T} \mathbb{P} c$. But this is a contradiction as $T = L_{a_{k_T}, \succ} \cap S_T$ and $T \subset S - b$.

Case 4: $S \notin \mathcal{S}$, $S - b \notin \mathcal{S}$. If there is no $S' \supset S$ such that $S \in \mathcal{S}$, then the claim is trivial. Suppose there exists such S' . Consider $T \subsetneq S - b$. By definition,

$$\mu(T|S) = \mu_{S' \supset S: S' \in \mathcal{S}}(T|S') \leq \mu_{S'' \supset S - b: S'' \in \mathcal{S}}(T|S'') = \mu(T|S - b).$$

Hence, the claim holds for all $T \subsetneq S - b$. We need to show that the claim also holds when $T = S - b$. Notice that if $\mu(T|S) = \mu(T|S - b)$ for all $T \subsetneq S - b$, then $\mu(S - b|S) \leq \mu(S - b|S - b)$ follows immediately. Hence suppose there is at least one T such that $\mu(T|S) < \mu(T|S - b)$.

Now if b is not \succ maximal in S , then $\mu(S - b|S) = 0$ and monotonicity is trivial. So suppose b is \succ maximal and $\mu(S - b|S) > \mu(S - b|S - b)$. For each $T \subsetneq S - b$ such that $\mu(T|S - b) > 0$, let a_{k_T} and S_T be such that $\mu(T|S - b) = \pi_{\text{obs}}(a_{k_T}|S_T)$. Let $a_{k_{S-b}}$ and S_{S-b} be such that $\mu(S - b|S) = \pi_{\text{obs}}(a_{k_{S-b}}|S_{S-b})$. Consider the collection $\{(a_{k_T}, S_T)\}$. Since $\mu(S - b|S) > \mu(S - b|S - b)$, if we sum choice probabilities over this collection, they add up to greater than one. This implies that $a_{k_T} \mathbb{P} c$ for some $T \subset S - b$ and for some $c \notin S - b$. But this is a contradiction since only sets of the form $T = L_{a_{k_T}, \succ} \cap S_T$ are considered with positive probability and $c \notin T$. ■

A.2.2 Estimation and Inference

In some cases not all choice problems $S \in \mathcal{X}$ may be observed for two reasons. First, some choice problems are ruled out *a priori* in the population due to, for instance, capacity or institutional constraints. Second, certain choice problems may not be available in a given finite sample due to sampling variability. Even if all choice problems are observed in the data, some may have only a few appearances, and usually are dropped from empirical analysis to avoid dealing with small (effective) sample sizes. We describe how our econometric methods (and assumptions) can also be adapted to situations of limited data.

Recall that $\mathcal{S} \subset \mathcal{X}$ denotes the collection of all observable choice problems. From an econometric perspective, this can also be seen as the collection of realized choice problems in the data. Assumption 2 now takes the following form.

Assumption 2' (DGP with Limited Data). The data is a random sample of choice problems Y_i and corresponding choices y_i , $\{(y_i, Y_i) : y_i \in Y_i, 1 \leq i \leq N\}$, generated by the underlying choice rule $\mathbb{P}[y_i = a | Y_i = S] = \pi(a|S)$, and that $\mathbb{P}[Y_i = S] \geq \underline{p} > 0$ for all $S \in \mathcal{S}$.

The most critical concern is on Assumption 1, which directly affects how the constraint matrix \mathbf{R}_\succ is constructed. We state a seemingly “stronger” version of that assumption.

Assumption 1' (Monotonic Attention with Limited Data). For any $A \subset S - T$, $\mu(T|S) \leq \mu(T|S - A)$.

When $\mathcal{S} = \mathcal{X}$, this assumption is equivalent to Assumption 1. To see this fact, pick an arbitrary $A \subset S - T$ and let a_1, a_2, \dots be an enumeration of elements of A . Then, $\mu(T|S) - \mu(T|S - A)$ can be written as

$\mu(T|S) - \mu(T|S - a_1) + \sum_{j \geq 1} [\mu(T|S - a_1 - \dots - a_j) - \mu(T|S - a_1 - \dots - a_j - a_{j+1})]$, where by Assumption 1 each summand is nonpositive, hence Assumption 1 implies Assumption 1'. The other direction is obvious: Assumption 1' implies Assumption 1.

If \mathcal{S} does not contain all possible choice problems, however, Assumption 1' provides a proper notion of monotonicity. To see the extra identification and statistical power implied by Assumption 1', two examples are provided after we give a proper definition of compatible preferences and the algorithm of constructing the constraint matrix in this context.

With limited data, unfortunately, there are two ways to generalize the previous definition, which are not equivalent. Recall that π_{obs} denotes the observed choice rule, defined only on \mathcal{S} , while we reserve the notation π to a choice rule that is defined on \mathcal{X} . Following is the first version.

Definition 6' (Compatible Preferences with Limited Data, I). Let π_{obs} be the underlying choice rule/data generating process. A preference \succ is compatible with π_{obs} on \mathcal{S} , denoted by $\succ \in \Theta_{\pi_{\text{obs}}}$, if $(\pi_{\text{obs}}, \succ)$ is a RAM on \mathcal{S} .

The other version is the following:

Definition 6'' (Compatible Preferences with Limited Data, II). Let π_{obs} be the underlying choice rule/data generating process. A preference \succ is compatible with π_{obs} on \mathcal{X} , denoted by $\succ \in \Theta_{\pi}$, if π_{obs} has an extension π to \mathcal{X} such that (π, \succ) is a RAM on \mathcal{X} .

When $\mathcal{S} = \mathcal{X}$, the two definitions agree, and also agree with Definition 6. If \mathcal{S} is a proper subset of \mathcal{X} , however, $\Theta_{\pi_{\text{obs}}}$ can be larger than Θ_{π} . Depending on the goal of analysis, both can be of interest, while Θ_{π} is much more difficult to characterize empirically.

The following algorithm generalizes the one given in the main text to the limited data scenario. The constraint matrix provided in this algorithm can be used to characterize $\Theta_{\pi_{\text{obs}}}$.

Algorithm 1' Construction of \mathbf{R}_{\succ} with Limited Data.

Ensure: Ordering of elements in π is set.

Require: Set a preference \succ .

$\mathbf{R}_{\succ} \leftarrow$ empty matrix

for S in \mathcal{S} **do**

for T in \mathcal{S} and $T \subset S$ **do**

for $a \prec S - T$ **do**

$\mathbf{R}_{\succ} \leftarrow$ add row corresponding to $\pi_{\text{obs}}(a|S) - \pi_{\text{obs}}(a|T) \leq 0$.

end for

end for

end for

Consider now two examples with limited data.

Example A.3 (Limited Data and Identification). Assume $\mathcal{S} = \{\{a, b, c, d\}, \{a, b\}\}$, and we are interested in the null hypothesis that $c \succ b \succ d \succ a$. Assumption 1 does not impose any constraint, since it only requires comparing choice problems that differ by one element. On the other hand, 1' implies the constraint that $\pi(a|\{a, b, c, d\}) \leq \pi(a|\{a, b\})$, since violating this constraint is equivalent to $a \succ c$ or $a \succ d$, which is incompatible with our hypothesis. \lrcorner

Example A.4 (Limited Data and Statistical Power). Consider $\mathcal{S} = \{\{a, b, c, d\}, \{a, b, c\}, \{a, b\}\}$. This is one example of limited data, but due to the special structure of \mathcal{S} , constraints implied by the two assumptions are equivalent *in population*. Consider the preference $a \succ c \succ d \succ b \succ$, then Assumption 1 gives two constraints, (i) $\pi(b|\{a, b, c, d\}) \leq \pi(b|\{a, b, c\})$ and (ii) $\pi(b|\{a, b, c\}) \leq \pi(b|\{a, b\})$. Assumption 1', however, will give the *extra* condition (iii) $\pi(b|\{a, b, c, d\}) \leq \pi(b|\{a, b\})$. In population, this extra constraint is redundant, since it is not possible to violate (iii) without violating at least one of (i) and (ii).

When applied to finite sample, however, the extra condition (iii) is no longer redundant, and may help to improve statistical power. To see this, assume (i) is violated by $\delta > 0$, a small margin, so is (ii). Then it is hard to reject any of them with finite sample. On the other hand, (iii) combines (i) and (ii), hence is violated by 2δ , which is much easier to detect in a finite sample. \lrcorner

Test statistics and critical values are constructed in a similar way based on $\mathbf{R}_{\succ} \hat{\pi}_{\text{obs}}$, hence not repeated here. We provide an analogous result of Theorem 5.

Theorem 5' (Validity of Critical Values with Limited Data I). Assume Assumption 2' holds.

1. Let Π_{obs} be a class of choice rules restricted to \mathcal{S} , and \succ a preference, such that: (i) for each $\pi_{\text{obs}} \in \Pi_{\text{obs}}$, $\succ \in \Theta_{\pi_{\text{obs}}}$; and (ii) $\inf_{\pi_{\text{obs}} \in \Pi_{\text{obs}}} \min(\sigma_{\pi_{\text{obs}}, \succ}) > 0$. Then

$$\limsup_{N \rightarrow \infty} \sup_{\pi_{\text{obs}} \in \Pi_{\text{obs}}} \mathbb{P}[\mathcal{J}(\succ) > c_{\alpha}(\succ)] \leq \alpha.$$

2. Let Π_{obs} be a class of choice rules restricted to \mathcal{S} and \mathcal{P} a class of preferences, such that: (i) for each $\pi_{\text{obs}} \in \Pi_{\text{obs}}$, $\mathcal{P} \subset \Theta_{\pi_{\text{obs}}}$; and (ii) $\inf_{\pi_{\text{obs}} \in \Pi_{\text{obs}}, \succ \in \mathcal{P}} \min(\sigma_{\pi_{\text{obs}}, \succ}) > 0$. Then

$$\limsup_{N \rightarrow \infty} \sup_{\pi_{\text{obs}} \in \Pi_{\text{obs}}} \mathbb{P}[\mathcal{J}(\mathcal{P}_{\forall}) > c_{\alpha}(\mathcal{P}_{\forall})] \leq \alpha.$$

3. Let Π_{obs} be a class of choice rules restricted to \mathcal{S} and \mathcal{P} a class of preferences, such that: (i) for each $\pi_{\text{obs}} \in \Pi_{\text{obs}}$, $\mathcal{P} \cap \Theta_{\pi_{\text{obs}}} \neq \emptyset$; and (ii) $\inf_{\pi_{\text{obs}} \in \Pi_{\text{obs}}, \succ \in \mathcal{P}} \min(\sigma_{\pi_{\text{obs}}, \succ}) > 0$. Then

$$\limsup_{N \rightarrow \infty} \sup_{\pi_{\text{obs}} \in \Pi_{\text{obs}}} \mathbb{P}[\mathcal{J}(\mathcal{P}_{\exists}) > c_{\alpha}(\mathcal{P}_{\exists})] \leq \alpha.$$

One natural question is whether it is possible to make any claim on Θ_{π} with limited data. Indeed, the three tests remain valid when applied to Θ_{π} (i.e. controls size uniformly), which can be easily seen from the fact that $\Theta_{\pi} \subset \Theta_{\pi_{\text{obs}}}$. We provide the following theorem. Also, for a choice rule π defined on \mathcal{X} , π_{obs} denotes its restriction to \mathcal{S} .

Theorem 5'' (Validity of Critical Values with Limited Data II). Assume Assumption 2' holds.

1. Let Π be a class of choice rules on \mathcal{X} , and \succ a preference, such that: (i) for each $\pi \in \Pi$, $\succ \in \Theta_{\pi}$; and (ii) $\inf_{\pi_{\text{obs}} \in \Pi_{\text{obs}}} \min(\sigma_{\pi_{\text{obs}}, \succ}) > 0$. Then

$$\limsup_{N \rightarrow \infty} \sup_{\pi \in \Pi} \mathbb{P}[\mathcal{J}(\succ) > c_{\alpha}(\succ)] \leq \alpha.$$

2. Let Π be a class of choice rules on \mathcal{X} and \mathcal{P} a class of preferences, such that: (i) for each $\pi \in \Pi$, $\mathcal{P} \subset \Theta_{\pi}$; and (ii) $\inf_{\pi_{\text{obs}} \in \Pi_{\text{obs}}, \succ \in \mathcal{P}} \min(\sigma_{\pi_{\text{obs}}, \succ}) > 0$. Then

$$\limsup_{N \rightarrow \infty} \sup_{\pi \in \Pi} \mathbb{P}[\mathcal{J}(\mathcal{P}_{\forall}) > c_{\alpha}(\mathcal{P}_{\forall})] \leq \alpha.$$

3. Let Π be a class of choice rules on \mathcal{X} and \mathcal{P} a class of preferences, such that: (i) for each $\pi \in \Pi$, $\mathcal{P} \cap \Theta_{\pi} \neq \emptyset$; and (ii) $\inf_{\pi_{\text{obs}} \in \Pi_{\text{obs}}, \succ \in \mathcal{P}} \min(\sigma_{\pi_{\text{obs}}, \succ}) > 0$. Then

$$\limsup_{N \rightarrow \infty} \sup_{\pi \in \Pi} \mathbb{P}[\mathcal{J}(\mathcal{P}_{\exists}) > c_{\alpha}(\mathcal{P}_{\exists})] \leq \alpha.$$

In Theorem 5'', we only need positive variance for moment conditions corresponding to the observed components of the choice rule. The reason is simple: In constructing the tests and critical values, we never use the unobserved part $\pi - \pi_{\text{obs}}$.

What is lost from Theorem 5' to Theorem 5''? With limited data, there may exist $\succ \in \Theta_{\pi_{\text{obs}}} - \Theta_{\pi}$ which is not rejected by the test statistic $\mathcal{J}(\succ)$ asymptotically.

A.3 Other Critical Values

There are many proposals for constructing critical values in the literature of testing moment inequalities (Canay and Shaikh, 2017; Ho and Rosen, 2017, and references therein). Here we discuss some of them for completeness. Throughout, let \mathbf{z}^* denote a random vector independent of the original data and $\mathbf{z}^* \sim \mathcal{N}(\mathbf{0}, \hat{\boldsymbol{\Omega}}/N)$. To save notation, define $(x)_+$ as the positive parts of x , and $(x)_-$ as the negative parts multiplied by -1 .

Plug-in Method

The first method simply plugs-in an estimate of $\mathbf{R}_{\succ} \boldsymbol{\pi}$ subject to the non-positivity constraint. Define

$$\mathcal{T}_{\text{PI}}^*(\succ) = \sqrt{N} \cdot \max \left((\mathbf{R}_{\succ} \mathbf{z}^* + (\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}})_-) \odot \hat{\boldsymbol{\sigma}}_{\succ} \right)_+,$$

then the critical value with level α with the plug-in method is

$$c_{\alpha, \text{PI}}(\succ) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{PI}}^*(\succ) \leq t] \geq 1 - \alpha \right\},$$

where \mathbb{P}^* denotes probability operator conditional on the data, and in practice, it is replaced by the simulated average $M^{-1} \sum_{m=1}^M \mathbf{1}(\cdot)$, where recall that $\mathcal{T}_{\text{PI}}^*(\succ)$ is simulated M times.

For the other two tests, critical values can be computed similarly. Define

$$\mathcal{T}_{\text{PI}}^*(\mathcal{P}_{\forall}) = \max_{\succ \in \mathcal{P}} \mathcal{T}_{\text{PI}}^*(\succ), \quad \mathcal{T}_{\text{PI}}^*(\mathcal{P}_{\exists}) = \min_{\succ \in \mathcal{P}} \mathcal{T}_{\text{PI}}^*(\succ),$$

then critical values are computed accordingly as the $1 - \alpha$ quantiles:

$$c_{\alpha, \text{PL}}(\mathcal{P}_{\forall}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{PL}}^*(\mathcal{P}_{\forall}) \leq t] \geq 1 - \alpha \right\}, \quad c_{\alpha, \text{PL}}(\mathcal{P}_{\exists}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{PL}}^*(\mathcal{P}_{\exists}) \leq t] \geq 1 - \alpha \right\}.$$

We note the critical values obtained here are *not* uniformly valid in the sense of Theorem 5.

Least Favorable Model

The critical values are nondecreasing in the centering, hence a conservative method is to consider the least favorable model, $\mathbf{R}_{\succ} \boldsymbol{\pi} = 0$, which assumes all the moment inequalities are binding. That is,

$$\mathcal{T}_{\text{LF}}^*(\succ) = \sqrt{N} \cdot \max \left((\mathbf{R}_{\succ} \mathbf{z}^*) \odot \hat{\boldsymbol{\sigma}}_{\succ} \right)_+,$$

and

$$c_{\alpha, \text{LF}}(\succ) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{LF}}^*(\succ) \leq t] \geq 1 - \alpha \right\}.$$

For the other two tests, critical values can be computed similarly. Define

$$\mathcal{T}_{\text{LF}}^*(\mathcal{P}_{\forall}) = \max_{\succ \in \mathcal{P}} \mathcal{T}_{\text{LF}}^*(\succ), \quad \mathcal{T}_{\text{LF}}^*(\mathcal{P}_{\exists}) = \min_{\succ \in \mathcal{P}} \mathcal{T}_{\text{LF}}^*(\succ),$$

then critical values are computed accordingly as the $1 - \alpha$ quantiles:

$$c_{\alpha, \text{LF}}(\mathcal{P}_{\forall}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{LF}}^*(\mathcal{P}_{\forall}) \leq t] \geq 1 - \alpha \right\}, \quad c_{\alpha, \text{LF}}(\mathcal{P}_{\exists}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{LF}}^*(\mathcal{P}_{\exists}) \leq t] \geq 1 - \alpha \right\}.$$

Undoubtedly, using such critical values may severely decrease the power of the test, especially when one or two moment restrictions are violated and the rest are far from binding. Next we introduce two methods seeking to improve power property of the test. The first one is based on moment selection, and the second one is based on replacing the unknown centering (moment inequalities) with an upper bound.

Generalized Moment Selection

Under the null hypothesis, a \sqrt{N} -scaled moment condition is either 0 or $-\infty$ asymptotically. Generalized moment selection exploits this observation. To describe, let $\lambda_N \geq 0$ and drift to ∞ slowly, and we represent row vectors (i.e. individual moment restrictions) in \mathbf{R}_{\succ} by $\{\mathbf{r}_{\succ,\ell} : 1 \leq \ell \leq L\}$. Then define

$$(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}})_{\text{GMS}} = \left[(\mathbf{r}'_{\succ,1} \hat{\boldsymbol{\pi}})_{\text{GMS}}, (\mathbf{r}'_{\succ,2} \hat{\boldsymbol{\pi}})_{\text{GMS}}, \dots, (\mathbf{r}'_{\succ,L} \hat{\boldsymbol{\pi}})_{\text{GMS}}, \right]',$$

with

$$(\mathbf{r}'_{\succ,\ell} \hat{\boldsymbol{\pi}})_{\text{GMS}} = \begin{cases} 0 & \text{if } \mathbf{r}'_{\succ,\ell} \hat{\boldsymbol{\pi}} / \hat{\sigma}_{\succ,\ell} \geq -\lambda_N, \\ -\infty & \text{if } \mathbf{r}'_{\succ,\ell} \hat{\boldsymbol{\pi}} / \hat{\sigma}_{\succ,\ell} < -\lambda_N. \end{cases}$$

In order to construct distributional approximation, $(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}})_{\text{GMS}}$ is used to replace the unknown population moment condition $\mathbf{R}_{\succ} \boldsymbol{\pi}$:

$$\mathcal{T}_{\text{GMS}}^*(\succ) = \sqrt{N} \cdot \max \left((\mathbf{R}_{\succ} \mathbf{z}^* + (\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}})_{\text{GMS}}) \odot \hat{\boldsymbol{\sigma}}_{\succ} \right)_+,$$

and

$$c_{\alpha, \text{GMS}}(\succ) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{GMS}}^*(\succ) \leq t] \geq 1 - \alpha \right\}.$$

For the other two tests, critical values can be computed similarly. Define

$$\mathcal{T}_{\text{GMS}}^*(\mathcal{P}_{\forall}) = \max_{\succ \in \mathcal{P}} \mathcal{T}_{\text{GMS}}^*(\succ), \quad \mathcal{T}_{\text{GMS}}^*(\mathcal{P}_{\exists}) = \min_{\succ \in \mathcal{P}} \mathcal{T}_{\text{GMS}}^*(\succ),$$

then critical values are computed accordingly as the $1 - \alpha$ quantiles:

$$c_{\alpha, \text{GMS}}(\mathcal{P}_{\forall}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{GMS}}^*(\mathcal{P}_{\forall}) \leq t] \geq 1 - \alpha \right\}, \quad c_{\alpha, \text{GMS}}(\mathcal{P}_{\exists}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{GMS}}^*(\mathcal{P}_{\exists}) \leq t] \geq 1 - \alpha \right\}.$$

The generalized moment selection method is similar to our construction, as both approaches “shrink” $\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}}$ to achieve uniformly valid inference compared to the plug-in method, while being less conservative than the least favorable method. The difference is that we propose to shrink the estimated moment conditions $\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}}$ continuously toward the least favorable case, while the generalized moment selection does the extreme: estimated moment conditions are replaced with their “asymptotic limits” based on a (increasing) threshold. See, for example, [Andrews and Soares \(2010\)](#), [Canay \(2010\)](#) and [Bugni \(2016\)](#) for more details.

Two-step Moment Selection

To illustrate this method, we first represent $\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}}$ by its individual components, as $\{\mathbf{r}'_{\succ,\ell} \hat{\boldsymbol{\pi}} : 1 \leq \ell \leq L\}$, where $\mathbf{r}'_{\succ,\ell}$ is the ℓ -th row of \mathbf{R}_{\succ} , so that there are in total L moment inequalities. The first step is to conduct moment selection. Let $0 < \beta < \alpha/3$, and the following set of indices of “active moment restrictions”:

$$\mathcal{L} = \left\{ \ell : \sqrt{N} \cdot \frac{\mathbf{r}'_{\succ,\ell} \hat{\boldsymbol{\pi}}}{\hat{\sigma}_{\succ,\ell}} \geq -2 \cdot c_{\beta, \text{LF}}(\succ) \right\},$$

then

$$\mathcal{T}_{\text{MS}}^*(\succ) = \sqrt{N} \cdot \max_{\ell \in \mathcal{L}} \left(\frac{\mathbf{r}'_{\succ,\ell} \mathbf{z}^*}{\hat{\sigma}_{\succ,\ell}} \right)_+,$$

and the critical value is computed as

$$c_{\alpha, \text{MS}}(\succ) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{MS}}^*(\succ) \leq t] \geq 1 - \alpha + 2\beta \right\}.$$

Constructing coupling statistics $\mathcal{T}^*(\mathcal{P}_{\forall})$ and $\mathcal{T}^*(\mathcal{P}_{\exists})$ with moment selection requires more delicate work. Note that simply taking maximum (or minimum) of individual $\mathcal{T}_{\text{MS}}^*(\succ)$ does not work, mainly due to the

two-step nature of the construction. More precisely, the first step moment selection controls error probability to be β for each individual preference, but not jointly, and the reason is that we used $c_{\beta,\text{LF}}(\succ)$ for moment selection, which is too small jointly for a collection of preferences.

For a collection of preferences, the correct moment selection is the following:

$$\mathcal{L}_{\succ,\mathcal{P}} = \left\{ \ell : \sqrt{N} \cdot \frac{\mathbf{r}'_{\succ,\ell} \hat{\boldsymbol{\pi}}}{\hat{\sigma}_{\succ,\ell}} \geq -2 \cdot c_{\beta,\text{LF}}(\mathcal{P}_{\forall}) \right\}, \quad \text{for each } \succ \in \mathcal{P}.$$

Now we use a much larger critical value for moment selection: $c_{\beta,\text{LF}}(\mathcal{P}_{\forall})$ instead of $c_{\beta,\text{LF}}(\succ)$. Then the coupling statistics are

$$\mathcal{T}_{\text{MS}}^*(\mathcal{P}_{\forall}) = \sqrt{N} \cdot \max_{\succ \in \mathcal{P}} \max_{\ell \in \mathcal{L}_{\succ,\mathcal{P}}} \left(\frac{\mathbf{r}'_{\succ,\ell} \mathbf{z}^*}{\hat{\sigma}_{\succ,\ell}} \right)_+, \quad \mathcal{T}_{\text{MS}}^*(\mathcal{P}_{\exists}) = \sqrt{N} \cdot \min_{\succ \in \mathcal{P}} \max_{\ell \in \mathcal{L}_{\succ,\mathcal{P}}} \left(\frac{\mathbf{r}'_{\succ,\ell} \mathbf{z}^*}{\hat{\sigma}_{\succ,\ell}} \right)_+,$$

then critical values are computed accordingly as the $1 - \alpha$ quantiles:

$$c_{\alpha,\text{MS}}(\mathcal{P}_{\forall}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{MS}}^*(\mathcal{P}_{\forall}) \leq t] \geq 1 - \alpha + 2\beta \right\},$$

$$c_{\alpha,\text{MS}}(\mathcal{P}_{\exists}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{MS}}^*(\mathcal{P}_{\exists}) \leq t] \geq 1 - \alpha + 2\beta \right\}.$$

See, for example, [Chernozhukov et al. \(2014\)](#) and references therein.

Two-step Moment Upper Bounding

This method uses a first step to construct a confidence region for $\mathbf{R}_{\succ} \boldsymbol{\pi}$, and the upper bound of such region is used as a conservative estimate for $\mathbf{R}_{\succ} \boldsymbol{\pi}$. Let $0 < \beta < \alpha$, and

$$\mathcal{T}_{\text{UB}}^*(\succ) = \sqrt{N} \cdot \max \left((\mathbf{R}_{\succ} \mathbf{z}^* + (\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} + c_{\beta,\text{LF}}(\succ) \hat{\boldsymbol{\sigma}}_{\succ} / \sqrt{N})_-) \odot \hat{\boldsymbol{\sigma}}_{\succ} \right)_+,$$

then

$$c_{\alpha,\text{UB}}(\succ) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{UB}}^*(\succ) \leq t] \geq 1 - \alpha + \beta \right\}.$$

Note that in the first step, we use the critical values from the least favorable method to construct upper bounds on the moment inequalities $\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} + c_{\beta,\text{LF}}(\succ) \hat{\boldsymbol{\sigma}}_{\succ} / \sqrt{N}$, which is guaranteed to have coverage $1 - \beta$. Then the significance level in the second step is adjusted to account for errors incurred in the first step. See, e.g., [Romano, Shaikh, and Wolf \(2014\)](#) for further details. In particular, they recommend to use $\beta/\alpha = 0.1$.

Same as the moment selection method, the upper bound (or confidence region for $\mathbf{R}_{\succ} \boldsymbol{\pi}$) constructed above only controls error probability for individual preference, but not jointly for a collection of preferences. Hence we need to make further adjustments. The solution is almost the same: replace the critical value used for constructing upper bounds by one that controls error probability jointly for the collection \mathcal{P} :

$$\mathcal{T}_{\text{UB}}^*(\mathcal{P}_{\forall}) = \sqrt{N} \cdot \max_{\succ \in \mathcal{P}} \max \left((\mathbf{R}_{\succ} \mathbf{z}^* + (\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} + c_{\beta,\text{LF}}(\mathcal{P}_{\forall}) \hat{\boldsymbol{\sigma}}_{\succ} / \sqrt{N})_-) \odot \hat{\boldsymbol{\sigma}}_{\succ} \right)_+,$$

$$\mathcal{T}_{\text{UB}}^*(\mathcal{P}_{\exists}) = \sqrt{N} \cdot \min_{\succ \in \mathcal{P}} \max \left((\mathbf{R}_{\succ} \mathbf{z}^* + (\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} + c_{\beta,\text{LF}}(\mathcal{P}_{\forall}) \hat{\boldsymbol{\sigma}}_{\succ} / \sqrt{N})_-) \odot \hat{\boldsymbol{\sigma}}_{\succ} \right)_+,$$

then critical values are computed accordingly as the $1 - \alpha$ quantiles:

$$c_{\alpha,\text{UB}}(\mathcal{P}_{\forall}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{UB}}^*(\mathcal{P}_{\forall}) \leq t] \geq 1 - \alpha + \beta \right\},$$

$$c_{\alpha,\text{UB}}(\mathcal{P}_{\exists}) = \inf \left\{ t : \mathbb{P}^* [\mathcal{T}_{\text{UB}}^*(\mathcal{P}_{\exists}) \leq t] \geq 1 - \alpha + \beta \right\}.$$

We would emphasize that for both of the coupling statistics, $\mathcal{T}_{\text{UB}}^*(\mathcal{P}_{\forall})$ and $\mathcal{T}_{\text{UB}}^*(\mathcal{P}_{\exists})$, the same first step critical value $c_{\beta,\text{LF}}(\mathcal{P}_{\forall})$ is used.

B Appendix B: Omitted Proofs

This appendix collects the proofs omitted from the main text to improve the exposition.

B.1 Proof of Theorem 3

We show that the set of extreme points of $\mathcal{MT}(\succ)$ is $\mathcal{AF}(\succ)$. Clearly, any $\Gamma \in \mathcal{AF}(\succ)$ is an extreme point. Pick a non-deterministic attention rule $\mu \in \mathcal{MT}(\succ)$. We show that μ cannot be an extreme point. Let $\mathcal{X}_\mu \subset \mathcal{X}$ stand for all sets for which $\mu(T|S) = 1$ for no $T \subset S$. We start by choosing $\varepsilon > 0$ small enough so that none of the non-binding constraints are affected whenever ε is added to or subtracted from $\mu(T|S)$ for all $T \subset S$ and $S \in \mathcal{X}$. Let $k_\mu = \min_{S \in \mathcal{X}_\mu} |S|$. Since μ is not deterministic, such k_μ exists.

We begin with the following simple observation that given S with $|S| = k_\mu$ we can have at most two subsets of S with $\mu(T|S) \in (0, 1)$. Moreover, it must be the case that $\mu(S|S) \in (0, 1)$.

Lemma A.1. Let S with $|S| = k_\mu$ be given. Then there exist at most two $T \subset S$ such that $\mu(T|S) \in (0, 1)$. Furthermore, $\mu(S|S) \in (0, 1)$.

Proof. Suppose there exist three such subsets: T_1, T_2 , and T_3 . Since μ is triangular the subsets which are considered with positive probability can be ordered by set inclusion. Hence, we can without loss of generality assume $T_1 \subset T_2 \subset T_3$. But then since μ is monotonic and $T_1 \subset T_2 \subset S$ it must be that $\mu(T_1|T_2) \in (0, 1)$ and $\mu(T_2|T_2) \in (0, 1)$. This contradicts the definition of k_μ . The same contradiction appears as long as $T_2 \subsetneq S$. Hence, $T_2 = S$. ■

Now for all sets $S \in \mathcal{X}_\mu$ with $|S| = k_\mu$, we define μ' and μ'' for such S as follows:

$$\begin{aligned}\mu'(T|S) &= \mu(T|S) + \varepsilon, \\ \mu'(S|S) &= \mu(S|S) - \varepsilon,\end{aligned}$$

and

$$\begin{aligned}\mu''(T|S) &= \mu(T|S) - \varepsilon, \\ \mu''(S|S) &= \mu(S|S) + \varepsilon\end{aligned}$$

where $T \subsetneq S$ with $\mu(T|S) \in (0, 1)$.

Suppose we have defined μ' and μ'' for all sets with $|S| \leq l$ and let S with $|S| = l + 1$ be given. If there exist no $T \subset S$ and $S_T \subset S$ such that $\mu'(T|S_T) \neq \mu''(T|S_T)$ and $\mu(T|S) = \mu(T|S_T)$, then we set $\mu(T|S) = \mu'(T|S) = \mu''(T|S)$ for all $T \subset S$. Otherwise, pick the smallest T for which such S_T exists. If $\mu'(T|S_T) > \mu''(T|S_T)$, then let $\mu'(T|S) = \mu(T|S) + \varepsilon$ and $\mu''(T|S) = \mu(T|S) - \varepsilon$ and if $\mu'(T|S_T) < \mu''(T|S_T)$, then let $\mu'(T|S) = \mu(T|S) - \varepsilon$ and $\mu''(T|S) = \mu(T|S) + \varepsilon$. If T is the only set for which such S_T exists, then let T' be the largest set for which $\mu(T'|S) \in (0, 1)$. Otherwise T' denotes the other set for which $S_{T'}$ satisfying the description exists. If $\mu'(T|S_T) > \mu''(T|S_T)$, then let $\mu'(T'|S) = \mu(T'|S) - \varepsilon$ and $\mu''(T'|S) = \mu(T'|S) + \varepsilon$ and if $\mu'(T|S_T) < \mu''(T|S_T)$, then let $\mu'(T'|S) = \mu(T'|S) + \varepsilon$ and $\mu''(T'|S) = \mu(T'|S) - \varepsilon$. For all other subsets μ, μ' , and μ'' agree. We proceed iteratively.

Lemma A.2. Suppose there exist $T \subset S$ and $S_T \subset S$ such that $\mu'(T|S_T) \neq \mu''(T|S_T)$ and $\mu(T|S) = \mu(T|S_T)$. Then either T is a singleton or we can set $S_T = T$.

Proof. The claim follows from Claim 1 when $|S| = k_\mu + 1$. Suppose the claim holds whenever $|S| \leq l$. We show that the claim holds when $|S| = l + 1$. Let $T \subset S$ and $S_T \subset S$ satisfy the description and suppose T is not a singleton. Since $\mu'(T|S_T) \neq \mu''(T|S_T)$, by construction, either T is the largest set satisfying $\mu(T|S_T) \in (0, 1)$ or there exists $S_{S_T} \subset S_T$ such that $\mu'(T|S_{S_T}) \neq \mu''(T|S_{S_T})$ and $\mu(T|S_T) = \mu(T|S_{S_T})$. If the first case is true, then since μ is monotonic, it must be the case that $\mu(T'|T) = \mu(T'|S_T)$ for all $T' \subset T$, and hence we are done. In the second case, the claim follows from induction. ■

Lemma A.3. For any S and non-singleton $T \subset S$, $\mu'(T|S) > \mu''(T|S)$ if and only if $\mu(T|S) = \mu(T|T)$ and $\mu'(T|T) > \mu''(T|T)$.

Proof. The “if” part of the claim follows from the construction. The “only if” part follows from the previous claim and the construction. ■

Lemma A.4. For any S , there exist either zero or two subsets satisfying $\mu'(T|S) \neq \mu''(T|S)$. Moreover if there are two sets satisfying the description, then $\mu'(T_1|S) > \mu''(T_1|S)$ if and only if $\mu'(T_2|S) < \mu''(T_2|S)$.

Proof. This first part of the claim is trivial when $|S| = k_\mu$. Suppose the claim is true for all S with $|S| \leq l$ and let S with $|S| = l + 1$ be given. If there is no T which satisfies the description in the construction, then no subset will be affected. Suppose there exists only one such T . We show that there exists $T' \supset T$ such that $\mu(T'|S) \in (0, 1)$. To see this notice that by monotonicity property $\mu(T''|S) \leq \mu(T''|S_T)$ for all $T'' \subset T$. Since by induction there are two subsets of S_T for which $\mu'(T|S_T) \neq \mu''(T|S_T)$ either $\mu(T''|S) < \mu(T''|S_T)$ for some $T'' \subset T$ or there exists $T''' \supset T$ such that $\mu(T'''|S_T) \in (0, 1)$. In both cases, $\sum_{T'' \subset T} \mu(T''|S) < 1$ follows. Hence, there is $T' \supset T$ such that $\mu(T'|S) \in (0, 1)$. The construction then guarantees that $\mu'(T'|S) \neq \mu''(T'|S)$ for some $T' \supset T$. Now suppose there are three subsets, T_1, T_2 , and T_3 , satisfying the description. Since μ is triangular, we can without loss of generality assume that $T_1 \subset T_2 \subset T_3$. By the previous claim, we can without loss of generality assume $S_{T_2} = T_2$ and $S_{T_3} = T_3$ (T_1 may be a singleton). But then since μ is monotonic, 3 subsets of S_{T_3} must satisfy the description, a contradiction to induction hypothesis.

To prove the second part of the claim, notice that the claim is follows from construction if $|S| = k_\mu$. Suppose the claim holds whenever $|S| \leq l$ and let $|S| = l + 1$ be given. If $T_2 = S$, then the claim follows from construction. If $T_2 \subsetneq S$, then the claim follows from induction and construction by considering the set T_2 . ■

It is clear that $\mu = \frac{1}{2}\mu' + \frac{1}{2}\mu''$. The previous lemmas also show that both μ' and μ'' are monotonic. Hence, no $\mu \in \mathcal{MT}(\succ) - \mathcal{AF}(\succ)$ can be an extreme point. This concludes the proof of Theorem 3.

B.2 Proof of Theorem 5

For completeness and notational clarity, we first present two lemmas, which follow from the central limit theorem.

Lemma A.5. Under Assumption 2,

$$\sqrt{N_S}(\hat{\pi}_S - \pi_S) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_{\pi,S}), \quad \mathbf{\Omega}_{\pi,S} = \text{diag}(\pi_S) - \pi_S \pi_S',$$

where $N_S = \sum_{1 \leq i \leq N} \mathbf{1}(Y_i = S)$ is the effective sample size of menu S , $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}_{\pi,S})$ is the $|S|$ -dimensional Gaussian distribution with mean zero and covariance $\mathbf{\Omega}_{\pi,S}$, and diag is the operator constructing diagonal matrices.

The asymptotic distribution is degenerate for two reasons. First, the choice rule, by construction, has to sum up to 1. Second, it is possible that some of the alternatives in S is never chosen (either in the sample or the population).

Lemma A.6. Under Assumption 2,

$$\sqrt{N}(\hat{\pi} - \pi) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_\pi),$$

where the asymptotic variance $\mathbf{\Omega}_\pi$ is block diagonal, with blocks given by $\frac{1}{\mathbb{P}[Y_i=S]} \mathbf{\Omega}_{\pi,S}$, $S \in \mathcal{X}$. (Block diagonality of the asymptotic variance follows directly from the fact that across choice problems, choice rules are estimated with independent samples, hence are independent.)

Recall that $\mathcal{F}(\succ) = \sqrt{N} \max((\mathbf{R}_\succ \hat{\pi}) \odot \hat{\sigma}_\succ)_+$. The next proposition shows that its distribution is approximated by the infeasible statistic $\sqrt{N} \max(\mathbf{R}_\succ(\mathbf{z}^* + \pi) \odot \hat{\sigma}_\succ)_+$. For clarity, we do not seek uniformity here. Later we will show how the same argument can be used to demonstrate distributional approximation uniformly in a class of DGPs (see Proposition A.2).

Proposition A.1. Suppose Assumption 2 holds and $\min(\sigma_{\pi, \succ}) > 0$. Then

$$\left| \mathbb{P} \left[\mathcal{F}(\succ) \leq t \right] - \mathbb{P}^* \left[\sqrt{N} \max(\mathbf{R}_\succ(\mathbf{z}^* + \boldsymbol{\pi}) \odot \hat{\boldsymbol{\sigma}}_\succ)_+ \leq t \right] \right| \rightarrow_{\mathbb{P}} 0, \quad \forall t \neq 0.$$

Remark A.1. We exclude $t = 0$ since the limiting distribution may have a point mass at the origin. \square

Proof. Let f be a bounded Lipschitz function. Without loss of generality assume its Lipschitz constant is 1, and $2 \cdot \|f\|_\infty = c$. For convenience, denote $\mathbb{E}[f(\cdot)]$ by $\mathbb{E}_f[\cdot]$.

By the central limit theorem, it is possible to construct a (sequence of) random vector $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\pi/N)$ such that $|\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) - \sqrt{N}\tilde{\mathbf{z}}| = O_{\mathbb{P}}(1/\sqrt{N})$ (we postpone the proof to the end). Further, let $\mathbf{w}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\sqrt{N})$ with suitable dimension such that $\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* \sim \mathbf{z}^*$. Then consider bounds on the following quantities.

$$\begin{aligned} & \left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \hat{\boldsymbol{\sigma}}_\succ \right)_+ \right] - \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| \\ & \leq \mathbb{E} \left[\left| \sqrt{N} \max \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \hat{\boldsymbol{\sigma}}_\succ \right)_+ - \sqrt{N} \max \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right| \wedge c \right] \\ & \leq \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \hat{\boldsymbol{\sigma}}_\succ \right)_+ - \sqrt{N} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right\|_\infty \wedge c \right] \\ & = \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \hat{\boldsymbol{\pi}} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \geq 0 \right) \right\|_\infty \wedge c \right] \\ & \leq \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right\|_\infty \wedge c \right] \\ & \quad + \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \boldsymbol{\pi} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \boldsymbol{\pi} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \geq 0 \right) \right\|_\infty \wedge c \right]. \end{aligned}$$

The first inequality uses Lipschitz property of f and the fact that the whole term is bounded by $2 \cdot \|f\|_\infty = c$. The second inequality uses basic property of the max operator. The third inequality follows from triangle inequality of the norm $\|\cdot\|_\infty$.

We further split in order to control the denominator:

$$\begin{aligned} & \text{previous display} \\ & \leq \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right\|_\infty \mathbf{1} \left(\min(\hat{\boldsymbol{\sigma}}_\succ) \geq \min(\boldsymbol{\sigma}_{\pi, \succ})/2 \right) \wedge c \right] \\ & \quad + \mathbb{P} \left[\min(\hat{\boldsymbol{\sigma}}_\succ) < \min(\boldsymbol{\sigma}_{\pi, \succ})/2 \right] \cdot c \\ & \quad + \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \boldsymbol{\pi} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \boldsymbol{\pi} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \geq 0 \right) \right\|_\infty \wedge c \right] \\ & = O \left(\frac{1}{\sqrt{N}} \right) + \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \boldsymbol{\pi} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \boldsymbol{\pi} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \geq 0 \right) \right\|_\infty \wedge c \right] \\ & \leq O \left(\frac{1}{\sqrt{N}} \right) + \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \boldsymbol{\pi} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \boldsymbol{\pi} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \boldsymbol{\pi} \geq -\frac{a_n}{\sqrt{N}} \right) \right\|_\infty \wedge c \right] \\ & \quad + \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \boldsymbol{\pi} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \boldsymbol{\pi} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \geq 0, \mathbf{R}_\succ \boldsymbol{\pi} < -\frac{a_n}{\sqrt{N}} \right) \right\|_\infty \wedge c \right] \\ & \leq O \left(\frac{1}{\sqrt{N}} \right) + \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_\succ \boldsymbol{\pi} \odot \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ \boldsymbol{\pi} \odot \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_\succ \boldsymbol{\pi} \geq -\frac{a_n}{\sqrt{N}} \right) \right\|_\infty \wedge c \right] \\ & \quad + \mathbb{E} \left[\left\| \mathbf{1} \left(\mathbf{R}_\succ \hat{\boldsymbol{\pi}} \geq 0, \mathbf{R}_\succ \boldsymbol{\pi} < -\frac{a_n}{\sqrt{N}} \right) \right\|_\infty \wedge c \right] \\ & = O \left(\frac{1}{\sqrt{N}} + \frac{a_n}{\sqrt{N}} + \frac{1}{a_n} \right), \end{aligned}$$

with $a_n \rightarrow \infty$ chosen so that the last line vanishes.

$$\begin{aligned}
& \left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\tilde{\mathbf{z}} + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| \\
& \leq \mathbb{E} \left[\left\| \sqrt{N} \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ - \sqrt{N} \left(\mathbf{R}_{\succ} (\tilde{\mathbf{z}} + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right\|_{\infty} \wedge c \right] \\
& \leq \mathbb{E} \left[\sqrt{N} \left\| \mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} (\tilde{\mathbf{z}} + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right\|_{\infty} \wedge c \right] \\
& = \mathbb{E} \left[\sqrt{N} \left\| \mathbf{R}_{\succ} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} \tilde{\mathbf{z}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right\|_{\infty} \wedge c \right] \\
& = O_{\mathbb{P}} \left(\frac{1}{\sqrt{N}} \right).
\end{aligned}$$

The first inequality uses Lipschitz continuity of f and property of the max operator. The second inequality drops positivity. The rate in the last line comes from the coupling requirement of $\tilde{\mathbf{z}}$.

$$\left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\tilde{\mathbf{z}} + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\boldsymbol{\Omega}_{\pi}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| = 0,$$

since the two terms have the same distribution.

$$\begin{aligned}
& \left| \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\boldsymbol{\Omega}_{\pi}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| \\
& \leq \mathbb{E}^* \left[\left\| \sqrt{N} \left(\mathbf{R}_{\succ} (\boldsymbol{\Omega}_{\pi}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ - \sqrt{N} \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right\|_{\infty} \wedge c \right] \\
& \leq \mathbb{E}^* \left[\sqrt{N} \left\| \mathbf{R}_{\succ} (\boldsymbol{\Omega}_{\pi}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right\|_{\infty} \wedge c \right] \\
& = \mathbb{E}^* \left[\sqrt{N} \left\| \mathbf{R}_{\succ} \boldsymbol{\Omega}_{\pi}^{1/2} \mathbf{w}^* \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} \hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* \otimes \boldsymbol{\sigma}_{\pi, \succ} \right\|_{\infty} \wedge c \right] \\
& = O_{\mathbb{P}} \left(\frac{1}{\sqrt{N}} \right).
\end{aligned}$$

The first inequality uses Lipschitz continuity of f and property of the max operator. The second inequality drops positivity. The rate in the last line comes from the fact $\sqrt{N} \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_{\pi}\|_{\infty} = O_{\mathbb{P}}(1)$. Also, by construction $\mathbf{w}^* = O_{\mathbb{P}^*}(1/\sqrt{N})$.

$$\begin{aligned}
& \left| \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \right] \right| \\
& \leq \mathbb{E}^* \left[\left\| \sqrt{N} \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ - \sqrt{N} \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \right\|_{\infty} \wedge c \right] \\
& = \mathbb{E}^* \left[\left\| \sqrt{N} \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \geq 0 \right) \right\|_{\infty} \wedge c \right] \\
& \leq \mathbb{E}^* \left[\left\| \sqrt{N} \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} \hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \right\|_{\infty} \wedge c \right] \\
& \quad + \mathbb{E}^* \left[\left\| \sqrt{N} \left(\mathbf{R}_{\succ} \boldsymbol{\pi} \otimes \boldsymbol{\sigma}_{\pi, \succ} - \mathbf{R}_{\succ} \boldsymbol{\pi} \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \odot \mathbf{1} \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \geq 0 \right) \right\|_{\infty} \wedge c \right] \\
& = O_{\mathbb{P}} \left(\frac{1}{\sqrt{N}} + \frac{a_n}{\sqrt{N}} + \frac{1}{a_n} \right).
\end{aligned}$$

The first inequality uses Lipschitz continuity of f and property of the max operator. The second inequality applies triangle inequality to the norm $\|\cdot\|_{\infty}$. The rest is essentially the same as that of the first part.

The only missing part is to show the existence of the coupling variable $\tilde{\mathbf{z}}$. Since the choice probabilities are averages of indicators, Corollary 4.1 of [Chen, Goldstein, and Shao \(2010\)](#) implies the following non-

asymptotic bound on the Wasserstein metric:

$$\inf \left\{ \mathbb{E} [|X - Y|] : X \sim \sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}), Y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\pi) \right\} \leq \text{Const.} \sqrt{\frac{1}{\min_{S \in \mathcal{X}} N_S}},$$

where the infimum is taken over all joint distributions with the given marginals, and the constant in the above display is universal. By Assumption 2, the rate on the RHS is proportional to $\sqrt{1/N}$. Existence of the coupling variable follows from the bounds on the Wasserstein metric.

Hence we showed that

$$\left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\hat{\boldsymbol{\Omega}}^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \hat{\boldsymbol{\sigma}}_{\succ} \right)_+ \right] \right| = o_{\mathbb{P}}(1).$$

■

Now we show how the previous result can be generalized to be uniform among a class of distributions. The main argument used in Proposition A.1 remains almost unchanged. Again we focus on the test statistic $\mathcal{T}(\succ)$.

Proposition A.2. Under the assumptions of part 1 in Theorem 5,

$$\sup_{\pi \in \Pi} \mathbb{P} \left[\left| \mathbb{P} \left[\mathcal{T}(\succ) \leq t \right] - \mathbb{P}^* \left[\sqrt{N} \max(\mathbf{R}_{\succ}(\mathbf{z}^* + \boldsymbol{\pi}) \otimes \hat{\boldsymbol{\sigma}}_{\succ})_+ \leq t \right] \right| > \varepsilon \right] \rightarrow 0, \quad \forall \varepsilon > 0, \forall t \neq 0.$$

Proof. The proof remains almost the same, while extra care is needed for the coupling argument, which we demonstrate below. We would like to bound the following quantity *uniformly*:

$$\left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\boldsymbol{\Omega}_\pi^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right|,$$

where f is a bounded function with Lipschitz constant 1. By the coupling argument in the proof of Proposition A.1, it is possible to construct, for each $\pi \in \Pi$, a random variable $\tilde{\mathbf{z}}_\pi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\pi/N)$ (the covariance matrix $\boldsymbol{\Omega}_\pi$ depends on π , indicated by the subscript), such that

$$\mathbb{E} \left[\left| \sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) - \sqrt{N}\tilde{\mathbf{z}}_\pi \right| \right] \leq \text{Const.} \sqrt{\frac{1}{\min_{S \in \mathcal{X}} N_S}}.$$

Then we can bound the aforementioned quantity by

$$\begin{aligned} & \left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\boldsymbol{\Omega}_\pi^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| \\ & \leq \left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\tilde{\mathbf{z}}_\pi + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| \\ & \quad + \left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\tilde{\mathbf{z}}_\pi + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f^* \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\boldsymbol{\Omega}_\pi^{1/2} \mathbf{w}^* + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right| \\ & = \left| \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}} \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] - \mathbb{E}_f \left[\sqrt{N} \max \left(\mathbf{R}_{\succ} (\tilde{\mathbf{z}}_\pi + \boldsymbol{\pi}) \otimes \boldsymbol{\sigma}_{\pi, \succ} \right)_+ \right] \right|, \end{aligned}$$

where the second line comes from triangle inequality, and the equality uses the fact that $\tilde{\mathbf{z}}_\pi$ and $\boldsymbol{\Omega}_\pi^{1/2} \mathbf{w}^*$ have the same distribution. Hence we have the bound:

$$\text{previous display} \lesssim \sqrt{\frac{1}{\min_{S \in \mathcal{X}} N_S}},$$

with the estimate in the above display being uniformly valid over $\pi \in \Pi$. Hence the proof reduces to bound

the probability (choose some $\varepsilon > 0$ small enough):

$$\begin{aligned}
& \sup_{\pi \in \Pi} \mathbb{P} \left[\min_{S \in \mathcal{X}} N_S < \varepsilon N \right] = \sup_{\pi \in \Pi} \mathbb{P} \left[\frac{N_S}{N} < \varepsilon, \exists S \right] \leq \sup_{\pi \in \Pi} \sum_{S \in \mathcal{X}} \mathbb{P} \left[\frac{N_S}{N} < \varepsilon \right] \\
& = \sup_{\pi \in \Pi} \sum_{S \in \mathcal{X}} \mathbb{P} \left[\frac{N_S - \mathbb{P}[Y_i = S]N}{N} < \varepsilon - \mathbb{P}[Y_i = S] \right] \\
& \leq \sup_{\pi \in \Pi} \sum_{S \in \mathcal{X}} \exp \left(-(\varepsilon - \mathbb{P}[Y_i = S])^2 \frac{N}{2} \right) \rightarrow 0.
\end{aligned}$$

For the inequality in the first line, we use union bound. The last line uses Hoeffding’s inequality, which is valid for any ε smaller than $\inf_{\pi \in \Pi} \min_{S \in \mathcal{X}} \mathbb{P}[Y_i = S] \geq \underline{p}$ (see Assumption 2). Hence we demonstrated that

$$\sqrt{\frac{1}{\min_{S \in \mathcal{X}} N_S}} \underset{\mathbb{P}}{\asymp} \sqrt{\frac{1}{N}}, \quad \text{uniformly in } \Pi.$$

■

Now we demonstrate how Theorem 5 follows from the previous proposition. We focus on the test statistic $\mathcal{T}(\succ)$, but the same argument applies to the other two tests, with straightforward modification.

Recall that in constructing $\mathcal{T}^*(\succ)$, we use $\kappa_N(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}})_-$ to replace the unknown $\mathbf{R}_{\succ} \boldsymbol{\pi}$. This unknown quantity is bounded above by 0 uniformly in Π , since this class only contains choice rules that are compatible with \succ . At the same time, $\kappa_N(\mathbf{R}_{\succ} \hat{\boldsymbol{\pi}})_-$ converges to 0 in probability uniformly for the class Π . Therefore asymptotically $\mathcal{T}^*(\succ)$ stochastically dominates $\mathcal{T}(\succ)$ uniformly in Π , which proves Theorem 5.

References

- AGUIAR, V. H. (2015): “Stochastic Choice and Attention Capacities: Inferring Preferences from Psychological Biases,” *SSRN Working Paper 2607602*.
- (2017): “Random Categorization and Bounded Rationality,” *Economics Letters*, 159, 46–52.
- AGUIAR, V. H., M. J. BOCCARDI, AND M. DEAN (2016): “Satisficing and Stochastic Choice,” *Journal of Economic Theory*, 166, 445–482.
- AHUMADA, A. AND L. ULKU (2017): “Luce Rule with Limited Consideration,” *SSRN Working Paper 2939901*.
- ANDREWS, D. W. AND G. SOARES (2010): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, 78, 119–157.
- BARBERÀ, S. AND B. GRODAL (2011): “Preference for Flexibility and the Opportunities of Choice,” *Journal of Mathematical Economics*, 47, 272–278.
- BLUNDELL, R., D. KRISTENSEN, AND R. MATZKIN (2014): “Bounding Quantile Demand Functions Using Revealed Preference Inequalities,” *Journal of Econometrics*, 179, 112–127.
- BRADY, R. L. AND J. REHBECK (2016): “Menu-Dependent Stochastic Feasibility,” *Econometrica*, 84, 1203–1223.
- BUGNI, F. A. (2016): “Comparison of Inferential Methods in Partially Identified Models in terms of Error in Coverage Probability,” *Econometric Theory*, 32, 187–242.
- BUGNI, F. A., I. A. CANAY, AND X. SHI (2015): “Specification Tests for Partially Identified Models Defined by Moment Inequalities,” *Journal of Econometrics*, 185, 259–282.

- CANAY, I. A. (2010): “EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity,” *Journal of Econometrics*, 156, 408–425.
- CANAY, I. A. AND A. M. SHAIKH (2017): “Practical and Theoretical Advances for Inference in Partially Identified Models,” in *Advances in Economics and Econometrics: Volumen 2: Eleventh World Congress*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, 271–306.
- CHEN, L. H., L. GOLDSTEIN, AND Q.-M. SHAO (2010): *Normal Approximation by Steins Method*, Springer.
- CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2014): “Testing Many Moment Inequalities,” *arXiv:1312.7614*.
- CHERNOZHUKOV, V., W. K. NEWWEY, AND A. SANTOS (2015): “Constrained Conditional Moment Restriction Models,” *arXiv:1509.06311*.
- DE CLIPPEL, G. AND K. ROZEN (2014): “Bounded Rationality and Limited Datasets,” *SSRN Working Paper 2018463*.
- DEAN, M., Ö. KIBRIS, AND Y. MASATLIOGLU (2017): “Limited Attention and Status Quo Bias,” *Journal of Economic Theory*, 169, 93–127.
- DEB, R., Y. KITAMURA, J. K. QUAH, AND J. STOYE (2017): “Revealed Price Preference: Theory and Stochastic Testing,” *Cowles Foundation Discussion Paper No. 2087*.
- ECHENIQUE, F. AND K. SAITO (2017): “General Luce Model,” *Caltech SS Working Paper 1407*.
- ECHENIQUE, F., K. SAITO, AND G. TSERENJIGMID (2014): “The Perception Adjusted Luce Model,” *Caltech SS Working Paper 1379*.
- FANG, Z. AND A. SANTOS (2015): “Inference on Directionally Differentiable Functions,” *arXiv:1404.3763*.
- FUDENBERG, D., R. IJIMA, AND T. STRZALECKI (2015): “Stochastic Choice and Revealed Perturbed Utility,” *Econometrica*, 83, 2371–2409.
- GAUNDRY, M. J. AND M. G. DAGENAIS (1979): “The Dogit Model,” *Transportation Research Part B: Methodological*, 13, 105–111.
- GOEREE, M. S. (2008): “Limited Information and Advertising in the US Personal Computer Industry,” *Econometrica*, 76, 1017–1074.
- GUL, F., P. NATENZON, AND W. PESENDORFER (2014): “Random Choice as Behavioral Optimization,” *Econometrica*, 82, 1873–1912.
- HAUSER, J. R. AND B. WERNERFELT (1990): “An Evaluation Cost Model of Consideration Sets,” *Journal of Consumer Research*, 16, 393–408.
- HAUSMAN, J. A. AND W. K. NEWWEY (2017): “Nonparametric Welfare Analysis,” *Annual Review of Economics*, 9, 521–46.
- HO, K. AND A. M. ROSEN (2017): “Partial Identification in Applied Research: Benefits and Challenges,” in *Advances in Economics and Econometrics: Volumen 2: Eleventh World Congress*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, 307–359.
- HONKA, E., A. HORTAÇSU, AND M. A. VITORINO (2017): “Advertising, Consumer Awareness, and Choice: Evidence from the US Banking Industry,” *The RAND Journal of Economics*, 48, 611–646.
- HORAN, S. (2013): “Random Consideration and Choice,” *Mimeo, CIRPÉE and Université du Québeca Montréal*.

- HOUTHAKKER, H. S. (1950): “Revealed Preference and the Utility Function,” *Economica*, 17, 159–174.
- IMBENS, G. W. AND C. F. MANSKI (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72, 1845–1857.
- KAWAGUCHI, K. (2017): “Testing rationality without restricting heterogeneity,” *Journal of Econometrics*, 197, 153–171.
- KITAMURA, Y. AND J. STOYE (2016): “Nonparametric Analysis of Random Utility Models,” *arXiv:1606.04819*.
- LLERAS, J. S., Y. MASATLIOGLU, D. NAKAJIMA, AND E. Y. OZBAY (2017): “When More Is Less: Limited Consideration,” *Journal of Economic Theory*, 170, 70–85.
- MANSKI, C. F. (2014): “Identification of Income-Leisure Preferences and Evaluation of Income Tax Policy,” *Quantitative Economics*, 5, 145–174.
- MANZINI, P. AND M. MARIOTTI (2014): “Stochastic Choice and Consideration Sets,” *Econometrica*, 82, 1153–1176.
- MASATLIOGLU, Y., D. NAKAJIMA, AND E. Y. OZBAY (2012): “Revealed Attention,” *American Economic Review*, 102, 2183–2205.
- MASATLIOGLU, Y. AND E. SULEYMANOV (2017): “Decision Making within a Product Network,” *Working Paper, University of Michigan*.
- MATZKIN, R. L. (2013): “Nonparametric Identification in Structural Economic Models,” *Annual Review of Economics*, 5, 457–486.
- REUTSKAJA, E., R. NAGEL, C. F. CAMERER, AND A. RANGEL (2011): “Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study,” *American Economic Review*, 101, 900–926.
- RIESKAMP, J., J. R. BUSEMEYER, AND B. A. MELLERS (2006): “Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice,” *Journal of Economic Literature*, 44, 631–661.
- ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2014): “A Practical Two-Step Method for Testing Moment Inequalities,” *Econometrica*, 82, 1979–2002.
- SALANT, Y. AND A. RUBINSTEIN (2008): “(A, f): Choice with Frames,” *Review of Economic Studies*, 75, 1287–1296.
- SIMON, H. A. (1955): “A Behavioral Model of Rational Choice,” *Quarterly Journal of Economics*, 69, 99–118.
- SULEYMANOV, E. (2017): “Stochastic Attention and Search,” *Mimeo, University of Michigan*.
- SUPPES, P. AND R. D. LUCE (1965): “Preference, Utility, and Subjective Probability,” *Handbook of Mathematical Psychology*, 3, 249–410.
- TVERSKY, A. (1972): “Elimination by Aspects: A Theory of Choice,” *Psychological Review*, 79, 281.
- VAN NIEROP, E., B. BRONNENBERG, R. PAAP, M. WEDEL, AND P. H. FRANSES (2010): “Retrieving Unobserved Consideration Sets from Household Panel Data,” *Journal of Marketing Research*, 47, 63–74.