Latent Utility and Latent Choices *

Roy Allen

Department of Economics

University of Western Ontario

rallen46@uwo.ca

April 29, 2026

Abstract

This paper studies measurement in a two stage discrete choice model. We observe a final discrete choice, but the first stage is latent such as a nest of alternatives, effort level, quantity of a related good, consideration set, application to schools, or search process. This paper integrates out the first stage, obtaining a tractable demand model for probabilities. The model is fully characterized by a version of Roy's identity in probabilities, and complementarity is allowed. Utility indices are nonparametrically identified using exclusion restrictions and we can conduct standard counterfactual and welfare analysis for covariate changes. Integrating out the latent stage avoids enumerating all possible latent choices, which is challenging in contexts with combinatorial structure such as consideration set and search.

*I thank Víctor Aguiar, Nail Kashaev, and Rory McGee for helpful comments.

1 Introduction

Classic discrete choice focuses on a single decision problem, to choose the alternative k that maximizes utility of the form

$$u_k(X_k) + \varepsilon_k.$$

Suppose the analyst cares ultimately about a single discrete choice. A key concern is that this choice does not occur in isolation from other choices. This leads to an endogeneity problem when the *latent choices* are related to both covariates X and the unobservables ε. This problem has been long-recognized in stochastic choice analysis in mathematical psychology and economics (Manski, 1977). Early analysis with structured endogeneity includes multi-stage choice models such as nested logit (Ben-Akiva, 1973), in which a subset of alternatives are chosen as a "first step" choice, and then the final choice is made. Multi-stage models were further developed in contexts of endogenous consideration sets (Hauser and Wernerfelt, 1990) and search (Weitzman, 1979), where an individual first chooses alternatives to consider/search. Other applications include effort choice as in certain (un)employment models, and application choice as in school choice models (Arcidiacono, 2005).

The presence of latent choices introduces several measurement questions. Intuitively, the model gets larger, but how much larger? Does "anything go?" or is there structure that remains? What can we identify when we do not observe the latent choices? Can we generally identify two stages of choice separately, or do we need additional assumptions? Does the model become more tractable or less tractable when we introduce latent choices?

This paper addresses these questions for a unifying *two stage additive random utility model*, where the "first stage" choice is unobserved, while the final discrete choice is observed. Hence, the setting involves latent utility and latent choices. The model here retains most structure of the classic one stage additive random utility model (McFadden, 1981), but is strictly more general when there are 3 or more final alternatives, since complementarity is allowed. The key assumption is that the latent choices are made prior to the final choice, to maximize expected utility. This leads to structured endogeneity between the additive shock in a traditional discrete choice analysis, and the latent action. The first result of this paper is to characterize the reduced form

once we integrate out the latent choices. It is "as if" a representative agent chooses probabilities p for the alternatives, conditional on covariates X, to maximize utility as in a perturbed utility model (PUM) of the form

$$\sum_{k=1}^{K} p_k u_k(X_k) + D(p).$$

Moreoever, the indirect utility of this representative agent problem is the average indirect utility, and hence has welfare content. The distribution of additive shocks in the first and second stages determines D. This novel equivalence result permits a direct application of demand tools, because this is a quasilinear utility setup if we interpret probabilities p as quantities, each utility index u_k as a negative price, and D as the "usual" utility representing substitution/complementarity patterns.

Using demand tools, we can conduct measurement of utility indices u_k, and changes in ex ante indirect utility and counterfactual bounds when covariates change. Importantly, this is *despite* the fact that we cannot generally seperate the two stages. In particular, I show that for the two stage model here, it is without loss of generality for the first stage latent actions to be *deterministic*, and hence we cannot generally recover a distribution over the latent actions. The simplest example of this is classic nested logit, which for certain parameter values can be written as either a one step model, or a two step model.

Analysis in this paper centers around a dual indirect utility approach. The key tool is the envelope theorem:

$$P(X) = \nabla \overline{V}(v)\big|_{v_k = u_k(X_k)},$$

where $\overline{V}$ is a convex function (dual to D) that measures ex ante welfare and P encodes primitive choice probabilities. Previous work has shown that an envelope theorem is implied by the one stage classic discrete choice model (McFadden, 1981; Sørensen and Fosgerau, 2022) and further that choice probabilities maximize a utility as in PUM (Hofbauer and Sandholm, 2002; Allen and Rehbeck, 2019a). This paper closes the gap by showing that when we introduce structured latent choices, the envelope theorem and equivalent PUM representation completely characterize the model.

The classic one step discrete choice model requires special structure on $\overline{V}$ forcing substitution and other alternating-signs properties (McFadden, 1981). The key difference

for the two stage model is that complementarity is allowed, so that making k more attractive can make j be chosen more often. This is because first stage complementarity can overwhelm the necessary second stage substitution, through reoptimization in the latent action. Thus, the common intuition that discrete choice "must" force substitution does not hold when we examine the comparative statistics of reduced form demand $P(X)$. In nested logit for example, making an alternative more attractive can raise the overall nest choice probability in a way that offsets the second stage substitution. Overall, the setup here provides a general foundation for complementarity in demand systems for market shares, as used in e.g. Fosgerau et al. (2024). Complementarity is ruled out in the demand systems of Berry (1994) and Berry et al. (1995).

I examine possible value functions that span behavior consistent with the model. For the two stage model, with finite support for covariates it is without loss for the value function to take the form:

$$\overline{V}^{2SM}(v) = \max_{j \in \{1,\ldots,J\}} \{p^{j\prime} v + c^j\}.$$

The vector v describes the utility index for each alternative. Here each p^j is a probability vector and we can take J to be equal to the size of the support of X, which importantly does not depend on the dimension of the final alternatives. Empirically, this means given an i.i.d. sample of size n, we need n pieces to fully span probabilities consistent with the model. This result allows us to avoid combinatorial type-enumeration methods that are widely used in the partial identification literature (McFadden and Richter, 1990; Honoré and Tamer, 2006; Manski, 2007; Kitamura and Stoye, 2018; Gu et al., 2025; Mbakop, 2023). New methods are in fact necessary here because it is not possible to enumerate types when the latent actions are of unrestricted size. For comparison with the classic one step model (which can be enumerated), I present a characterization showing the value function instead can take the expected maximum form:

$$\overline{V}^{1SM}(v) = \sum_{j=1}^{J} p^j \max_k \{v_k + c_k^j\}.$$

This corresponds to a discrete distribution with scalar probability p^j on shock vector c^j. We need more types to span the model, but with this novel description, the

number of types J is still *linearly* increasing in the support of X (or n in the case of empirical distributions).

This paper treats the first stage abstractly. The first stage could be complicated, and could involve an entire latent process called the "action." Recall a central finding is that we still have a quasilinear demand model. It is helpful comparing this with the existing finding that the Weitzman (1979) sequential search model can be written as a classic one stage discrete choice model (Kleinberg et al., 2016; Armstrong, 2017; Choi et al., 2018). Likewise, Matějka and McKay (2015) and Fosgerau et al. (2020) show equivalence of certain rational inattention models and one stage discrete choice models. This paper differs from this work because we study a general class of models, permitting complementarity. A cost of exploiting only the general structure is that when we do not explicitly model or estimate the latent action process, we cannot assess all possible counterfactuals involving the latent process. Instead, the counterfactuals this framework can address involve changes in covariates X, allowing the endogenous latent action to reoptimize.

The endogeneity problem tackled in this paper requires the latent and observed choices to be part of an unified choice problem. Specifically, the value function of the second stage enters the first stage, which is *essential* here. Indeed, the core technical structure of this paper is to combine the envelope theorem, law of total probability, and the chain rule. A unified choice structure arises in certain models of search (Weitzman, 1979), dynamic discrete choice (Rust, 1987),[1] endogenous consideration sets (Hauser and Wernerfelt, 1990), and school choice (Arcidiacono, 2005). This differentiates the work from equilibrium models in which different actors make different choices, and so the setup here is not designed to cover price endogeneity (Berry, 1994; Berry et al., 1995; Tebaldi et al., 2023), which is commonly addressed using instrumental variables. Likewise, the setup here does not generally match two sided matching contexts that have second stage shifters that do not shift the first stage (He et al., 2024; Agarwal and Somaini, 2022), or models of limited consideration where first stage consideration set formation does not reference the second stage value function (Goeree, 2008; Abaluck and Adams-Prassl, 2021). Latent states have

[1] Identification is notoriously difficult in dynamic choice models (Rust, 1994; Magnac and Thesmar, 2002); this paper differs by observing only composite choice probabilities and not each stage separately, but does not study discounting or transitions in observable state variables.

been analyzed in dynamic discrete choice in Hu and Shum (2012), which uses repeated measurements while here we observe some choices in a single period.

In addition to papers focusing specifically on search and consideration sets,[2] there is a small literature in static discrete choice allowing some unobservable choices. Thompson (1989) shows that in a one step discrete choice model, observing one binary choice probability suffices to identify other choice probabilities when we know each utility index and it satisfies rich support conditions. Fosgerau and Ranjan (2017) show this result generalizes if instead we observe the sum of choice probabilities for a collection of alternatives. Kono et al. (2025) study empirical content with unobservable alternatives in the abstract random utility model with menu variation, with no covariates. The literature on mixed discrete/continuous choice as in Hanemann (1984) studies discrete choices as related to other choices, but canonically all choices are observed. The dynamic discrete choice literature

The rest of the paper is organized as follows. Section 2 presents the two stage model and provides examples. Section 3 describes the observational equivalence of the two stage model and PUM, illustrating how we can take a demand approach. Section 4 analyzes the value functions in each of the stages, presents envelope relations, and shows how the value functions are linked through the chain rule. Section 5 shows connections with the classic one stage model, shows the role of the number of latent actions, and shows how piecewise linear value functions suffice to trace out behavior consistent with the model.

2 The Two Stage Model

The model involves two stages. The second stage follows the classic discrete choice model in using additive separability between covariates and random latent components. The difference is that this additive latent component is endogenized, and can depend on the first stage action. The first stage is to optimally choose an action to maximize expected utility, net of an additive cost.

The core empirical primitives for this paper are covariates X in the final stage and choice probabilities over the final alternatives, but not the first stage latent action.

[2]For surveys see Honka et al. (2019) and Crawford et al. (2021).

In some empirical settings, researchers also have access to covariates W that shift the costs of the first stage. Choice probabilities for the observed choices are collected in a vector $P(X, W)$ with k-th choice probability denoted $P_k(X, W)$. The "classic" static case has $W = w$ degenerate. The first part of the paper treats the probabilities as known to study model structure and identification questions. Loosely, such choice probabilities can be estimated from an independent and identically distributed sample of choices together with corresponding covariates (X, W).

Assumption 1 (Empirical Primitive)**.** *The analyst identifies $P(X, W)$ over the support of (X, W).*

I now describe the two stages in detail. In the first stage, the individual chooses a latent action $A \in \mathcal{A}$ to maximize utility:

$$A \in M^F(X, \xi, W, \omega) := \underset{A \in \mathcal{A}}{\operatorname{argmax}} V_A(X, \xi) + C(A, W, \omega),$$

where $V_A(X, \xi)$ is a value associated with the second stage that will be described shortly. This setup allows utility ties with an arbitrary rule for breaking utility ties. First stage choice probabilities are then given by the $|\mathcal{A}|$-dimensional vector

$$P^F(X, W) \in \mathbb{E}[\text{Conv}\{e_B^{|\mathcal{A}|} \mid B \in M^F(X, \xi, W, \omega)\} \mid X, W].$$

We encode the event that B is chosen through the basis vector $e_B^{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$. The convex hull describes all possible mixtures (probabilities) over elements of the argmax set,[3] which we need here because we allow utility ties. The expectation is the conditional Aumann expectation, taken over all possible measurable selections from the convex hull (see e.g. Beresteanu et al. (2011) or Sørensen and Fosgerau (2022)).

We assume for any (X, W, ξ, ω) there is some A that attains the maximum, which further yields finite utility. Covariates are treated as known in the first stage. The function $C : \mathcal{A} \times \text{supp}(W) \times \Omega \to \mathbb{R} \cup \{-\infty\}$ is termed a *value distortion function*, and depends on observable shifters W, the action A, and an unobservable ω. A leading interpretation here is the cost a first stage action, in which lower values of C are more costly. Alternatively, it can be interpreted as the flow utility of action A in the first period. The action set $\mathcal{A}$ is finite, so that we can accurately say this is a two

[3]The convex hull of a set Γ, denoted $\text{Conv}(\Gamma)$, is the smallest convex set containing Γ.

stage discrete choice model. The interpretation of $C(A, W, \omega) = -\infty$ is that action A is infeasible or arbitrarily undesirable given (W, ω). This allows random constraints on A. Because we assume there is a maximizer, we rule out cases in which random constraints eliminate all A.

In the final stage, given a first stage action A, the individual chooses an alternative k^* to maximize utility as:

$$k^* \in M^S(X, A, \eta, \xi) := \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \{u_k(X_k) + \varepsilon_k(A, \eta, \xi)\}.$$

The second stage choice probability vector then satisfies

$$P^S(X, W, A) \in \mathbb{E}[\operatorname{Conv}\{e_k^K \mid k \in M^S(X, A, \eta, \xi)\} \mid X, W, A].$$

Here, e_k^K represents that alternative k is chosen. The shock ε_k and its arguments are latent. The unobservable η is known to the individual only in the second stage, while the other variables are known in both stages. Endogeneity arises here because A can depend on X and ξ. There are finitely many alternatives K, X_k is a vector of covariates, $X = (X_1', \ldots, X_K')'$ collects all covariates, $u_k : \operatorname{supp}(X_k) \to \mathbb{R}$ is a price-like utility index for good k, and $\varepsilon_k : \mathcal{A} \times \mathcal{H} \times \Xi \to \mathbb{R} \cup \{-\infty\}$ is an additive disturbance that can depend on the first stage choice A and latent shocks. Random constraints that depend on the action and unobservables are represented via combinations satisfying $\varepsilon_k(A, \eta, \xi) = -\infty$.[4] We assume for any $(X, A, \eta, \xi, \omega)$ that there exists some k such that $\varepsilon_k(A, \eta, \xi)$ is finite. Thus, the final stage maximization problem has a solution that attains finite utility.

Before η is realized, the value of action A is

$$V_A(X, \xi) = \mathbb{E}\left[\max_k \{u_k(X_k) + \varepsilon_k(A, \eta, \xi)\} \mid A, X, \xi\right].$$

We impose throughout that expectations are finite. We maintain the following exogeneity conditions.

Assumption 2 (Exogeneity). *(i) η is independent of (A, X, W, ξ, ω).*

(ii) (ξ, ω) is independent of (X, W).

[4] We define $b + -\infty = -\infty$ for scalar (finite) b.

Part (i) means we can interpret η as an exogenous shock arriving after the first stage action A. In particular, W cannot predict η and is excluded from each value V_A. In the general nonparametric model, the exogenous unobservables (η, ξ, ω) enter their respective functions in arbitrary nonseparable ways, and have unrestricted dimension. In general, ξ enters both stages, and so this paper does not require full independence between unobservables in the two stages. Moreover, ξ and ω can be arbitrarily related. We can interpret ξ as a preference shifter and/or part of the information set in the first stage ε_k, so that $\varepsilon \mid A, X, \xi$ has a distribution that varies flexibly with ξ.

The overall choice probability for each alternative k can be written as

$$P_k(X, W) = \sum_{A \in \mathcal{A}} P_k^S(X, W, A) P_A^F(X, W).$$

These are conditional probabilities, so that P_k is the conditional probability of choosing k overall, P_k^S is the second stage probability conditional on A, and P_A^F is the probability of choosing A in the first stage. This setup is termed the *2-stage additive random utility model* (2S-ARUM), or for shorthand just the two stage model (2SM). It covers a variety of existing models including the classic one step additive random utility model of McFadden (1981), which corresponds to the first stage action set being a singleton.[5] This paper shows that the general structure of 2SM is tractable.

2.1 Examples

Example 1 (Nested Logit)**.** *Given a partition of alternatives* $\{1, \ldots, K\}$ *into mutually exclusive nests, the standard nested logit formula reads:*

$$\begin{aligned} P_k(X) &= \frac{e^{u_k(X_k)/\lambda(A)}}{\sum_{j \in K(A)} e^{u_j(X_j)/\lambda(A)}} \frac{e^{V_A(X)}}{\sum_{B \in \mathcal{A}} e^{V_B(X)}} \\ &= P_k^S(X, A) P_A^F(X) \end{aligned}$$

where k is in nest $K(A)$. The action A in the first stage is to select the nest $K(A)$. A classic example is to select a neighborhood A that consists of a collection of homes (McFadden, 1978). The first stage has the usual logit formula, corresponding to

[5] Technically, the $|\mathcal{A}| = 1$ case is discrete choice in the extended reals Allen (2026), which does not assume choice probabilities go to 0/1 at extreme values of utility indices.

$C(A, W, \omega) = \omega_A$ for shocks $\{\omega_A\}$ that are i.i.d. type one extreme value. The second stage sets $\varepsilon_k(A, \eta) = \lambda(A)\eta_k$ for any alternative k in nest $K(A)$, and is $-\infty$ otherwise. Shocks $\{\eta_k\}$ are also i.i.d. extreme value. The values are given by the well-known log-sum formula

$$\begin{aligned} V_A(X) &= \mathbb{E}\left[\max_{k \in K(A)} u_k(X_k) + \varepsilon_k(A, \eta) \mid X, A\right] \\ &= \lambda(A)\mathbb{E}\left[\max_{k \in K(A)} u_k(X_k)/\lambda(A) + \eta_k \mid X, A\right] \\ &= \lambda(A) \ln\left(\sum_{k \in K(A)} e^{u_k(X_k)/\lambda(A)}\right). \end{aligned}$$

The maximum here is taken over $k \in K(A)$ because other alternatives have utility $-\infty$. We thus see that nested logit is a 2-stage additive random utility model, for any scale function $\lambda(\cdot) > 0$.

One motivation of nested logit is sequential choices as here (Ben-Akiva, 1973; Anderson and De Palma, 1992).[6] *This model can be written as a one step model when each nest satisfies $0 < \lambda(A) \leqslant 1$.*[7] *From the description above, we can also write this as a two step model where $P_A^F(X) < 1$. Importantly, this means we generally cannot separate the two stages nonparametrically, since for certain parameter values nested logit can be rationalized by a nontrivial two step model or one step classic discrete choice model.*

Similar analysis holds for generalizations of nested logit in which alternatives show up in multiple nests, such as paired combinatorial logit (Koppelman and Wen, 2000) and the inverse product differentiation logit of Fosgerau et al. (2024). For the purposes of this paper, these generalizations better match the setup here because when an alternative is in multiple nests, we do not conclusively observe the first stage action.

Example 2 (Limited Observability of Demand)**.** *Suppose we model a discrete choice over K alternatives, but there are goods A chosen that we do not observe. Write final*

[6] See also Kovach and Tserenjigmid (2022), who study the empirical content of nested logit and generalizations.

[7] Train (2009) treats "utility maximization" as synonymous with the one step model.

realized utility for final alternative k^* and unobserved quantities A as:

$$\sum_{k=1}^{K} 1\{k^* = k\} u_k(X_K) + \sum_{k=1}^{K} 1\{k^* = k\} \varepsilon_k(A, \eta, \xi) + C(A, W, \omega).$$

This covers the bundles model of Gentzkow (2007), allowing limited observability of parts of the bundle. This setup also includes quasilinear demand, where X_k would include prices of the observed K goods, and W would include the prices of the unobserved goods.[8] *Total expenditure is the sum of expenditure on k and A, which matches our separability assumptions.*

Here, the primitive $P(X, W)$ requires observing choices over final alternatives, shifters X for those choices, and covariates W that shift the desirability of the unobserved bundle. If covariates W are not observed but are independent of X, then they can be absorbed into the first stage unobservables ω. Importantly, generally it is not appropriate to just condition on W and use traditional discrete choice tools. The reason is that probabilities can exhibit complementarity between the final alternatives due to the latent action A, and hence we need a more general model.

Example 3 (Costly Effort)**.** *Consider a binary choice setting in which an individual can exert costly effort A as a first stage action. For simplicity, A is an intensity and the first stage objective function is*

$$V_A(X) - A^2,$$

where the value distortion function is just $-A^2$. The intensity alters the distribution of utility shocks in the second stage. If $A \geqslant \eta^A$, the shock is η^H, and otherwise it is η^L, where $\eta^H \geqslant \eta^L$ almost surely. Thus, higher effort leads probabilistically to a higher utility shock associated with alternative 1. The second stage unobservables are collected in $\eta = (\eta^A, \eta^H, \eta^L)$. Overall, the second stage shock is encoded as $\varepsilon_1(A, \eta)$ and the overall utility of alternative 1 in the second stage is

$$u_1(X_1) + \varepsilon_1(A, \eta).$$

[8] Generally, one should not model price as exogenous. This comment is about how to relate the models if prices were exogenous and observed.

The utility of alternative 2 is an "outside option" set to 0. This example readily generalizes to more than two outcomes, where the effort level A is then a vector of efforts associated with different possible outcomes.

Example 4 (Endogenous Consideration Sets / Simultaneous Search)**.** *Similar to nested logit, let the first stage be a choice of consideration set $K(A) \subseteq \{1, \ldots, K\}$, indexed by the action A. Unlike nested logit, the consideration set need not be an element of a partition; any possible subset of $\{1, \ldots, K\}$ can be considered. The optimal choice of consideration set is specified via maximization of*

$$V_A(X, \xi) + C(A, W, \omega).$$

The value functions for the set $K(A)$ is given by

$$V_A(X, \xi) = \mathbb{E}\left[\max_{k \in K(A)} u_k(X_k) + \varepsilon_k(A, \eta, \xi) \mid A, X, \xi\right],$$

where $k \notin K(A)$ has $\varepsilon_k(A, \eta, \xi) = -\infty$, so that the choice must lie in $K(A)$. Special cases of this include Hauser and Wernerfelt (1990), Honka (2014), and the simultaneous search models in Honka et al. (2019).[9]

This paper integrates out A, and so we do not need to ever solve the first stage problem. This is important in consideration set / simultaneous search models, because enumerating all possible consideration sets grows as 2^K, which is prohibitive when K is moderately large. For example, with $K = 40$ there are over a trillion first stage sets. Each increase of 10 increases the number of sets by a factor of $2^{10} = 1048$.

Example 5 (School Choice)**.** *Consider a school application setting, as in Arcidiacono (2005), Howell (2010), Fu (2014), and Walters (2018). In the first stage action is which schools to apply to. Schools are chosen to maximize*

$$V_A(X, \xi) + C(A, W, \omega).$$

The potential application combinations are $A = \{0, 1\}^K$, where $A_k = 1$ denotes that the individual applied to school k. After the application, admissions only depend on A and a lottery device η independent of everything else, resulting in a choice set

[9] See also Roberts and Lattin (1991) and Ajayi and Sidibe (2020).

$K(A, \xi, \eta)$. Notably, this can depend on latent variables ξ. Finally, shocks are drawn and the individual chooses a school k to maximize

$$u_k(X_k) + \varepsilon_k(A, \eta, \xi).$$

Recall that ε_k is allowed to be $-\infty$ to encode constraints. Thus, $\varepsilon_k(A, \eta, \xi) = -\infty$ whenever $k \notin K(A, \xi, \eta)$. In the general model, η can conceptually be thought of as having two pieces: the lottery device describing the feasibility set, and an additive preference shock describing the numerical value of utility for a school.

One challenge in the literature is that we may not observe all applications. For example, Arcidiacono (2005) uses NLSY class of 1972 data, and observes reports of the top three choices, but does not observe all applications. A second issue is numerical: similar to the consideration set / simultaneous search setting, fully specified estimation of the two stages is difficult, because we must enumerate all possible actions because they grow as 2^K, where K is the number of schools. For numerical reasons, Arcidiacono (2005) restricts attention to 8 schools, and models that possible application lists have 3 schools. Che et al. (2025) reviews the broader literature on school choice, including other methods designed to address the combinatorial curse of dimensionality. This paper sidesteps the observability and numerical issues by integrating out the latent action and studying empirical content directly. The cost of this is that this paper does not study counterfactuals that can change the structure of the first stage, focusing instead on counterfactual/welfare questions involving covariate changes.

3 Observational Equivalence and Demand Approach

When we introduce latent choicess as in the two stage model, we unify quasilinear demand and discrete choice.

To state the result, define the probability simplex here as

$$\Delta^{K-1} = \left\{ p \mid \sum_{k=1}^{K} p_k = 1, p_k \geqslant 0 \right\}.$$

Definition 1 (Perturbed Utility Model)**.** *$P(X,W)$ is consistent with the one stage perturbed utility model (PUM) with finite-valued utility indices u if the following conditions hold: For every $(x,w) \in \text{supp}(X,W)$,*

$$P(x,w) \in \underset{p \in \Delta^{K-1}}{\text{argmax}} \sum_{k=1}^{K} p_k u_k(x_k) + D(p,w)$$

for a function $D : \mathbb{R}^K \times \mathbb{R}^{dim(W)} \to \mathbb{R} \cup \{-\infty\}$ such that for each $w \in \mathbb{R}^{dim(W)}$, $D(p,w) > -\infty$ for some $p \in \Delta^{K-1}$.

Theorem 1. *Consider the following conditions.*

(i) $P(X,W)$ is consistent with the 2-stage additive random utility model, with utility indices u.

(ii) $P(X,W)$ is consistent with the one stage perturbed utility model, with utility indices u.

Condition (i) always implies (ii). If in addition $\text{supp}(X,W)$ is finite, then conditions (i) and (ii) are equivalent, and it is without loss of generality to consider $\mathcal{A} = |\text{supp}(X,W)|$ latent actions with degenerate first stage shocks (ξ,ω) and degenerate first stage choices $P_A^F(X,W) \in \{0,1\}$.

Theorem 1 formalizes that the two stages collapse to a single maximization problem, so we can view PUM as the reduced form of 2SM. The requirement of finite support for the complete characterization is because the first stage set of actions is discrete, so that we can call this a two stage discrete choice model. The fact that we can rationalize 2SM with $P_A^F(X,W) \in \{0,1\}$ means it is impossible to conclusively separate the two stages at this level of generality when we do not observe A. That is, we cannot point identify P^F or P^S in general, and we cannot identify the distribution of (ξ,ω) or test if these shocks are degenerate. This motivates the focus on utility indices u and ex ante welfare (formalized soon), which are nonparametrically identified, instead of stage-specific objects.

Classic discrete choice (McFadden, 1981) shows connections between demand and one step discrete choice, with a key difference being that one step discrete choice forces substitution. Here we see that when we introduce a latent action, we can in fact

view discrete choice as equivalent to quasilinear demand, which allows complementarity. In more detail, PUM here is analogous to quasilinear utility, by mapping u_k to negative price of k, P_k to quantity, and D to the "utility" controlling complementary/substitutability patterns.

The term perturbed utility follows McFadden and Fosgerau (2012) in a demand context and Fudenberg et al. (2015) and Allen and Rehbeck (2023) in stochastic choice. Fudenberg et al. (2015) study a version without covariates, focusing on a disturbance $D(p) = \sum_{k=1}^{K} d(p_k)$ that models goods as symmetric substitutes, which overlaps with the one step additive random utility model precisely in the logit case $d(p_k) = p_k \ln(p_k)$ (Hofbauer and Sandholm, 2002). Machina (1985) advances that stochastic choice is potentially due to deliberate randomization, and PUM is a formalism of this idea. Chapter 3 in Rehbeck (2017) interprets u as a local utility as in Machina (1982). Mattsson and Weibull (2002) interpret D as a control cost of implementing a choice. An early appearance of the PUM representation in a market share context is Anderson et al. (1988). Recent market share work includes Fosgerau et al. (2024).

Remark 1 ($\mathcal{A}$ and Classic Discrete Choice). Observational equivalence holds when we do not restrict the first stage in any way other than the stated separability assumptions. Additional restrictions on the cardinality of $\mathcal{A}$ place additional empirical restrictions, as examined further in Section 5. A leading example is $\mathcal{A}$ a singleton, so that the first stage choice is degenerate, which reduces to familiar discrete choice.

Remark 2 (Role of W). Theorem 1 establishes that the covariates W in the first stage enter in a fully unstructured way, because the function $D(\cdot, w)$ in the perturbed utility representation can change arbitrarily when we go from w to some other $\tilde{w}$. Importantly though, the utility indices u are the same at different values of conditioning variables. Thus, while complementarity/substitutability patterns can change *arbitrarily* with changes in W, the relative desirability of covariate changes is stable.

3.1 Binary Choice

When $K = 2$, 2SM is equivalent to an extended binary choice model when we condition on W. A key takeaway is that endogeneity that arises due to latent actions is not a problem in binary choice, *provided* the latent actions maximize expected utility as formalized here.

To state the equivalence, the *extended binary choice model* takes the form:

$$(Y_1, Y_2) \in \underset{(y_1,y_2)\in\{0,1\},y_1+y_2=1}{\operatorname{argmax}} y_1(u_1(X_1) + \eta_1) + y_2(u_2(X_2) + \eta_2),$$

where conditional on W, (X_1, X_2) and (η_1, η_2) are independent. Here, (η_1, η_2) has support in $[-\infty, \infty)^2$, but $P(\eta_1 = -\infty, \eta_2 = -\infty \mid W) = 0$ so that at most one shock is $-\infty$. Here, Y_k is an indicator for whether alternative k is chosen.

Proposition 1. *Suppose $K = 2$ and $\operatorname{supp}(X, W)$ is finite. The following are equivalent:*

(i) $P(X, W)$ is consistent with the 2-stage additive random utility model, with utility indices u.

(ii) $P(X, W)$ is consistent with the one stage perturbed utility model, with utility indices u.

(iii) $P(X, W)$ is consistent with the extended binary choice model, with utility indices u.

(iv) $P(X, W)$ is monotone: for $w \in \operatorname{supp}(W)$ and any $x, \tilde{x} \in \operatorname{supp}(X \mid W = w)$, we have

$$(P(\tilde{x}, w) - P(x, w))'(u(\tilde{x}) - u(x)) \geqslant 0.$$

In fact, (iii) and (iv) are always equivalent (for $K = 2$), regardless of whether $\operatorname{supp}(X, W)$ is finite.

In familiar examples that rule out utility ties, $P_1(X, W) = H(u_1(X_1) - u_2(X_2); W)$, where H is the cumulative distribution function for the difference in the additive shocks $(\eta_2 - \eta_1)$, and can depend on W here. When we allow utility ties and work in the extended reals (to handle random exogenous constraints), we generally lose this representation. The equivalence of (iii) and (iv) is related to the equivalence result of Vytlacil (2002), which connects monotonicity, binary choice thresholding crossing models, and maximization. Weak monotonicity in (iv) can be rewritten as the implication $u_2(\tilde{x}_2) - u_1(\tilde{x}_1) > u_1(x_1) - u_2(x_2)$ implies $P_1(\tilde{x}, w) \geqslant P_1(x, w)$.

3.2 Complementarity

Hofbauer and Sandholm (2002) prove that the one stage additive random utility model is strictly smaller than the perturbed utility model when $K > 3$, which immediately implies the same result for 2SM. The examples they present involve substitution between alternatives that is not strong enough. Another way to see why 2SM is bigger than the one stage model is that when $K > 2$, the model permits complementarity. I illustrate this in a simple nested logit example, which is a special case of Example 1.

Consider a partition of alternatives into nests $\{1, 2\}$ and $\{3\}$. Here there is a single λ, which controls the nest $\{1, 2\}$. Further, suppose for simplicity that $u_k(X_k) = X_k$ is scalar so that we can formalize complementarity directly in terms of changes in u_k. The probability of choosing $k = 1$ is given by

$$P_1(X) = P_1^S(X, A_{12}) P_{A_{12}}^F(X),$$

where A_{12} denotes that the nest $\{1, 2\}$ is chosen. Differentiating with respect to X_2 yields the cross-derivative:

$$\frac{\partial P_1(X)}{\partial X_2} = \frac{\partial P_1^S(X, A_{12})}{\partial X_2} P_{A_{12}}^F(X) + P_1^S(X, A_{12}) \frac{\partial P_{A_{12}}^F(X)}{\partial X_2}.$$

The first term is negative and reflects *within-nest* substitution. Indeed, conditional on the action we have substitution here. The second term is positive and reflects *nest-level* complementarity, since making X_2 more attractive raises the inclusive value of the $\{1, 2\}$ nest and thus the probability of choosing A_{12} in the first stage. Which term dominates is an empirical question that depends on λ.

The overall cross derivative is positive, indicating complementarity, precisely when

$$\frac{\lambda - 1}{\lambda} > P_1(X) + P_2(X).$$

See Equation 24 in Allen and Rehbeck (2016) for a generalization to all nested logit models: complementarity occurs in small nests in terms of total choice probabilities. Complementarity requires $\lambda > 1$ (Feng et al., 2018). In the context of the two stage representation of nested logit, $\lambda > 1$ means that the variance of the second stage shock is higher than the first stage shock. See Fosgerau et al. (2024) and Monardo

(2025) for demand systems that generalize nested logit and allow complementarity, and are in the PUM class (and hence also 2SM).

4 Value Functions

This section develops core model structure through value functions.

The results use tools from convex analysis such as subgradients and convex conjugates, which are reviewed more completely in Appendix B. A key object is the subdifferential ∂f, which is a multivalued mapping that is the derivative when f is convex and differentiable, and more generally is given by

$$\partial f(v) = \{y \mid f(\tilde{v}) \geqslant f(v) + y'(\tilde{v} - v), \forall \tilde{v}\}.$$

Elements of ∂f are called subgradients. The convex conjugate of f, denoted f^*, is given by:

$$f^*(v) = \sup_{y \in \mathbb{R}^L} y'v - f(y).$$

4.1 Setup

This section defines the key value functions, and then describes basic convexity properties.

We first need a change of variables to go between u and X space. This will allow us to describe local changes in utility indices and present envelope relations. To that end, define

$$\overline{V}_A^S(v \mid X, \xi) = \mathbb{E}\left[\max_k \{v_k + \varepsilon_k(A, \eta, \xi)\} \mid A, X, \xi\right].$$

The conditional expectation is over η. This value satisfies the change of variables relation $\overline{V}_A^S(u(X) \mid X, \xi) = \overline{V}_A^S(u(X) \mid \xi) = V_A(X, \xi)$, because X does not predict η and only serves to set $v_k = u_k(X_k)$ once we condition on (A, ξ). Here, $\overline{V}_A^S$ describes how the value of the action varies with the final stage utility indices v, which does not depend on the first stage distortion function C. The utility indices are collected in $u(X) = (u_1(X_1), \ldots, u_K(X_K))'$. Define now a value associated with the first stage:

$$\overline{V}^F(\tau \mid X, W, \omega) = \max_{A \in \mathcal{A}} \{\tau_A + C(A, W, \omega)\},$$

where τ has dimension equal to the size of $\mathcal{A}$. When τ varies, this function describes how the ex ante value function changes in the first stage. Finally, define the composition

$$\overline{V}(v \mid X, W, \xi, \omega) = \overline{V}^F\left(\overline{V}^S(v \mid X, \xi) \mid X, W, \omega\right), \tag{1}$$

where $\overline{V}^S(v \mid X, \xi) = \left(\overline{V}^S_A(v \mid X, \xi)\right)_{A \in \mathcal{A}}$ collect the first stage values. The core structure that this paper exploits is that the value function $\overline{V}$ is the composition of two convex functions.

Now define the following counterparts that integrate out the exogenous unobservables:

$$\begin{aligned}
\overline{V}^S_A(v \mid X, W) &:= \mathbb{E}\left[\max_k \{v_k + \varepsilon_k(A, \eta, \xi)\} \mid A, X, W\right] \\
\overline{V}^F(\tau \mid X, W) &:= \mathbb{E}\left[\max_{A \in \mathcal{A}} \{\tau_A + C(A, W, \omega)\} \mid X, W\right] \\
\overline{V}(v \mid X, W) &:= \mathbb{E}\left[\overline{V}(v \mid X, W, \xi, \omega) \mid X, W\right] \\
&= \mathbb{E}\left[\overline{V}^F\left(\overline{V}^S(v \mid X, \xi) \mid X, W, \omega\right) \mid X, W\right].
\end{aligned}$$

The maintained Assumption 2 imposes exogeneity conditions that yield the following exclusion restrictions, so that conditioning on X does not alter the conditional distribution of unobservables.

Lemma 1. *$\overline{V}^F(\tau \mid X, W) = \overline{V}^F(\tau \mid W)$ and $\overline{V}(v \mid X, W) = \overline{V}(v \mid W)$. When ξ is degenerate, $\overline{V}^S(v \mid X, W) = \overline{V}^S(v)$, so that $\overline{V}(v \mid W) = \overline{V}^F(\overline{V}^S(v) \mid W)$.*

The primary target here is $\overline{V}$. Evaluating at $v = u(X)$, we obtain that $\overline{V}(u(X) \mid W)$ is the average indirect utility conditional only on W and X,[10] but not the shocks or first stage action. This is the primary welfare object studied in this paper because its difference as X changes can be nonparametrically identified from $P(X, W)$, as discussed more in Remark 3 below. Importantly for interpretation, fixing W we can vary X and change utility indices only and not the distribution of unobservable exogenous variables. Hence, there is a stable distribution of heterogeneity that can be used for welfare analysis.

To state the next result, let $\text{supp}(X)$ be the support of X, i.e. the smallest closed set B satisfying $P(X \in B) = 1$, and similarly for the other random variables. Recall

[10] Conditioning on X only serves to set $v = u(X)$.

we assume throughout that all expectations stated here are finite. Say a function f is 1-Lipschitz if $\|f(\tilde{v}) - f(v)\|_\infty \leqslant \|\tilde{v} - v\|_\infty$, where $\|a\|_\infty = \max_\ell |a_\ell|$ is the supremum norm of a vector. Say $f : \mathbb{R}^L \to \mathbb{R}$ is translation invariant if $f(v + c1_L) = f(v) + c$ for any $v \in \mathbb{R}^L$ and $c \in \mathbb{R}$, where $1_L \in \mathbb{R}^L$ is the one vector $(1,\ldots,1)$.

Lemma 2. *(i)* $\overline{V}^S_A : \mathbb{R}^K \times \text{supp}(X, W) \to \mathbb{R}$ *is weakly increasing, convex, translation invariant, and* 1*-Lipschitz in* v*, for each* $A \in \mathcal{A}$*.*

(ii) $\overline{V}^F : \mathbb{R}^{|\mathcal{A}|} \times \text{supp}(W) \to \mathbb{R}$ *is weakly increasing, convex, translation invariant, and* 1*-Lipschitz in* τ*.*

(iii) $\overline{V} : \mathbb{R}^K \times \text{supp}(W) \to \mathbb{R}$ *is weakly increasing, convex, translation invariant, and* 1*-Lipschitz in* v*.*

Weakly increasing for $\overline{V}^S_A$ means that if $\tilde{v} \geqslant v$ component-wise, then $\overline{V}^S_A(\tilde{v} \mid X) \geqslant \overline{V}^S_A(v \mid X)$ for any $\xi \in \text{supp}(\xi)$. The extra properties relative to quasilinear demand are that the value functions are translation invariant and 1-Lipschitz in $\|\cdot\|_\infty$.[11] The literature on convex risk measures has presented a number of mathematical results that exploit these features, including parametric examples of such convex functions, focusing on infinite-dimensional outcomes (Eichhorn and Römisch, 2005; Filipović and Kupper, 2007).

4.2 Envelopes

For envelope theorem intuition, plug in the utility indices evaluated at covariates X to obtain

$$\begin{aligned}\overline{V}^S_A(u(X) \mid X, W) &= \mathbb{E}\left[\max_k \{u_k(X_k) + \varepsilon_k(A, \eta, \xi)\} \mid A, X, W\right] \\ &= \sum_{k=1}^K P_k(X, W, A) u_k(X_k) + \mathbb{E}[\varepsilon_{k*}(A, \eta, \xi) \mid A, X, W],\end{aligned}$$

[11] Here, $\overline{V}^S_A$ and $\overline{V}^F$ also have the shape of classic discrete choice value functions. Under smoothness, the second order mixed partials of each $\overline{V}^S_A$ and $\overline{V}^F$ must be weakly negative. This is because within-stage choices force substitution. Other conditions are required, including the alternating signs conditions that formalize a type of strong substitution (McFadden, 1981).

where k^* denotes the choice so that ε_{k*} is the shock at the corresponding choice.[12] Envelope theorem intuition suggests that a small change in u_k will change the value at rate P_k. This is indeed the case, and similar results hold for the other value functions. I record the formal result below.

To state the result, collect the values in the final stage in $\overline{V}^S(u(X) \mid X, W) = \left(\overline{V}^S_A(u(X) \mid X, W)\right)_{A \in \mathcal{A}}$. Let $\partial_\tau \overline{V}^F\left(\overline{V}^S(u(X) \mid X, W) \mid W\right)$ denote the subdifferential of $\overline{V}^F(\cdot \mid W)$, evaluated at the vector $\overline{V}^S(u(X) \mid X, W)$, and similar for $\partial_v \overline{V}$.

Theorem 2. *The following relations hold for 2S-ARUM:*

(i) We have: $P^S(X, W, A) \in \partial_v \overline{V}^S_A(u(X) \mid X, W)$. If in addition conditional on $(X, W, A) = (x, w, a)$, utility ties in the second stage are probability zero events, then $P^S(x, w, a) = \nabla_v \overline{V}^S_a(u(x) \mid x, w)$.

(ii) If ξ is degenerate: $P^F(X, W) \in \partial_\tau \overline{V}^F\left(\overline{V}^S(u(X)) \mid W\right)$. If in addition conditional on $(X, W) = (x, w)$, utility ties in the first stage are probability zero events, then $P^F(x, w) = \nabla_\tau \overline{V}^F\left(\overline{V}^S(u(x)) \mid w\right)$.

(iii) We have: $P(X, W) \in \partial_v \overline{V}(u(X) \mid W)$. If in addition conditional on $(X, W) = (x, w)$, utility ties in the first and second stage are probability zero events, then $P(x, w) = \nabla_v \overline{V}(u(x) \mid w)$.

Part (iii) describes the key structure for our purposes, which references only the empirical primitive $P(X, W)$. As formalized in the proof and developed more in Section 4.5, the chain rule for convex functions plays a central role in relating the stages.

These are envelope theorems, familiar from static and dynamic discrete choice (McFadden, 1981; Rust, 1988). Classic analysis imposes a density assumption to ensure no utility ties in a one step discrete choice setting (McFadden, 1981). If we interpret P as quantities, and $u(X)$ as a negative price vector, then (iii) is a version of Roy's identity from demand (noting the absence of income effects).

Remark 3 (Identification of $(u, \overline{V})$). Theorem 2(iii) only references choice probabilities concerning observables and not the latent action A. Appendix C outlines how

[12] See Galichon and Salanié (2022) for a related decomposition in a matching context.

Theorem 2(iii), with extra differentiability restrictions, can be used to nonparametrically identify u and differences in $\overline{V}$ using Slutsky symmetry as in Allen and Rehbeck (2019a). We need each alternative k to have at least one continuous regressor that is excluded from the other utility indices (so in particular, $X_k \neq X_j$ generally), K location assumptions such as $u_k(0) = 0$ for each k, and 1 scale assumption such as $u_1(\overline{X}_1) - u_1(0) = 1$ for a reference value $\overline{X}_1$.

4.3 Dual Envelopes

Dual versions of the envelope results are available, which may be interpreted as a type of first order condition concerning optimal probabilities. To state them, let $\overline{V}^{F*}(\cdot \mid W)$ be the convex conjugate of $\overline{V}^{F}(\cdot \mid W)$, treating W as a parameter, and similarly for $\overline{V}_A^{S*}$ and $\overline{V}^{*}$.

Theorem 3. *The following relations hold for 2S-ARUM:*

(i) $u(X) \in \partial_v \overline{V}_A^{S*}(P^S(X, W, A) \mid X, W) \qquad \forall A \in \mathcal{A}.$

(ii) If ξ *is degenerate,* $\overline{V}^{S}(u(X) \mid X, W) \in \partial_\tau \overline{V}^{F*}\left(P^F(X, W) \mid W\right).$

(iii) $u(X) \in \partial_v \overline{V}^{*}\left(P(X, W) \mid W\right).$

The effective domains satisfy $\operatorname{dom}(\overline{V}_A^{S*}(\cdot \mid X, W)) \subseteq \Delta^{K-1}$ *for each* $A \in \mathcal{A}$, $\operatorname{dom}(\overline{V}^{F*}(\cdot \mid W)) \subseteq \Delta^{|\mathcal{A}|-1}$, *and* $\operatorname{dom}(\overline{V}^{*}(\cdot \mid W)) \subseteq \Delta^{K-1}$.

Related inversion results have been used for identification and estimation previously in Hotz and Miller (1993), Berry (1994), Chiong et al. (2016), Galichon and Salanié (2022), and Allen and Rehbeck (2019a), among others. Extending these ideas, suppose a researcher fully specifies the latent pieces: the first and second stage ε_k and C functions, and the distribution of latent shocks (but not A directly). Specifying these objects then also specifies *all* of the value functions here, as well as their convex conjugates.

Unlike the value functions, which are always finite here, the conjugates are ∞ for probabilities off the simplex. This introduces a technical wrinkle for inverting the results: the conjugates here are never fully differentiable. Instead, each ∂ in Theorem 3 includes a one-dimensional affine space because adding a constant to each utility index

does not alter the optimal choice probability.[13] The analogue of differentiability then (for inversion) is to assume that we have exactly a one-dimensional space, and then we can identify utility indices by a location assumption.

Proposition 2. *(i) Given $(x, w) \in \text{supp}(X, W)$ and known $P(x, w)$ in the relative interior of Δ^{K-1} and known $\overline{V}^*$, and assuming $\partial_v \overline{V}^*(P(x, w) \mid w)$ is a one dimensional affine space, we identify $u_j(x) - u_k(x)$ for any j, k. Thus we can identify u for a specified x by either of the following location assumptions: (a) $\sum_{k=1}^K u_k(x_k) = 0$, (b) $u_1(x_1) = 0$.*

(ii) Given $u(X)$ and known differentiable $\overline{V}_A^S$, we identify P^S through

$$P^S(X, W, A) = \nabla_v \overline{V}_A^S(u(X) \mid X, W).$$

(iii) Assume ξ is degenerate. Given $u(X)$, $\overline{V}^S = (\overline{V}_A^S)_{A \in \mathcal{A}}$, and known differentiable $\overline{V}^F$, we identify P^F through

$$P^F(X, W) = \nabla_\tau \overline{V}^F \left(\overline{V}^S(u(X)) \mid W \right).$$

Part (i) is an injectivity result, stating that the only way to keep choice probabilities fixed is to add c to *each* utility index. An alternative way to obtain injectivity in terms of a derivative condition is to reparametrize and consider derivatives in directions that keep us on the probability simplex, as in see Section S.6 in Allen and Rehbeck (2019b). In the one stage model, injectivity is related to the distribution of shocks being connected (cf. Corollary 1 in Norets and Takahashi (2013)). See Sørensen and Fosgerau (2022) for comprehensive analysis of the inverse mapping in the classic one step model.

The purpose of Proposition 2 is to connect with existing literature that specifies a value function and then inverts it. Note that this invertibility approach does not use exclusion restrictions concerning which parts of X enter which u_k utility index terms, and further all covariates here can be discrete. An important caveat of this approach is that we must specify the precise shape of the value functions, and to decompose the stages here we must specify both value functions. Indeed, part (i)

[13] Compare with Theorem 3 in Chiong et al. (2016), in the context of dynamic discrete choice.

is an exact identification results when we do not use exclusion restrictions and have general nonparametric u. In this sense, the decomposition is similar to a normalization where we *define* the u by specifying the other pieces.

For a simple logit example for part (i) that applies with degenerate W, consider $\overline{V}^{\Lambda *}(p) = -\sum_{k=1}^{K} p_k \ln p_k + \iota_{\Delta^{K-1}}(p)$. Here, the convex indicator $\iota_{\Delta^{K-1}}(p)$ encodes the constraint, and is 0 for $p \in \Delta^{K-1}$ and ∞ otherwise. Inverting this gives $\ln(P_k(X)) - \ln(P_1(X)) = u_k(X_k) - u_1(X_1)$, e.g. Hotz and Miller (1993), Berry (1994). See Fosgerau et al. (2024) for a more general class of parametric models in the PUM class that can readily be inverted.

4.4 From Envelopes to Maximization

A dual perspective to build theory-consistent choice probabilities is to take (sub)derivatives of a convex value function in the spirit of McFadden (1981). It is natural then to wonder what utility matches the envelope relations. The answer is below. Here, we use the inner product notation $\langle p, u(X) \rangle = \sum_{k=1}^{K} p_k u_k(X_k)$.

Theorem 4. *The following relations hold for 2S-ARUM:*

$$
\begin{aligned}
P^S(X,W,A) &\in \arg\max_{p \in \Delta^{K-1}} \Big\{ \langle p, u(X) \rangle - \overline{V}^{S*}_A(p \mid X, W) \Big\} \qquad \forall A \in \mathcal{A}, \\
P^F(X,W) &\in \arg\max_{\pi \in \Delta^{|\mathcal{A}|-1}} \Big\{ \langle \pi, \overline{V}^S(u(X) \mid X, W) \rangle - \overline{V}^{F*}(\pi \mid W) \Big\} \text{ if } \xi \text{ is degenerate}, \\
P(X,W) &\in \arg\max_{p \in \Delta^{K-1}} \Big\{ \langle p, u(X) \rangle - \overline{V}^{\,*}(p \mid W) \Big\}, \qquad (2) \\
u(X) &\in \arg\max_{v \in \mathbb{R}^K} \Big\{ \langle v, P^S(X,W,A) \rangle - \overline{V}^S_A(v \mid X, W) \Big\} \qquad \forall A \in \mathcal{A}, \\
\overline{V}^S(u(X) \mid X, W) &\in \arg\max_{\tau \in \mathbb{R}^{|\mathcal{A}|}} \Big\{ \langle \tau, P^F(X,W) \rangle - \overline{V}^F(\tau \mid W) \Big\} \text{ if } \xi \text{ is degenerate}, \\
u(X) &\in \arg\max_{v \in \mathbb{R}^K} \Big\{ \langle v, P(X,W) \rangle - \overline{V}(v \mid W) \Big\}.
\end{aligned}
$$

Relation 2 shows why 2S-ARUM implies PUM, once we set $D(p,W) = -\overline{V}^*(p \mid W)$, establishing the (i) $\implies$ (ii) part of Theorem 1. Galichon and Salanié (2022) present related duality results in a one stage matching context. The effective domain results in Theorem 3 state that the conjugates are finite only over a subset of the simplex. This means that the constraints, e.g. $p \in \Delta^{K-1}$, are technically redundant.

4.5 Chain Rule and Law of Total Probability

Recall that $\overline{V}$ is an explicit composition of the first and second stage value functions when we condition on ξ (Equation 1). When ξ is degenerate, we obtain from Lemma 1 the simplification

$$\overline{V}(u(X) \mid W) = \overline{V}^F(\overline{V}^S(u(X)) \mid W).$$

We can see from this a close connection between the chain rule applied to the ex ante value function $\overline{V}$, and the law of total probability. I first state the equivalence with differentiability.

Proposition 3. *Assume $\overline{V}$, $\overline{V}^F$, and $\overline{V}^S$ are differentiable and ξ is degenerate. We have the following relations:*

$$\begin{aligned} P^S(X, W, A) &= \nabla_v \overline{V}^S_A(u(X)) \qquad \forall A \in \mathcal{A} \\ P^F(X, W) &= \nabla_\tau \overline{V}^F\left(\overline{V}^S(u(X)) \mid W\right) \\ P(X, W) &= \nabla_v \overline{V}(u(X) \mid W). \end{aligned}$$

Given these relations, the following are equivalent:

$$\begin{aligned} P(X, W) &= \sum_{A \in \mathcal{A}} P^F_A(X, W)\, P^S(X, W, A) \\ \nabla_v\, \overline{V}(u(X) \mid W) &= \sum_{A \in \mathcal{A}} \nabla_{\tau_A}\, \overline{V}^F\big(\overline{V}^S(u(X)) \mid W\big)\, \nabla_v\, \overline{V}^S_A(u(X)). \end{aligned}$$

Equivalence here means one result directly implies the other, without other machinery.

The law of total probability has been extensively used to tackle measurement questions, e.g. Manski (1989). As formalized through the chain rule, here we *require* that the stages are linked through closely related maximization problems, as in standard dynamic programming models (Rust, 1987). There are two step models that use different structure. In the special cases of consideration set models, Abaluck and Adams-Prassl (2021) use deviations from Slutsky symmetry to identify how consideration sets change with covariates. That approach does not apply here, because 2SM satisfies the Slutsky restrictions as seen from $P(X, W) = \nabla_v \overline{V}(u(X) \mid W)$; recall that $\overline{V}$ is convex and thus must have a positive semidefinite and symmetric Hessian.

When we drop differentiability, we have the following analogue.

Theorem 5 (Chain Rule)**.** *If ξ is degenerate, we have*

$$\begin{aligned}
&\partial_v \overline{V}(u(X) \mid W) \\
&\quad = \left\{ \sum_{A=1}^{|\mathcal{A}|} p_A^F p^{S,A} \mid p^F \in \partial_\tau \overline{V}^F(\overline{V}^S(u(X)) \mid W), p^{S,A} \in \partial_v \overline{V}_A^S(u(X)), \forall A \in \mathcal{A} \right\} \\
&\quad = \bigcup_{p^F \in \partial_\tau \overline{V}^F(\overline{V}^S(u(X)))} \left(\sum_{A=1}^{|\mathcal{A}|} p_A^F \partial_v \overline{V}_A^S(u(X)) \right).
\end{aligned}$$

5 Piecewise Linear Value Functions

The multi-stage model can be parametrized through the ex ante value function $\overline{V}$, as shown in Theorem 2. This section presents a parametrization of the value function in terms of finitely many maxima that applies with finite support of covariates. I use this to show how the cardinality ("flexibility") of the latent action $\mathcal{A}$ restricts the form of the value function, providing a clear relationship between the classic one stage additive random utility model McFadden (1981) and the two stage model.

As a warmup, we start with the 1-stage additive random utility model, which is a special case of the 2-stage model restricted so that $\mathcal{A}$ a singleton.[14] To state the result, let $\mathcal{V}^{1SM}(J)$ be the set of piecewise linear functions of the form:

$$\overline{V}^{1SM}(v) = \sum_{j=1}^{J} p^j \max_k \{v_k + c_k^j\},$$

where for each j, $p^j \in \Delta^{J-1}$, $c^j \in \mathbb{R}^K$, and $c_0^j = 0$. The coefficients in $\overline{V}^{1SM}$ describe a distribution over the additive shocks, allowing an unspecified tie-breaking rule. Then p^j describes the probability that the shock vector equals $(c_1^j, \ldots, c_K^j)$. We see that $\overline{V}^{1SM}$ is the familiar expected maximum, or "Emax" formulation, calculated for this discrete distribution of shocks. For comparability, W is degenerate because it is not present in the classic model.

[14] Technically, $|\mathcal{A}| = 1$ should be called the extended additive random utility model, since we work in the extended reals. Here, $\text{supp}(X)$ is finite and so from Proposition 1(ii) in Allen (2026), it is without loss of generality for the shocks to all be finite.

Proposition 4. *Assume W is degenerate and* $\text{supp}(X)$ *is finite. The following are equivalent:*

(i) $P(X)$ is consistent with the 1-stage additive random utility model with utility indices u.

(ii) There is a J and $\overline{V}^{1SM} \in \mathcal{V}^{1SM}(J)$ such that for each $x \in \text{supp}(X)$,

$$P(x) \in \partial \overline{V}^{1SM}(u(x)).$$

Moreover, it is without loss to set $J \leqslant |\text{supp}(X)|(K-1)+1$.

Previous literature has established that the additive random utility model can be described by a finite mixture when covariates have finite support and there are no utility ties (Koning and Ridder, 2003; Tebaldi et al., 2023).[15] More broadly, a common approach in the identification literature of choice models is to enumerate choice patterns and then tie these to identified moments through a finite mixture formulation (Honoré and Tamer, 2006; Manski, 2007; Kitamura and Stoye, 2018; Gu et al., 2025; Mbakop, 2023).[16] Dedicated hypothesis testing methods are now available (Kitamura and Stoye, 2018; Fang et al., 2023). Proposition 4 is a dual formulation, showing that we can work in indirect utility space as in classical discrete choice (McFadden, 1981). One feature here is that the scaling of coefficients is *linear* in the support of X, as seen in the bound $J \leqslant |\text{supp}(X)|(K-1)+1$, which follows from Caratheodory's Theorem. This bound describes the number of active types needed to rationalize the data, which is smaller than the number of choice patterns consistent with the model as in the minimal relevant partition (Koning and Ridder, 2003; Tebaldi et al., 2023).

Turning to two stage models, we can first examine how the flexibility of the 2-stage model varies with the cardinality of the first stage action set, denoted $|\mathcal{A}|$. That is, here we consider a subset of the 2SM that restricts the number of latent actions. Define now $\mathcal{V}^{Q-2SM}(J)$ as the set of piecewise linear functions of the form

$$\overline{V}^{Q-2SM}(v) = \max\left\{\overline{V}^{S}_{A_1}(v), \ldots, \overline{V}^{S}_{A_Q}(v)\right\},$$

[15] For the general random utility model, see Falmagne (1978) and McFadden and Richter (1990). Yang and Kopylov (2023) study empirical content of a version of ARUM; see Equation 14 for a linear programming formulation.

[16] For earlier work without explicit choice models, see Balke and Pearl (1994, 1997).

where each $\overline{V}^S_{A_q} \in \mathcal{V}^{1SM}(J)$. These value functions suggest the interpretation of choosing an action A_q in the first stage, where the action is a distribution of shocks in the second stage. It is helpful to connect with distributionally robust discrete choice models, as studied in Natarajan et al. (2009), Feng et al. (2017), and Müller et al. (2023), among others.

Definition 2 (Distributionally Robust Discrete Choice)**.** *$P(X)$ is consistent with the distributionally robust model (DRM) with utility indices u if the following condition holds: For every $x \in \operatorname{supp}(X)$,*

$$P(x) \in \partial \overline{V}^{DR}(u(x)),$$

where $\overline{V}^{DR}$ is everywhere finite, and satisfies $\overline{V}^{DR}(u(x)) = \max_{H \in \mathcal{H}} \mathbb{E}[\max_k \{u_k(x) + \zeta_k\}]$ for some class of distributions $\mathcal{H}$.

This model can be interpreted as seeing X, then choosing an H, then drawing $\zeta \sim H$, then making a final discrete choice. Hence, the distribution of shocks is endogenous, but the set of candidates $\mathcal{H}$ is exogenous. This model is obviously a special case of 2SM, which can be seen by setting $A = H$, $\mathcal{A} = \mathcal{H}$, and (ξ, ω) degenerate.

Proposition 5. *Assume W is degenerate and $\operatorname{supp}(X)$ is finite. For any natural number Q, the following are equivalent:*

(i) $P(X)$ is consistent with the 2-stage additive random utility model with utility indices u, $|\mathcal{A}| = Q$ possible actions in the first stage, and degenerate (ξ, ω).

(ii) $P(X)$ is consistent with the distributionally robust model with utility indices u, and $|\mathcal{H}| = Q$ possible distributions of shocks.

(iii) There is a J and $\overline{V}^{Q-2SM} \in \mathcal{V}^{Q-2SM}(J)$ such that for each $x \in \operatorname{supp}(X)$,

$$P(x) \in \partial \overline{V}^{Q-2SM}(u(x)).$$

Moreover, it is without loss to set $J \leqslant |\operatorname{supp}(X)|(K-1)+1$.

The focus of this paper is on settings in which the first stage is a flexible "action" meant to capture a range of phenomena, and in particular the cardinality of the possible latent actions is unrestricted. It is not possible to enumerate types describing

the first and second stages when $\mathcal{A}$ has unrestricted size. Enumerating would require fleshing out the first stage explicitly and describing explicitly what $\mathcal{A}$ is, while the focus of this paper is on treating the first stage in a general way that is not explicitly modeled. Hence we need new methods.

For our purposes, the key analogue of the above results is below. First define $\mathcal{V}^{2SM}(J)$ as the set of all functions of the form

$$\overline{V}^{2SM}(v) = \max_{j \in \{1,\dots,J\}} \{p^{j\prime} v + c^j\},$$

where each $p^j \in \Delta^{K-1}$, and each $c^j \in \mathbb{R}$. A key finding is that we have fewer parameters to consider when we let Q be large. That is, by convexifying the model further we have a framework that is in some sense more parsimonious.

The set $\mathcal{V}^{2SM}(J)$ has several useful properties: each element is weakly increasing, convex, translation invariant, 1-Lipschitz in $\|\cdot\|_\infty$, finite-valued, and has all subgradients within the probability simplex. I say such functions are "well-behaved" and below establish it is without loss of generality to consider well-behaved value functions. This is a common finding in the revealed preference literature (Varian, 1982). Framed differently, these extra assumptions are not testable here. Here we re-introduce W since it is present in 2SM.

Theorem 6. *Assume* $\operatorname{supp}(X, W)$ *is finite. For any natural number* $Q \geqslant |\operatorname{supp}(X, W)|$*, the following are equivalent:*

(i) $P(X, W)$ *is consistent with the 2-stage additive random utility model with utility indices* u*,* $|\mathcal{A}| = Q$ *possible latent actions, degenerate* (ξ, ω)*, and* $P^F(A \mid X, W) \in \{0, 1\}$ *for each* $A \in \mathcal{A}$*.*

(ii) $P(X, W)$ *is consistent with the 2-stage additive random utility model with utility indices* u *(without restricting* $\mathcal{A}$*).*

(iii) For each $w \in \operatorname{supp}(W)$*, there is a convex function* $V^w : \mathbb{R}^K \to \mathbb{R}$ *such that for each* $x \in \operatorname{supp}(X \mid W = w)$*,*

$$P(x, w) \in \partial V^w(u(x)).$$

(iv) *For each* $w \in \text{supp}(W)$*, there is a* J^w *and* $\overline{V}^{2SM} \in \mathcal{V}^{2SM}(J^w)$ *such that for each* $x \in \text{supp}(X \mid W = w)$*,*

$$P(x, w) \in \partial \overline{V}^{2SM,w}(u(x)).$$

Moreover, it is without loss of generality to set $J^w \leqslant |\text{supp}(X \mid W = w)|$*.*

(v) *For each* $w \in \text{supp}(W)$ *and for every cycle* $x^1, \ldots, x^L, x^{L+1} \in \text{supp}(X \mid W = w)$ *(where* $x^{L+1} = x^1$*),*

$$\sum_{\ell=1}^{L} P(x^\ell, w)'(u(x^{\ell+1}) - u(x^\ell)) \leqslant 0.$$

(vi) *For each* $w \in \text{supp}(W)$*, there exist finite-valued scalars* $\{\tau_x^w\}_{x \in \text{supp}(X|W=w)}$ *such that for every* $x, \tilde{x} \in \text{supp}(X \mid W = w)$*, we have*

$$\tau_{\tilde{x}}^w \geqslant \tau_x^w + P(x, w)'(u(\tilde{x}) - u(x)).$$

For checking model consistency, I recommend part (vi) because it is linear program that is tractable in moderate sized problems. Directly checking (v) in contrast would require a very large number of cycles.

Part (i) shows that "enough" latent actions can fully span the model, when the actions enter in a flexible way. This is important empirically, because if we interpret the distributions above as empirical distributions (as would arise in estimation problems), the theorem establishes that we need at most one action per observation if we use a flexible class of indirect utility such as $\mathcal{V}^{2SM}$. Thus, by integrating out the latent actions and studying the generaly structure, we entirely avoid the challenging combinatorial natural of certain search and consideration set models (for other approaches, see Aguiar and Kashaev (2025) and Chib and Shimizu (2025)). The degenerate case in (i) is immediately equivalent to DRM upon relabeling, with the generalization that the set of distributions $\mathcal{H}$ must depend on w. Part (iii) formalizes that the envelope theorem provides a complete account of the general two stage model when we only observe the overall choice probabilities P. Note that in (iv), when W is degenerate we can restrict $J \leqslant |\text{supp}(X)|$ whereas the one stage model requires the bounds

$J \leqslant |\operatorname{supp}(X)|(K-1)+1$, which increases with the number of alternatives K. From (iv), we see that 2SM can be parametrized with a $\overline{V}^{2SM}$ that is convex in the (p, c) parameters. In contrast, the other models above have terms like $p^j \max_k\{v_k + c_k^j\}$, which is not jointly convex in the parameters. The shape restriction (v) is a type of cyclic monotonicity, as has been used in Shi et al. (2018) in cross-sectional and panel discrete choice without latent actions. We can invert (v) to fully characterize the identified set for u. Pakes and Porter (2024) show that the cyclic monotonicity inequalities are not sharp in general for the classic additive random utility model (i.e. they do not exhaust empirical content, or equivalently do not exhaust information on admissible u). The inequalities are sharp for the two stage model studied here. The subgradient inequalities in (vi) are analogous to those in quasilinear demand (Browning, 1989; Brown and Calsamiglia, 2007) and PUMs (McFadden and Fosgerau, 2012; Allen and Rehbeck, 2023), which are further related to the Afriat inequalities (Afriat, 1967).

The equivalence of (i) and (iii) is related to Theorem 1 in Feng et al. (2017), which formalizes equivalence of value functions relating PUM and the distributionally robust model.[17] The equivalence of (iii)-(vi) uses a construction due to Rockafellar (2015).

Remark 4 (Minimal Q). We have that $Q = 1$ is the one step additive random utility model, and with $Q = |\operatorname{supp}(X)|$ we have the general 2SM. The minimal Q^* can be interpreted as a measure of deviations from the one stage model, or a measure of action-complexity needed in the first stage. We can calculate Q^* as the smallest integer such that the conditions in Proposition 5 hold.

Remark 5 (ReLU Parametrizations). The rectified linear (ReLU) function $g(v) = \max\{0, v\}$ for scalar v has been extensively used to parametrize artificial neural networks. This section does not parametrize probabilities directly, but instead parametrizes the value functions, through input-convex neural networks related to the analysis of Amos et al. (2017). The $\overline{V}$ parametrizations in this section are all formed by affine compositions of ReLU, and the resulting $\overline{V}$ are input-convex because they consist of compositions of weakly increasing convex functions, noting that the weights p are always nonnegative. This is related to Grzeskiewicz (2025), which parametrizes *utility functions* through input-concave neural networks, in a deterministic demand

[17] Feng et al. (2017) call the distributionally robust model the semiparametric choice model following Natarajan et al. (2009), and call PUM the representative agent model.

setting. Nguyen (2025) uses neural networks to approximate value functions in a dynamic discrete choice model.

Remark 6 (Variation in W). Theorem 6(iv) can be written as stating that we can study value functions of the form $\overline{V}^{2SM}(v;w) = \max_{j\in\{1,\ldots,J\}}\{p^j(w)'v_j + c^j(w)\}$, reflecting dependence on w. The restriction here is that $p^j(w)$ must always be on the probability simplex. In practice, one way to ensure this is to build a logit (or softmax) model for the p^j weights, where w enters through a linear index. For example, consider $p^j(w) = \frac{e^{\gamma^{j\prime}w^j}}{\sum_{\ell=1}^{J} e^{\gamma^{\ell\prime}w^\ell}}$, where w is partitioned into j-specified shifters.

6 Counterfactual and Welfare Bounds

Section 5 just presented empirical content characterizations given u, which immediately characterize sharp bounds on candidate probabilities at new values of covariates that are not seen, again conditional on u.[18] With known u, a linear programming formulation of sharp bounds on welfare and counterfactual objects is characterized for quasilinear utility in Allen and Rehbeck (2020a), and those results can then be applied to PUM (and hence also 2SM). The translation is that u becomes negative price, and probabilities become quantities. For 2SM, this section generalizes those results to one-step bounds when we do not know u, but instead have a linear index structure $u_k(X_k) = \beta_k' X_k$. Overall though, the core logic here is that conditional on utility indices we can obtain sharp bounds on welfare/counterfactuals, and we have sharp bounds on β by using Theorem 6.[19] When we combine these two steps into one, we obtain an optimization-based formulation of the bounds.

For counterfactual analysis, we require that β is not identically zero. The reason is that we allow utility ties, and $\beta = 0$ can always be rationalized, in which case the model delivers no nontrivial counterfactual predictions at new values of covariates. For welfare analysis, we specify the scale of β so that we can state welfare in units that are specified ex ante. To accomplish both tasks we specify $\beta_{1,1} = 1$, i.e. the first regressor of the first alternative is known to be relevant, and with a known

[18] Tebaldi et al. (2023) present similar bounds when u is negative price, in 1SM, using instrumental variables. Cyclic monotonicity based bounds with a given u are presented in Allen and Rehbeck (2019a) and Chiong et al. (2021).

[19] Shi et al. (2018) characterize the identified set for slopes when using cyclic monotonicity. That paper has the same β_k for each across alternative, and focuses on a panel setup.

magnitude. If price is somehow exogenous and the first regressor, we would instead use -1, and welfare would be in units of dollars (presuming price does not otherwise enter into X as a squared term, for example). The parameter space overall is given by $\beta = (\beta_1', \ldots, \beta_K')'$ with $\beta_{1,1} = 1$, and is denoted Θ. A natural restriction in many contexts is the additional assumption that β_k is the same across alternatives k. Write $u(x, \beta) = (\beta_1' x_1, \ldots, \beta_K' x_K)'$ for the vector of utility indices. The analysis below works with other parametrizations, and in fact u can be in an infinite-dimensional class, but for exposition we focus on the linear index case because it enhances tractability and matches empirical practice.

6.1 Counterfactuals

Consider a new value of covariates (x^*, w^*) where $w^* \in \text{supp}(W)$ but $x^* \notin \text{supp}(X \mid W = w^*)$. Let $P^*(x^*, w^*)$ denote the counterfactuals we wish to bound. We bound objects like $\lambda' P^*(x^*, w^*)$, where λ is a vector describing the direction we wish to bound. A leading example is $\lambda = e_k^K$ for a basis vector, which bounds the k-th probability. Another is λ equal to a price vector of interest, in which case $\lambda' P^*(x^*, w^*)$ is expenditure or revenue. Upper bounds are given by the value function:

$$
\begin{aligned}
&\max_{\tau, \beta \in \Theta,\, p} \quad \lambda' p \\
&\text{s.t.} \quad \tau_{\tilde{x}}^{w} \;\geqslant\; \tau_{x}^{w} + P(x, w)'\big(u(\tilde{x}, \beta) - u(x, \beta)\big), \qquad \forall\, (\tilde{x}, w), (x, w) \in \text{supp}(X, W), \\
&\qquad \tau_{x^*}^{w^*} \;\geqslant\; \tau_{x}^{w^*} + P(x, w^*)'\big(u(x^*, \beta) - u(x, \beta)\big), \qquad \forall\, x \in \text{supp}(X \mid W = w^*), \\
&\qquad \tau_{x}^{w^*} \;\geqslant\; \tau_{x^*}^{w^*} + p'\big(u(x, \beta) - u(x^*, \beta)\big), \qquad \forall\, x \in \text{supp}(X \mid W = w^*), \\
&\qquad\quad p \in \Delta^{K-1}.
\end{aligned}
$$

Lower bounds replace the max with a min. Throughout this section we assume the support of (X, W) is finite so that we have a finite-dimensional optimization problem.

In general, all pairs of support points that share the same w discipline β, as reflected in the first inequalities. Given a candidate β though, the quantity bounds only involve covariates that match the counterfactual value w^*, as formalized in the middle two inequalities. In particular, the bounds at new values of w outside the support are always trivial even if we knew β, because the model does not restrict how w enters.

These bounds are generally not linear programs because of the interaction $p'(u(x,\beta) - u(x^*,\beta))$, which involves the product of p and the β pieces. Hence this part is bilinear. With this said, for *fixed* β, the problem above would be a linear program. Thus if we only had price and it enters in a known way (cf. the analysis of 1SM in Tebaldi et al. (2023)), then counterfactual bounds would be a linear program. The unknown β case brings us closer to Gu et al. (2025), but that paper does not obviously apply here because we cannot enumerate the two stage model in general when $\mathcal{A}$ has unknown size.

6.2 Welfare

For 2SM, when $(x^0, w^*), (x^1, w^*) \in \text{supp}(X, W)$, we can bound the welfare change when we change from $x^0 \to x^1$, conditional on w^*. The upper bound on our welfare object $\overline{V}(u(x^1,\beta), w^*) - \overline{V}(u(x^0,\beta), w^*)$ is given by the value of the following linear program:

$$\begin{aligned} \max_{\tau,\beta\in\Theta} \quad & \tau_{x^1}^{w^*} - \tau_{x^0}^{w^*} \\ \text{s.t.} \quad & \tau_{\tilde{x}}^{w} \geqslant \tau_{x}^{w} + P(x,w)'\big(u(\tilde{x},\beta) - u(x,\beta)\big), \qquad \forall\, (\tilde{x},w),(x,w) \in \text{supp}(X,W). \end{aligned}$$

The minimum over the same problem yields the sharp lower bounds. Thus we arrive at direct bounds on welfare. All w serve to help bound β, but for a fixed β, only values with $w = w^*$ serve to bound the welfare change. Unlike the other examples in this section, this is a linear program even when we do not know β, because we never have two unknowns multiplying.

For general changes in covariates outside the support, we need to modify the program. We maintain that $w^* \in \text{supp}(W)$ but now consider general x^0 and x^1 values. The

new bounds are given by:

$$\begin{aligned}
&\max / \min_{\tau, \beta \in \Theta,\, p^0,\, p^1} \quad \tau_{x^1}^{w^*} - \tau_{x^0}^{w^*} \\
&\quad \text{s.t.} \quad \tau_{\tilde{x}}^{w} \;\geqslant\; \tau_{x}^{w} + P(x,w)'\big(u(\tilde{x},\beta) - u(x,\beta)\big), \qquad \forall\, (\tilde{x},w),(x,w) \in \operatorname{supp}(X,W), \\
&\qquad\quad \tau_{x^\ell}^{w^*} \;\geqslant\; \tau_{x}^{w^*} + P(x,w^*)'\big(u(x^\ell,\beta) - u(x,\beta)\big), \qquad \forall\, x \in \operatorname{supp}(X \mid W = w^*),\ \ell \in \{0,1\}, \\
&\qquad\quad \tau_{x}^{w^*} \;\geqslant\; \tau_{x^\ell}^{w^*} + (p^\ell)'\big(u(x,\beta) - u(x^\ell,\beta)\big), \qquad \forall\, x \in \operatorname{supp}(X \mid W = w^*),\ \ell \in \{0,1\}, \\
&\qquad\quad \tau_{x^1}^{w^*} \;\geqslant\; \tau_{x^0}^{w^*} + (p^0)'\big(u(x^1,\beta) - u(x^0,\beta)\big), \\
&\qquad\quad \tau_{x^0}^{w^*} \;\geqslant\; \tau_{x^1}^{w^*} + (p^1)'\big(u(x^0,\beta) - u(x^1,\beta)\big), \\
&\qquad p^0,\ p^1 \in \Delta^{K-1}.
\end{aligned}$$

The inequality constraints encode that each in-support P and out of support p^0 and p^1 must be optimal. As with the counterfactuals, this setup is not a linear program because of the interaction of the p^ℓ and β pieces.

Appendix A Proofs of Results in Main Text

Proof of Theorem 1. The implication (i) $\implies$ (ii) is obtained from Theorem 3, setting $D(p,W) = -\overline{V}^*(p \mid W)$.

To prove that (ii) $\implies$ (i), start with the PUM representation. Fix $w \in \operatorname{supp}(W)$ and consider $x, \tilde{x} \in \operatorname{supp}(X \mid W = w)$. Since $D(p,w) > -\infty$ for some $p \in \Delta^{K-1}$, we have that the maximizers attain finite utility and hence $D(P(x,w),w)$ and $D(P(\tilde{x},w),w)$ are both finite. Define

$$V^w(v) = \max_{p \in \Delta^{K-1}} p'v + D(p,w).$$

This is finite for v equal to $u(x)$ or $u(\tilde{x})$. We have

$$V^w(u(x)) = P(x,w)'u(x) + D(P(x,w),w),$$

and also

$$V^w(u(\tilde{x})) \geqslant P(x,w)'u(\tilde{x}) + D(P(x,w),w).$$

Combining these and rearranging yields

$$V^w(u(\tilde{x})) \geqslant V^w(u(x)) + P(x,w)'(u(\tilde{x}) - u(x)).$$

This takes the form

$$\tau^w_{\tilde{x}} \geqslant \tau^w_x + P(x,w)'(u(\tilde{x}) - u(x)).$$

This holds for any $w \in \text{supp}(W)$ and any $x, \tilde{x} \in \text{supp}(X \mid W = w)$. We match Theorem 6 condition (vi), which implies 2SM. Further, from Theorem 6 it is without loss of generality to set (ξ, ω) degenerate and set $P^F(A \mid X, W) \in \{0,1\}$ for each $A \in \mathcal{A}$. □

Proof of Proposition 1. The equivalence of (i) and (ii) is a special case of Theorem 1.

Next we prove that (i) and (iv) are equivalent. From Theorem 6, we have that conditional on $W = w$, part (i) is equivalent to cyclic monotonicity, which is equivalent to monotonicity from p. 240 in Rockafellar (2015) using the fact that probabilities sum to 1. In more detail, write $P(x,w) = (P_1(x,w), 1 - P_1(x,w))'$ so that each cycle in Theorem 6(vi) can be rewritten as:

$$\sum_{\ell=1}^{L} P(x^\ell, w)'(u(x^{\ell+1}) - u(x^\ell)) = \sum_{\ell=1}^{L} P_1(x^\ell, w)\Big(h(x^{\ell+1}) - h(x^\ell)\Big),$$

where $h(x) = (u_1(x_1) - u_2(x_2))$. This rewrites so that scalar-valued $P_1(\cdot, w)$ is cyclically monotone in h, which again from p. 240 in Rockafellar (2015) is equivalent to monotonicity in h since P_1 is scalar. Since this holds for any w, we have that in general (i) and (iv) are equivalent.

Next we show that (iii) and (iv) are equivalent, regardless of whether $\text{supp}(X, W)$ is finite. First we show that (iii) implies (iv). Condition on $w \in \text{supp}(W)$ and let $\zeta := \eta_2 - \eta_1$. Here, set $\zeta = 0$ when $\eta_2 = -\infty$ and $\eta_1 = -\infty$. Recall $P(\eta_1 = -\infty, \eta_2 = -\infty \mid W = w) = 0$. Under (iii), for each $x \in \text{supp}(X \mid W = w)$ there exists $\tau(x,w) \in [0,1]$ such that

$$P_1(x,w) = \Pr(\zeta < h(x) \mid W = w) + \tau(x,w) \Pr(\zeta = h(x) \mid W = w).$$

Take $x, \tilde{x}$ with $h(\tilde{x}) > h(x)$. Then

$$P_1(x,w) \leqslant \Pr(\zeta \leqslant h(x) \mid W = w) \leqslant \Pr(\zeta < h(\tilde{x}) \mid W = w) \leqslant P_1(\tilde{x},w).$$

So $P_1(\tilde{x},w) - P_1(x,w) \geqslant 0$ whenever $h(\tilde{x}) - h(x) > 0$. The same argument with x and $\tilde{x}$ swapped gives $P_1(\tilde{x},w) - P_1(x,w) \leqslant 0$ whenever $h(\tilde{x}) - h(x) < 0$. Therefore

$$\big(P(\tilde{x},w) - P(x,w)\big)\big(h(\tilde{x}) - h(x)\big) \geqslant 0.$$

Finally, since

$$P(\tilde{x},w) - P(x,w) = \big(P_1(\tilde{x},w) - P_1(x,w), -(P_1(\tilde{x},w) - P_1(x,w))\big)',$$

we have

$$(P(\tilde{x},w) - P(x,w))'(u(\tilde{x}) - u(x)) = \big(P_1(\tilde{x},w) - P_1(x,w)\big)\big(h(\tilde{x}) - h(x)\big) \geqslant 0,$$

which is exactly (iv).

To complete the proof we show that (iv) implies (iii), using Lemma 10 in Appendix B.2. To that end, condition on $w \in \text{supp}(W)$ and consider the collection $\{(P_1(x,w), u_1(x_1) - u_2(x_2))\}_{x \in \text{supp}(X)}$. When (iv) holds, this collection forms the graph of a multi-valued monotone mapping (see Appendix B.2), and hence from Lemma 10 there exists a distribution H^w over extended real-valued ζ, and tie-breaking rule $\tau(x,w)$ such that

$$P_1(x,w) = \Pr_{H^w}(\zeta < u_1(x_1) - u_2(x_2)) + \tau(x,w) \Pr_{H^w}(\zeta = u_1(x_1) - u_2(x_2)).$$

This is just the binary choice model in the extended reals (so that the CDF of H^w need not limit to 0 or 1 at extremes), where we allow utility ties. Since we can do this for any w, and we have defined the binary choice model to have an arbitrary distribution conditional on w, we conclude that (iv) implies (iii). □

Proof of Lemma 1. From Assumption 2(ii), ω is independent of (X, W) and so the

$\overline{V}^F$ exclusion restriction holds. Next we see that

$$\begin{aligned}\overline{V}(v \mid X, W) &= \mathbb{E}\left[\overline{V}^F\left(\overline{V}^S(v \mid \xi) \mid W, \omega\right) \mid X, W\right] \\ &= \mathbb{E}\left[\overline{V}^F\left(\overline{V}^S(v \mid \xi) \mid W, \omega\right) \mid W\right] \\ &= \overline{V}(v \mid W),\end{aligned}$$

because (ξ, ω) is independent of (X, W). Again leveraging independence, we can exclude X. When ξ is degenerate, the only remaining randomness in $\overline{V}^S$ is due to η, which is independent of (A, X, W), and so $\overline{V}^S($ exclusion restriction holds.

For the final part, recall when ξ is degenerate we can directly integrate out ω and use the previous exclusion restrictions to get

$$\overline{V}(v \mid W) = \overline{V}^F\left(\overline{V}^S(v) \mid W\right)$$

□

Proof of Lemma 2. Consider the function

$$\phi(v, A, \eta, \xi) = \max_k\{v_k + \varepsilon_k(A, \eta, \xi)\}.$$

We see that ϕ is weakly increasing, convex, translation invariant, and 1-Lipschitz in v. These properties are defined as inequalities, and are preserved under conditional expectations when the expectations exist and are finite, due to monotonicity of conditional expectations (see Theorem 34.2(iii) in Billingsley (2017)). Thus taking expectations yields that $\overline{V}^S_A(v \mid X, W)$ is weakly increasing, convex, translation invariant, and 1-Lipschitz in v, yielding part (i). Part (ii) follows from the same argument. For part (iii), note that $\overline{V}$ is a composition of two weakly increasing convex functions, and hence is convex from Lemma 4(i). Likewise, $\overline{V}$ is weakly increasing, translation invariant, and 1-Lipschitz since it is a composition of two weakly increasing, translation invariant, and 1-Lipschitz functions.

Next we use the exclusion restrictions in Lemma 1 to state the result. □

Proof of Theorem 2. **Proof of part (i).** Define $\psi(v; A, \eta, \xi) = \max_k\{v_k + \varepsilon_k(A, \eta, \xi)\} = \max_{p \in \Delta^{K-1}} \sum_{k=1}^K p_k v_k + \sum_{k=1}^K p_k \varepsilon_k(A, \eta, \xi)$ as a function of v, which

is finite-valued because we assume a maximizer exists that attains finite utility. We have that $\tilde{k} \in \text{argmax}_k\{v_k + \varepsilon_k(A, \eta, \xi)\}$ implies $e^K_{\tilde{k}} \in \partial_v \psi(v; A, \eta, \xi)$; here $\partial_v \psi$ can be interpreted as mixture distributions over the maximizers. Hence:

$$\begin{aligned} P^S(X, W, A, \xi, \omega) &\in \mathbb{E}[\partial_v \psi(u(X); A, \eta, \xi) \mid X, W, A, \xi, \omega] \\ &= \partial_v \mathbb{E}[\psi(u(X); A, \eta, \xi) \mid X, W, A, \xi, \omega] \\ &= \partial_v \overline{V}^S_A(u(X) \mid X, W, \xi). \end{aligned}$$

The first inclusion follows from taking conditional expectations (leveraging that realized k^* satisfies $e^K_{k*} \in \partial_v \psi(u(X); A, \eta, \xi)$). The ∂ interchanges follows from Proposition 2.2 in Bertsekas (1973) (see also Theorem 1 in Sørensen and Fosgerau (2022)). The second equality uses the definition of $\overline{V}^S_A$ and the fact that ω is independent of η. We can use Proposition 2.2 in Bertsekas (1973) again to integrate out (ξ, ω) to obtain

$$P^S(X, W, A) \in \partial_v \overline{V}^S_A(u(X) \mid X, W).$$

Now if in addition conditional on $(A, X, W) = (a, x, w)$ we assume that utility ties are probability 0 events, we have that conditionally, $\partial_v \psi(u(x); a, \eta, \xi) = \{\nabla_v \psi(u(x); a, \eta, \xi)\} = \{e^K_{k*}\}$ almost surely from Theorem 25.1 in Rockafellar (2015). Likewise,

$$\begin{aligned} P^S(x, w, a, \xi, \omega) &= \mathbb{E}[\nabla_v \psi(u(x); a, \eta, \xi) \mid x, w, a, \xi, \omega] \\ &= \nabla_v \mathbb{E}[\psi(u(x); a, \eta, \xi) \mid x, w, a, \xi, \omega] \end{aligned}$$

almost surely, where again the interchange is justified via Proposition 2.2 in Bertsekas (1973). By repeating the steps above, we obtain

$$P^S(x, w, a) = \nabla_v \overline{V}^S_a(u(X) \mid X = x, W = w).$$

Proof of part (ii). By analogous arguments, we obtain when ξ is degenerate that

$$\begin{aligned} P^F(X, W, \omega) &\in \partial_\tau \overline{V}^F(\overline{V}^S(u(X) \mid X) \mid X, W, \omega) \\ &= \partial_\tau \overline{V}^F(\overline{V}^S(u(X)) \mid X, W, \omega). \end{aligned}$$

Hence integrating out ω we obtain

$$P^F(X,W) \in \partial_\tau \overline{V}^F(\overline{V}^S(u(X)) \mid W),$$

where we use Lemma 1 to exclude X from conditioning.

If conditional on $(X,W) = (x,w)$, utility ties are a singleton in the first stage almost surely, then we have conditionally that

$$P^F(x,w,\omega) = \nabla_\tau \overline{V}^F(\overline{V}^S(u(x) \mid x) \mid x,w,\omega)$$

almost surely so that

$$P^F(x,w) = \nabla_\tau \overline{V}^F(\overline{V}^S(u(X) \mid X = x) \mid W = w).$$

Proof of part (iii). From the law of total probability, we have for each k that

$$P_k(X,W,\xi,\omega) = \sum_{A \in \mathcal{A}} P_A^F(X,W,\xi,\omega) P_k^S(X,W,A,\xi,\omega).$$

Recall that

$$\overline{V}(v \mid X,W,\xi,\omega) = \overline{V}^F\left(\overline{V}^S(v \mid X,\xi) \mid X,W,\omega\right).$$

That is, we have a composition of convex and weakly increasing functions. From the law of total probability and the chain rule (Lemma 4),

$$P(X,W,\xi,\omega) \in \partial_v \overline{V}(u(X) \mid X,W,\xi,\omega).$$

Integrate out (ξ,ω) to obtain:

$$P(X,W) \in \partial_v \overline{V}(u(X) \mid X,W).$$

The step uses the interchange $\mathbb{E}[\partial_v \overline{V}(u(X) \mid X,W,\xi,\omega) \mid X,W] = \partial_v \overline{V}(u(X) \mid X,W)$, which is valid from from Proposition 2.2 in Bertsekas (1973). Finally, use Lemma 1 to write $\overline{V}(u(X) \mid X,W) = \overline{V}(u(X) \mid W)$.

When we assume that conditional on $(X,W) = (x,w)$, the first and second stages have

utility ties with probability zero, we have that $\partial_v \overline{V}(u(X) \mid X, W, \xi, \omega)$ is a singleton almost surely, which from Theorem 25.1 in Rockafellar (2015) is the derivative. Hence conditional on $(X, W) = (x, w)$,

$$P(x, w, \xi, \omega) = \nabla_v \overline{V}(u(x) \mid x, w, \xi, \omega)$$

almost surely. Again, integrate out (ξ, ω) and use Lemma 1 (to exclude x) to obtain:

$$P(x, w) = \nabla_v \overline{V}(u(x) \mid W = w).$$

□

Proof of Theorem 3. Parts (i)-(iii) follow from Theorem 2 and Lemma 3(iii).

To complete the proof, recall from Lemma 2, $\overline{V}_A^S$, $\overline{V}^F$, and $\overline{V}$ are each finite-valued, translation invariant, and 1-Lipschitz in $\|\cdot\|_\infty$. Lemma 9 tells us then that the effective domains are in the probability simplex. □

Proof of Proposition 2. For part (i), Theorem 3 tells us $u(x) \in \partial_v \overline{V}^*(P(x, w) \mid w)$, and we see that $u(x) + c1_K \in \partial_v \overline{V}^*(P(x, w) \mid w)$ for any scalar c (here, $c1_K = (c, \ldots, c)'$) because $\overline{V}$ is translation invariant. Thus from the assumption $\partial_v \overline{V}^*(P(x, w) \mid w)$ is a one dimensional affine space, we fully characterize the space via $\{u(x) + c1_K\}_{c \in \mathbb{R}}$. Hence, either (a) or (b) uniquely pin down $u(x)$.

Parts (ii) and (iii) are immediate from Theorem 2, once we apply Theorem 25.1 in Rockafellar (2015) to replace the subdifferential with the derivative. □

Proof of Theorem 4. The bottom three relations follow from Theorem 3 and Lemma 3. The top three relations, except with $p \in \mathbb{R}^K$ and $\pi \in \mathbb{R}^{|\mathcal{A}|}$ as the constraints, follow from Theorem 2 and Lemma 3. The dom relations established in Theorem 3 show that the constraints $p \in \Delta^{K-1}$ and $\pi \in \Delta^{|\mathcal{A}|-1}$ are redundant, e.g. $\overline{V}^*(y \mid W) = \infty$ when $y \notin \Delta^{K-1}$, and hence adding the explicit simplex constraints is without loss. □

Proof of Proposition 3. The first three derivative relations follow from Theorem 2 and the fact that the subdifferential collapses to the derivative under differentiability

(Theorem 25.1 in Rockafellar (2015)).

To show the equivalence, start with the law of total probability. Replace P terms with derivatives. We obtain the chain rule. Likewise, given the chain rule we can replace terms using the envelope results and obtain the law of total probability. □

Proof of Theorem 5. From Lemma 1 we have

$$\overline{V}(v \mid W) = \overline{V}^F(\overline{V}^S(v) \mid W).$$

Lemma 2 yields that $\overline{V}^F$ is convex and weakly increasing, and that $\overline{V}^S$ is convex. All functions are finite-valued. We can thus apply the chain rule, Lemma 4(ii). □

Proof of Proposition 4. First we prove that (ii) implies (i). Define $h(v; c) = \max_k\{v_k + c_k\}$. Since $P(x) \in \partial\overline{V}^{1SM}(u(x))$ by assumption,

$$P(x) \in \partial\overline{V}^{1SM}(u(x)) = \sum_{j=1}^{J} p^j \partial h(u(x); c^j) = \sum_{j=1}^{J} p^j \operatorname{Conv}\left\{e_\ell^K : \ell \in \operatorname*{argmax}_k\{u_k(x_k) + c_k^j\}\right\}.$$

Recall e_ℓ^K is the unit vector with ℓ component 1, and zero elsewhere. The first equality uses Lemma 6, leveraging that $p^j \geqslant 0$ so that each term in the sum is convex, and the fact that subgradients scale as $\partial(p^j h)(v; c^j) = p^j \partial h(v; c^j)$. Note the subgradient of h is with respect to the v argument, while each c^j is treated as a parameter. The second equality uses Lemma 5 to write out $\partial h(v; c)$ as a mixture (convex combination) over maximizers. Overall, the distribution $\{(p^j, c^j)\}_{j=1}^J$ rationalizes $P(x)$ as an additive random utility model, allowing an arbitrary tie-breaking rule when the argmax is not unique.[20]

Now we prove that (i) implies (ii). Since p is consistent with ARUM, we have that choice probabilities are given by

$$P(X) = P^S(X) \in \mathbb{E}[\operatorname{Conv}\{e_k^K \mid k \in M^S(X, \eta)\} \mid X],$$

where we suppress dependence on the first stage action choice A because it is degenerate in ARUM. Likewise, absorb ξ into η because they enter the maximization

[20]Note that at any point where the maximizer is unique, the convex hull part collapes to an indicator for the unique utility-maximizing choice.

problem an identical manner. Further write the tie-breaking probabilities, conditional on covariates and the shock, as

$$P(X,\eta) \in \text{Conv}\{e_k^K \mid k \in M^S(X,\eta)\}.$$

Define now an equivalence relation over the shocks: $\eta \sim \tilde{\eta}$ if $M^S(x,\eta) = M^S(x,\tilde{\eta})$ for each $x \in \text{supp}(X)$. Because the support of X is finite, the set of equivalence classes of shocks is finite, and thus we can rationalize with a measure μ with at most one point for each equivalence class. Write one such distribution as $\{(p^j, c^j)\}_{j=1}^J$ (later we will show a bound on J, but for now we just establish it is finite). Thus for each $x \in \text{supp}(X)$ we can write

$$\begin{aligned} P(x) &= \sum_{j=1}^J p^j \mathbb{E}[P(x,\eta) \mid x, \eta \sim c^j], \\ &\in \sum_{j=1}^J p^j \,\text{Conv}\left\{e_\ell : \ell \in \operatorname*{argmax}_k \{u_k(x_k) + c_k^j\}\right\}. \end{aligned} \tag{3}$$

The inclusion leverages the fact that for all $\eta \sim c^j$, the argmax set is the same for each $x \in \text{supp}(X)$. Combining the inclusion in (3) and the first part, which shows how to write this convex hull as a subgradient of a function of the form of $\overline{V}$, we complete the proof that (i) implies (ii).

The last part is to show that when (ii) holds, it is without loss of generality to set $J \leqslant |\,\text{supp}(X)|(K-1)+1$. Enumerate the support of X to write

$$\overline{p} = \left(p(x^1), \ldots, p\left(x^{|\,\text{supp}(X)|}\right)\right)',$$

where $\overline{p}$ lies in the affine space

$$\left\{\left(z^1, \ldots, z^{|\,\text{supp}(X)|}\right)' \in \mathbb{R}^{|\,\text{supp}(X)|K} : \sum_{k=1}^K z_k^m = 1, m \in \{1, \ldots, |\,\text{supp}(X)|\}\right\}.$$

This has dimension $|\,\text{supp}(X)|(K-1)$, because of the $|\,\text{supp}(X)|$ distinct equality

constraints. Recall we assume in (ii) that

$$P(x) \in \partial \overline{V}^{1SM}(u(x)) = \sum_{j=1}^{J} p^j \partial h(u(x); c^j),$$

so that for each $x^m \in \text{supp}(X)$ and c^j, there is some $d^j(x^m) \in \partial h(u(x^m); c^j)$ such that

$$P(x^m) = \sum_{j=1}^{J} p^j d^j(x^m).$$

Now write $\overline{d}^j = \left(d^j(x^1), \ldots, d^j\left(x^{|\text{supp}(X)|}\right)\right)'$, which is again in a dimension $|\text{supp}(X)|(K-1)$ affine space, the same dimension as $\overline{p}$. We compactly can write $\overline{p} \in \text{Conv}(\{\overline{d}^1, \ldots, \overline{d}^J\})$. From Caratheodory's Theorem, we have that $\overline{p}$ can then be written as a convex combination of at most $|\text{supp}(X)|(K-1)+1$ elements of $\left\{\overline{d}^j\right\}_{j=1}^{J}$. It is thus without loss of generality to set $J \leqslant |\text{supp}(X)|(K-1)+1$. □

Proof of Proposition 5. First we prove (ii) $\implies$ (i). Let $\mathcal{H} = \{H_1, \ldots, H_Q\}$ and set $\mathcal{A} = \{A_1, \ldots, A_Q\}$. For action A_q, define the second-stage shock distribution to be H_q, and choose

$$k^* \in \underset{k}{\text{argmax}}\{u_k(X_k) + \zeta_k\}.$$

Then the value of A_q can be written as

$$V_{A_q}(u(X)) = \mathbb{E}_{H_q}\left[\max_k\{u_k(x_k) + \zeta_k\}\right].$$

DRM states that the action $H_q = A_q$ is chosen to maximize this. Thus we have a 2SM with $|\mathcal{A}| = Q$, yielding the same $P(X)$. To complete the translation, write $\varepsilon_k(A_q, \eta) = \zeta_k$, set $C(A, \omega) = 0$, and set ξ degenerate.

Next we prove (i) $\implies$ (iii). Index actions by $q \in \{1, \ldots, Q\}$. For each fixed q, apply Proposition 4 to the one-step final-stage problem associated with action A_q. This constructively yields a finite distribution of shocks

$$\{(p_q^j, c_q^j)\}_{j=1}^{J_q}$$

and the corresponding value function

$$\overline{V}^S_{A_q}(v) = \sum_{j=1}^{J_q} p_q^j \max_k \{v_k + c_{qk}^j\} \in \mathcal{V}^{1SM}(J_q)$$

such that

$$P^S(x, A_q) \in \partial \overline{V}^S_{A_q}(u(x)), \qquad x \in \operatorname{supp}(X),$$

with

$$J_q \leqslant |\operatorname{supp}(X)|(K-1)+1.$$

Set $J = \max_q J_q$, and add zero-weight terms as needed so each $\overline{V}^S_{A_q} \in \mathcal{V}^{1SM}(J)$. Define

$$\overline{V}^{Q-2SM}(v) := \max_{q \in \{1,\ldots,Q\}} \overline{V}^S_{A_q}(v).$$

Let

$$M(v) := \operatorname*{argmax}_{q \in \{1,\ldots,Q\}} \left\{ \overline{V}^S_{A_q}(v) + C(A_q) \right\},$$

which uses the assumption that (ξ, ω) are degenerate. The first stage $P^F(x)$ is a mixture over maximizing actions allowing ties, so

$$P^F(x) \in \operatorname{Conv}\{e_q^{|\mathcal{A}|} : q \in M(u(x))\}.$$

Note here that this relationship is only over $x \in \operatorname{supp}(X)$. We observe that M is constructed from the composition of an outer max function over Q, and inner $\overline{V}^S_{A_q}$ functions, so we can use the law of total probability and chain rule to conclude that

$$P(x) \in \partial \overline{V}^{Q-2SM}(u(x)).$$

Observe that $\overline{V}^S_{A_q}(v) + C(A_q) \in \mathcal{V}^{1SM}(J)$ because adding a constant keeps us in the class. Hence (iii) holds. Also, since each $J_q \leqslant |\operatorname{supp}(X)|(K-1)+1$, we have

$$J = \max_q J_q \leqslant |\operatorname{supp}(X)|(K-1)+1.$$

Finally we prove (iii) $\implies$ (ii). Let

$$\overline{V}^{Q-2SM}(v) = \max_{q\in\{1,\dots,Q\}} \overline{V}^{S}_{A_q}(v), \qquad \overline{V}^{S}_{A_q} \in \mathcal{V}^{1SM}(J).$$

By the definition of $\mathcal{V}^{1SM}(J)$, each term can be written as

$$\overline{V}^{S}_{A_q}(v) = \sum_{j=1}^{J} p_q^j \max_k\{v_k + c_{qk}^j\} = \mathbb{E}_{H_q}\Big[\max_k\{v_k + \zeta_k\}\Big]$$

for a discrete distribution H_q that places mass p_q^j on shock vector c_q^j. Set

$$\mathcal{H} := \{H_1, \dots, H_Q\},$$

allowing duplicate labels if needed to keep the cardinality Q. Define

$$\overline{V}^{DR}(v) := \max_{H\in\mathcal{H}} \mathbb{E}_{\zeta\sim H}\Big[\max_k\{v_k + \zeta_k\}\Big] = \max_q \overline{V}^{S}_{A_q}(v) = \overline{V}^{Q-2SM}(v).$$

From (iii), for each $x \in \text{supp}(X)$,

$$P(x) \in \partial\overline{V}^{Q-2SM}(u(x)) = \partial\overline{V}^{DR}(u(x)),$$

which is exactly the DRM condition with $|\mathcal{H}| = Q$. Therefore (ii) holds. □

Proof of Theorem 6.

Step 1. Equivalence of (iii)-(vi). The implication (iii) $\implies$ (vi) is immediate from the definition of the subgradient, (vi) $\implies$ (v) from summing up the inequalities (and (v) $\implies$ (vi) from Theorem 1 in Allen and Rehbeck (2020b)), and clearly (iv) implies (iii). To complete the first part we prove that (vi) implies (iv). Thus define:

$$\begin{aligned} g^w(v) &:= \max_{x\in\text{supp}(X|W=w)} \{\tau_x^w + P(x,w)'(v - u(x))\} \\ &= \max_{x\in\text{supp}(X|W=w)} \{P(x,w)'v + (\tau_x^w - P(x,w)'u(x))\}, \end{aligned}$$

which we see satisfies $g^w \in \mathcal{V}^{2SM}(|\,\text{supp}(X \mid W = w)|)$. We have the attainment $g^w(u(\tilde{x})) \geqslant \tau_{\tilde{x}}^w + P(\tilde{x},w)'(u(\tilde{x}) - u(\tilde{x})) = \tau_{\tilde{x}}^w$, and from (vi) we have $g^w(u(\tilde{x})) \leqslant \tau_{\tilde{x}}^w$, so $g^w(u(\tilde{x})) = \tau_{\tilde{x}}^w$ for any $x \in \text{supp}(X \mid W = w)$. Hence, the subgradient inequalities

(vi) establish $P(x,w) \in \partial g^w(u(x))$ by substituting $g^w(u(x))$ for each τ_x^w.

Step 2. (vi) $\implies$ (i). Use the previous g^w construction. Index the actions then as $A^{x,w}$, which involves a second stage distribution of shocks that satisfies $\varepsilon_k(A^{x,w},\eta) = -\infty$ for all alternatives except one, which then attains utility $c^{x,w} = \tau_x^w - P(x,w)'u(x)$, which is finite. Match so that the probability of $\varepsilon_k(A^{x,w},\eta) = c^{x,w}$ equals $P_k(x,w)$. The value function of each $A^{x,w}$ takes the form $V^S_{A^{x,w}}(v) = P(x,w)'v + c^{x,w}$. Now build $C(A,W)$ to equal 0 for actions $A^{x,\tilde{w}}$ with $\tilde{w} = w$, and set $C(A,W) = -\infty$ otherwise. We obtain for any $\tilde{x} \in \operatorname{supp}(X \mid W = w)$ that the corresponding action satisfies $V^S_{A^{\tilde{x},w}}(u(x)) + C(A^{\tilde{x},w},w) = P(\tilde{x},w)'u(x) + c^{\tilde{x},w} \leqslant g^w(u(x))$. Equality is attained at $A^{x,w}$ and so the labeled action $A^{x,w}$ maximizes the first stage among all $A^{\tilde{x},w}$, and hence also among all $A^{\tilde{x},\tilde{w}}$ because of the construction of C. Constructively we can set $P^F_{A^{x,w}}(x,w) = 1$ for the associated $A^{x,w}$, and $P^S_k(x,A^{x,w}) = P_k(x,w)$. The total actions are $\mathcal{A} = \cup_{(x,w)\in\operatorname{supp}(X,W)}\{A^{x,w}\}$ so that $|\mathcal{A}| = |\operatorname{supp}(X,W)|$. There is no (ξ,ω) shock in this rationalization. Adding redundant actions as needed to reach Q, we thus match (i).

Step 3. (i) $\implies$ (ii) $\implies$ (iii). Clearly (i) $\implies$ (ii). Finally, (ii) $\implies$ (iii) from Theorem 2. □

Appendix B Tools

We employ convex-analytic tools in the tradition of Rockafellar (2015), which allow us to handle the possibility of multiple maximizers and instances in which there are random constraints so that we must work in the extended reals.

A set $S \subseteq \mathbb{R}^L$ is *convex* if $v, \overline{V} \in S$ implies $\alpha v + (1-\alpha)\tilde{v} \in S$ for any $\alpha \in [0,1]$. A function $f : \mathbb{R}^L \to \mathbb{R} \cup \{\infty\}$ is *convex* if it has a convex *epigraph* $\{(v,r) \in \mathbb{R}^{L+1} \mid f(v) \leqslant r\}$, and f is *proper* if it is not identically ∞. The *effective domain* $\operatorname{dom} f = \{v \mid f(v) < \infty\}$ can be interpreted as feasible values, and allows integrated analysis of constrained and unconstrained optimization. When f is real-valued, convexity reduces to $f(\alpha v + (1-\alpha)\tilde{v}) \geqslant \alpha f(v) + (1-\alpha)f(\tilde{v})$ for all $v, \tilde{v} \in \mathbb{R}^L$ and $\alpha \in [0,1]$. For a convex function $f : \mathbb{R}^L \to \mathbb{R} \cup \{\infty\}$, the *subdifferential* is the multi-valued mapping ∂f defined on $\operatorname{dom} f$ via

$$\partial f(v) = \left\{ t \in \mathbb{R}^L \mid f(\tilde{v} \geqslant f(v) + t'(\tilde{v} - v), \forall \tilde{v} \right\},$$

where the inner product of vectors a and b is written as $a'b$. Each element of $\partial f(v)$ is a *subgradient* of f at v. There is a unique subgradient at v if and only if f is differentiable at v by Theorem 25.1 in Rockafellar (2015), in which case the gradient satisfies $\{\nabla f(v)\} = \partial f(v)$. A function f is *concave* if $-f$ is convex, and then $\partial(-f)$, i.e. the subdifferential of $-f$, is called the *superdifferential* of f, with elements termed *supergradients*. The convex conjugate of a function $f : \mathbb{R}^L \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(v) = \sup_{y \in \mathbb{R}^L} y'v - f(y).$$

Thus, f^* is the value function or indirect utility function of a specific maximization problem. Note that f^* is always convex, regardless of whether f is. Finally, $f : \mathbb{R}^L \to \mathbb{R} \cup \{\infty\}$ is lower-semicontinuous if $\{y \in \mathbb{R}^K \mid f(y) > \alpha\}$ is open for each scalar α. Everywhere-finite convex functions are continuous and hence automatically lower-semicontinuous.

The following result presents core duality relating objective functions, value functions, subgradients, and optimal choices.

Lemma 3 (Rockafellar (2015), Theorem 23.5)**.** *Let $f : \mathbb{R}^L \to \mathbb{R} \cup \{\infty\}$ be proper, convex, and lower semi-continuous. For any pair $(v, v^*) \in \mathbb{R}^L \times \mathbb{R}^L$:*

(i) $\partial f^*(v) = \operatorname{argmax}_{y \in \mathbb{R}^L} y'v - f(y)$.

(ii) $\partial f(v^*) = \operatorname{argmax}_{y \in \mathbb{R}^L} v^{*\prime}y - f^*(y)$.

(iii) $v^* \in \partial f^*(v) \iff v \in \partial f(v^*)$

Part (i) is an envelope theorem, tying variation in a value function to an optimal choice. This lemma has a number of implications. For example, if the optimal choice in (i) at v is unique and denoted v^*, then we have

$$v^* = \nabla f^*(v).$$

We also employ a composition and chain rule to link the value function between stages. The classic chain rule requires the inner and outer functions to both be differentiable. A number of generalizations are known. We use the following version.

Lemma 4 (Composition and Chain Rule)**.** *Let $f : \mathbb{R}^L \to \mathbb{R}$ be convex and weakly increasing in each argument, and let $g_l : \mathbb{R}^M \to \mathbb{R}$ be convex for $l \in \{1, \ldots, L\}$. Define $h : \mathbb{R}^M \to \mathbb{R}$ as*

$$h(v) = f(g_1(v), \ldots, g_L(v)).$$

(i) h is convex.

(ii) (Corollary 14 in Burke et al. (2021)) Let $v \in \mathbb{R}^M$ and set $\overline{g} = (g_1(v), \ldots, g_L(v))$. We have:

$$\partial h(v) = \left\{ \sum_{l=1}^{L} \gamma_l \delta_l \mid \gamma \in \partial f(\overline{g}), \delta_l \in \partial g_l(v) \text{ for each } l \in \{1, \ldots, L\} \right\}$$

Proof. To prove (i), let $v, \tilde{v} \in \mathbb{R}^M$ and let $\alpha \in [0, 1]$. We obtain

$$\begin{aligned} h(\alpha v + (1-\alpha)\tilde{v}) &= f(g_1(\alpha v + (1-\alpha)\tilde{v}), \ldots, g_L(\alpha v + (1-\alpha)\tilde{v})) \\ &\leqslant f(\alpha g_1(v) + (1-\alpha) g_1(\tilde{v}), \ldots, \alpha g_L(v) + (1-\alpha) g_L(\tilde{v})) \\ &\leqslant \alpha f(g_1(v), \ldots, g_L(v)) + (1-\alpha) f(g_1(\tilde{v}), \ldots, g_L(\tilde{v})) \\ &= \alpha h(v) + (1-\alpha) h(\tilde{v}). \end{aligned}$$

The first inequality uses the assumption each g_l is finite-valued and convex, and so

$$g_l(\alpha v + (1-\alpha)\tilde{v}) \leqslant \alpha g_l(v) + (1-\alpha) g_l(\tilde{v}),$$

and f is weakly increasing in each argument. The second inequality uses the fact that f is convex. Since h is finite-valued, we conclude it is convex.

For part (ii), see Corollary 14 in Burke et al. (2021). We can apply that corollary because f and each g_l are everywhere finite, so that the conditions in Theorem 12(b) in Burke et al. (2021) are automatically satisfied. □

Monotonicity of the outer function f is important for part (i). For example, $f(v) = -v$ and $g_1(v) = v^2$ are both convex, but $f(g_1(v)) = -v^2$ is not convex.

B.1 Convex Analysis Lemmas

Lemma 5 (Corollary 8.25 and Example 8.26 in Rockafellar and Wets (1998))**.** *Let* $f(v) = \max_{\ell \in \{1,\dots,L\}}\{v_\ell\}$*. Then* f *is the support function of the probability simplex* Δ^{L-1}*, so that*

$$\partial f(v) = \operatorname*{argmax}_{w \in \Delta^{L-1}} \sum_{\ell=1}^{L} w_\ell v_\ell = \operatorname{Conv}\left\{ e_\ell : \ell \in \operatorname*{argmax}_{\ell \in \{1,\dots,L\}}\{v_\ell\} \right\},$$

where $e_\ell \in \mathbb{R}^L$ *is the unit vector with a* 1 *in position* ℓ *and zero elsewhere.*

Lemma 6 (Theorem 23.8 in Rockafellar (2015))**.** *Let* $f_1, \dots, f_m$ *be proper convex functions on* $\mathbb{R}^L$ *and* $f = f_1 + \cdots + f_m$*. If each* f_m *is everywhere finite, then*

$$\partial f = \partial f_1 + \cdots + \partial f_m.$$

The next three results are basic. The literature on convex risk measures uses similar properties, focusing on infinite dimensional contexts, e.g. Filipović and Kupper (2007).

Lemma 7. *Let* $f : \mathbb{R}^L \to \mathbb{R}$ *be convex. Suppose* f *is translation invariant in the sense that* $f(v + c1_L) = f(v) + c$ *for every* $v \in \mathbb{R}^L$ *and scalar* c*. It follows that* $\operatorname{dom}(f^*) \subseteq \{y \in \mathbb{R}^L \mid \sum_{\ell=1}^{L} y_\ell = 1\}$.

Proof of Lemma 7. Let $y, v \in \mathbb{R}^L$ and let $c \in \mathbb{R}$. We see:

$$y'(v + c1_L) - f(v + c1_L) = y'v - f(v) + c(y'1_L - 1).$$

Suppose that $\sum_{\ell=1}^{L} y_\ell = y'1_L > 1$. Then sending $c \to \infty$ increases this expression without bound and we see that $f^*(y) = \infty$. When instead $y'1_L < 1$ we can send $c \to -\infty$ to see $f^*(y) = \infty$. □

The following result is a type of known duality between unit balls in $\|\cdot\|_\infty$ and $\|\cdot\|_1$ (here $\|y\|_1 = \sum_{\ell=1}^{L} |y_\ell|$).

Lemma 8. *Let* $f : \mathbb{R}^L \to \mathbb{R}$ *be* 1*-Lipschitz in* $\|\cdot\|_\infty$*. It follows that* $\operatorname{dom}(f^*) \subseteq \{y \in \mathbb{R}^L \mid \sum_{\ell=1}^{L} |y_\ell| \leqslant 1\}$.

Proof. Suppose now that $\sum_{\ell=1}^{L} |y_\ell| > 1$. We wish to show that $f^*(y) = \infty$. Let $s \in \{-1, 1\}^L$ satisfy $s_\ell = 1$ when $y_\ell \geqslant 0$, and be -1 otherwise. Set $v = cs$ so that $\|v\|_\infty = c$. Since f is 1-Lipschitz we have $f(v) \leqslant f(0) + \|v\|_\infty = f(0) + c$. We obtain

$$y'v - f(v) \geqslant c\|y\|_1 - (f(0) + c) = c\,(\|y\|_1 - 1) - f(0).$$

Send $c \to \infty$ to conclude $f^*(y) = \infty$. □

Lemma 9. *Let $f : \mathbb{R}^L \to \mathbb{R}$ be translation invariant and 1-Lipschitz in $\|\cdot\|_\infty$. It follows that $\mathrm{dom}(f^*) \subseteq \Delta^{L-1}$.*

Proof. This follows from the previous two lemmas and the fact that $\sum_{\ell=1}^{L} y_\ell = 1$ and $\sum_{\ell=1}^{L} |y_\ell| \leqslant 1$ force $y \in \Delta^{K-1}$. □

B.2 Motonicity and Binary Choice with Ties

To cover discrete and continuous cases as well as utility ties, we use the machinery of a multi-valued mapping $\rho : \mathbb{R} \rightrightarrows [0, 1]$. Hence $\rho(v)$ is a subset of $[0, 1]$, possibly empty. Because we need to allow for empty sets, this is not a correspondence. To relate to this paper, interpret $v = u_1(x_1) - u_2(x_2)$ as a utility difference and $\rho(v)$ as the set of all probabilities for good 1 that have the same utility difference; this would be a singleton if utility ties were ruled out. Empty $\rho(v)$ corresponds to utility differences outside the support. Thus with finite support of covariates, $\rho(v)$ is empty for typical v. Say that ρ is monotone if for any $(p, v), (\tilde{p}, \tilde{v}) \in \mathrm{gph}(\rho)$, we have

$$(\tilde{p} - p)(\tilde{v} - v) \geqslant 0.$$

Here for any multivalued mapping $\nu : \mathbb{R} \rightrightarrows [0, 1]$, the graph is defined as

$$\mathrm{gph}(\nu) := \{(p, v) \in [0, 1] \times \mathbb{R} : p \in \nu(v)\}.$$

Lemma 10. *Let $\rho : \mathbb{R} \rightrightarrows [0, 1]$ be a monotone multivalued mapping. Then there exists a monotone extension $\rho^{Ext} : \mathbb{R} \rightrightarrows [0, 1]$ such that $\mathrm{gph}(\rho) \subseteq \mathrm{gph}(\rho^{Ext})$ and $\rho^{Ext}(v) \neq \varnothing$ for all $v \in \mathbb{R}$, and the extension also satisfies the property that there exists*

a probability distribution H on $\overline{\mathbb{R}} = [-\infty, \infty]$ such that for any $(p, v) \in \mathrm{gph}(\rho^{Ext})$,

$$p = \Pr_H(\zeta < v) + \tau \Pr_H(\zeta = v)$$

for some $\tau \in [0, 1]$.

Proof. We construct the extension directly.

Step 1. Build a CDF from ρ. For each $v \in \mathbb{R}$, define

$$S(v) := \{p \in [0, 1] : \exists t > v \text{ with } p \in \rho(t)\}, \qquad F(v) := \inf S(v),$$

with the convention $\inf \varnothing := 1$.

Because $S(v_2) \subseteq S(v_1)$ when $v_1 < v_2$, F is nondecreasing. We next show that F is right-continuous. If $S(v) = \varnothing$, then $F(v) = 1$ and $F(s) = 1$ for every $s > v$. If $S(v) \neq \varnothing$, pick $\delta > 0$ and choose (p_δ, t_δ) with $t_\delta > v$, $p_\delta \in \rho(t_\delta)$, and $p_\delta < F(v) + \delta$. For any $s \in (v, t_\delta)$, we have $p_\delta \in S(s)$ so $F(s) \leqslant p_\delta < F(v) + \delta$, while monotonicity of F gives $F(s) \geqslant F(v)$. Hence $\lim_{s \downarrow v} F(s) = F(v)$.

Step 2. Define the extension. Let

$$\rho^{Ext}(v) := [F(v^-), F(v)],$$

where $F(v^-) = \lim_{t \uparrow v} F(t)$. This is nonempty since F is nondecreasing.

Now show $\rho \subseteq \rho^{Ext}$. Take $(p, v) \in \mathrm{gph}(\rho)$. For the upper bound, if $q \in \rho(t)$ with $t > v$, monotonicity of ρ gives $q \geqslant p$, so $p \leqslant \inf S(v) = F(v)$. For the lower bound, if $t < v$, then (p, v) itself implies $p \in S(t)$, so $F(t) \leqslant p$. Taking $\sup_{t<v}$ yields $F(v^-) \leqslant p$. Therefore $p \in [F(v^-), F(v)] = \rho^{Ext}(v)$.

The extension is monotone: if $v < \tilde{v}$, $p \in \rho^{Ext}(v)$, and $\tilde{p} \in \rho^{Ext}(\tilde{v})$, then

$$p \leqslant F(v) \leqslant F(\tilde{v}^-) \leqslant \tilde{p},$$

so $(\tilde{p} - p)(\tilde{v} - v) \geqslant 0$.

Step 3. Construct a shock distribution and tie representation. Since F is

nondecreasing, right-continuous, and bounded in $[0,1]$, define

$$F_- := \lim_{t\to-\infty} F(t), \qquad F_+ := \lim_{t\to+\infty} F(t).$$

Let $U \sim \text{Unif}[0,1]$, and define $\zeta \in \overline{\mathbb{R}}$ by

$$\zeta = \begin{cases} -\infty, & U \leqslant F_-, \\ \inf\{t \in \mathbb{R} : F(t) \geqslant U\}, & U > F_-, \end{cases}$$

with $\inf \varnothing := +\infty$. Let H be the distribution of ζ. We obtain for each v,

$$\Pr_H(\zeta \leqslant v) = F(v), \quad \Pr_H(\zeta < v) = F(v^-), \quad \Pr_H(\zeta = v) = F(v) - F(v^-).$$

Take any $(p,v) \in \text{gph}(\rho^{Ext})$. Then $p \in [F(v^-), F(v)]$, so choose

$$\tau = \begin{cases} \dfrac{p - F(v^-)}{F(v) - F(v^-)} & \text{if } F(v) > F(v^-), \\ 0 & \text{if } F(v) = F(v^-). \end{cases}$$

Then $\tau \in [0,1]$ and

$$p = F(v^-) + \tau\big(F(v) - F(v^-)\big) = \Pr_H(\zeta < v) + \tau \Pr_H(\zeta = v).$$

This is exactly a binary threshold model with ties: choose alternative 1 if $v > \zeta$, alternative 2 if $v < \zeta$, and at $v = \zeta$ choose alternative 1 with probability τ. □

Appendix C Identification

Theorem 1 establishes that results for point or partial identification of utility indices u or ex ante welfare $\overline{V}$ can be ported between the models. Allen and Rehbeck (2019a) have provided conditions for point identification in PUM and so those conditions also apply to the multistage models. I outline the ideas here, because they have not been applied to the general two stage model. Allen and Rehbeck (2019a) study PUM without the first period shifters W, but their analysis holds when we fix $W = w$. I first review their arguments and then generalize to the case in which W is not

degenerate. This is needed to identify $\overline{V}$ when we vary w.

C.1 Identification Fixing $W = w$

The key part is the envelope condition from Theorem 2:

$$P(X, w) = \nabla_v \overline{V}(u(X) \mid w).$$

Suppose for illustration each X_k is scalar, varies continuously, and is alternative-specific and thus can move separately for each good. Taking partial derivatives, we obtain for alternatives k, j that

$$\frac{\partial P_k(X, w)}{\partial X_j} = \nabla_{j,k} \overline{V}(u(X); w) u_j'(X_j).$$

This uses the exclusion restriction that X_j only shifts the utility index u_j. Since partial derivatives are symmetric, $\nabla_{j,k} = \nabla_{k,j}$, we can interchange the role of k and j and divide to obtain

$$\frac{\frac{\partial P_k(X,w)}{\partial X_j}}{\frac{\partial P_j(X,w)}{\partial X_k}} = \frac{u_j'(X_j)}{u_k'(X_k)}. \tag{4}$$

Allen and Rehbeck (2019a) show that with one scale assumption such as $|u_j'(0)| = 1$, and K locations such as $u_k(0) = 0$ for each k, we can identify all u_k by integrating this equation. Allen and Rehbeck (2019a) also generalize to the case in which X_k is not scalar using an injectivity argument, so that X_k can have certain components shared between alternatives, such as demographic variables. The key though is that each alternative must have at least one continuous regressor that is excluded from the other utility indices. All other covariates can be discrete and/or shared across alternatives, such as demographic variables.

Allen and Rehbeck (2019a) note that a testable implication of (4) is that the derivative ratio must be the same as X_ℓ varies for any ℓ that is distinct from k, j, because X_ℓ is not present on the right hand side, but is on the left hand side.[21] Note this testable implication requires at least 3 alternatives. An additional testable implication is that

[21] This is not subject to the no empirical content result of Theorem 2 in Allen and Rehbeck (2023) because we impose smoothness conditions and have a unique maximizer of the PUM representation, obtained when we assumed differentiability of $\overline{V}$.

the derivative ratio is the same as w varies.

Suppose now that we have identified each utility index u_k, either by this argument or a priori knowledge of the utility indices. We can directly "integrate" the relation

$$P(X, w) \in \partial_v \overline{V}(u(X); w),$$

by varying X. This recovers differences in $\overline{V}$. See Theorem 4 in Allen and Rehbeck (2019a) for formal details in PUM, which also applies to the multistage models and does not require $\overline{V}$ to be differentiable. Section 4.2 in Allen and Rehbeck (2019a) presents a direct formula for welfare changes that is amenable to plug-in estimation: a researcher need only estimate $P(X, w)$ and derivatives to estimate welfare differences.

C.2 Identification Varying W

For identification, the change relative to Allen and Rehbeck (2019a) is that now W is present and can vary. First note we can reproduce the analysis above to identify u_k^w at a specific $W = w$ value. If we repeat this argument at another $\tilde{w}$ value, we necessarily have that $u^w = \lambda u^{\tilde{w}} + c1$ for some $\lambda > 0$ and $c1 = (c, \ldots, c)'$ vector. The question arises whether we need to assume separate location/scale assumptions for each w conditioning value, or whether we can find a common $u_k = u_k^w$ with just a scale/location assumption at one w. In general we need to pick a location/scale for each w when we treat first stage costs C as unrestricted. To see this, use the PUM representation and consider a class $\lambda(w)(D(p, w) + c(w))$. Here, changing $\lambda(w)$ changes the scale of u^w via $u = u^w/\lambda(w)$ and changing $c(w)$ changes the location. We generally cannot rule out such rescalings. We see that it is a normalization to pick one w^* that sets the scale and location for the whole system, so that $u^w = u^{w^*}$.

References

Abaluck, J. and A. Adams-Prassl (2021): "What do consumers consider before they choose? Identification from asymmetric demand responses," *The Quarterly Journal of Economics*, 136, 1611–1663.

Afriat, S. N. (1967): "The construction of utility functions from expenditure data," *International economic review*, 8, 67–77.

Agarwal, N. and P. J. Somaini (2022): "Demand Analysis under Latent Choice Constraints," Tech. rep., National Bureau of Economic Research.

Aguiar, V. H. and N. Kashaev (2025): "Identification and estimation of discrete choice models with unobserved choice sets," *Journal of Business & Economic Statistics*, 43, 204–215.

Ajayi, K. and M. Sidibe (2020): "School choice under imperfect information," *Economic Research Initiatives at Duke (ERID) Working Paper*.

Allen, R. (2026): "Exogenous consideration and extended random utility," *Journal of Econometrics*, 253, 106166.

Allen, R. and J. Rehbeck (2016): "Complementarity in perturbed utility models," *University of California-San Diego*.

——— (2019a): "Identification with additively separable heterogeneity," *Econometrica*, 87, 1021–1054.

——— (2019b): "Supplement to "Identification with additively separable heterogeneity" (Online Appendix)," Wiley Online Library, supporting information, supplementary material for Econometrica, Vol. 87, No. 3.

——— (2020a): "Counterfactual and welfare analysis with an approximate model," *arXiv preprint arXiv:2009.03379*.

——— (2020b): "Satisficing, aggregation, and quasilinear utility," *Available at SSRN 3180302*.

——— (2023): "Revealed stochastic choice with attributes," *Economic Theory*, 75, 91–112.

Amos, B., L. Xu, and J. Z. Kolter (2017): "Input convex neural networks," in *International conference on machine learning*, PMLR, 146–155.

Anderson, S. P. and A. De Palma (1992): "Multiproduct firms: A nested logit approach," *The Journal of Industrial Economics*, 261–276.

Anderson, S. P., A. De Palma, and J.-F. Thisse (1988): "A representative consumer theory of the logit model," *International Economic Review*, 461–466.

Arcidiacono, P. (2005): "Affirmative action in higher education: How do admission and financial aid rules affect future earnings?" *Econometrica*, 73, 1477–1524.

Armstrong, M. (2017): "Ordered consumer search," *Journal of the European Economic Association*, 15, 989–1024.

Balke, A. and J. Pearl (1994): "Counterfactual probabilities: Computational methods, bounds and applications," in *Uncertainty in artificial intelligence*, Elsevier, 46–54.

——— (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American statistical Association*, 92, 1171–1176.

Ben-Akiva, M. E. (1973): "Structure of passenger travel demand models." Ph.D. thesis, Massachusetts Institute of Technology.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821.

Berry, S., J. Levinson, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, 242–262.

Bertsekas, D. P. (1973): "Stochastic optimization problems with nondifferentiable cost functionals," *Journal of Optimization Theory and Applications*, 12, 218–231.

Billingsley, P. (2017): *Probability and measure*, John Wiley & Sons.

Brown, D. J. and C. Calsamiglia (2007): "The nonparametric approach to applied welfare analysis," *Economic Theory*, 31, 183–188.

Browning, M. (1989): "A nonparametric test of the life-cycle rational expections hypothesis," *International Economic Review*, 979–992.

Burke, J. V., H. Tim, and Q. V. Nguyen (2021): "A study of convex convex-composite functions via infimal convolution with applications," *Mathematics of Operations Research*, 46, 1324–1348.

Che, Y.-K., J. Grenet, and Y. He (2025): "Allocating students to schools: theory, methods, and empirical insights," *Handbook of the Economics of Matching*, 2, 307–407.

Chib, S. and K. Shimizu (2025): "Scalable estimation of multinomial response models with random consideration sets," *Journal of the American Statistical Association*, 1–28.

Chiong, K. X., A. Galichon, and M. Shum (2016): "Duality in dynamic discrete-choice models," *Quantitative Economics*, 7, 83–115.

Chiong, K. X., Y.-W. Hsieh, and M. Shum (2021): "Bounds on counterfactuals in semiparametric discrete-choice models," *Handbook of Research Methods and Applications in Empirical Microeconomics*, 223–237.

Choi, M., A. Y. Dai, and K. Kim (2018): "Consumer search and price competition," *Econometrica*, 86, 1257–1281.

Crawford, G. S., R. Griffith, and A. Iaria (2021): "A survey of preference estimation with unobserved choice set heterogeneity," *Journal of Econometrics*, 222, 4–43.

Eichhorn, A. and W. Römisch (2005): "Polyhedral risk measures in stochastic programming," *SIAM Journal on Optimization*, 16, 69–95.

Falmagne, J.-C. (1978): "A representation theorem for finite random scale systems," *Journal of Mathematical Psychology*, 18, 52–72.

Fang, Z., A. Santos, A. M. Shaikh, and A. Torgovitsky (2023): "Inference for Large-Scale Linear Systems With Known Coefficients," *Econometrica*, 91, 299–327.

Feng, G., X. Li, and Z. Wang (2017): "On the relation between several discrete choice models," *Operations research*, 65, 1516–1525.

——— (2018): "On substitutability and complementarity in discrete choice models," *Operations Research Letters*, 46, 141–146.

Filipović, D. and M. Kupper (2007): "Monotone and cash-invariant convex functions and hulls," *Insurance: Mathematics and Economics*, 41, 1–16.

Fosgerau, M., E. Melo, A. De Palma, and M. Shum (2020): "Discrete choice and rational inattention: A general equivalence result," *International economic review*, 61, 1569–1589.

Fosgerau, M., J. Monardo, and A. De Palma (2024): "The inverse product differentiation logit model," *American Economic Journal: Microeconomics*, 16, 329–370.

Fosgerau, M. and A. Ranjan (2017): "A note on identification in discrete choice models with partial observability," *Theory and Decision*, 83, 283–292.

Fu, C. (2014): "Equilibrium tuition, applications, admissions, and enrollment in the college market," *Journal of Political Economy*, 122, 225–281.

Fudenberg, D., R. Iijima, and T. Strzalecki (2015): "Stochastic choice and revealed perturbed utility," *Econometrica*, 83, 2371–2409.

Galichon, A. and B. Salanié (2022): "Cupid's invisible hand: Social surplus and identification in matching models," *The Review of Economic Studies*, 89, 2600–2629.

Gentzkow, M. (2007): "Valuing new goods in a model with complementarity: Online newspapers," *American Economic Review*, 97, 713–744.

Goeree, M. S. (2008): "Limited information and advertising in the US personal computer industry," *Econometrica*, 76, 1017–1074.

Grzeskiewicz, M. (2025): "Uncovering utility functions from observed outcomes," *arXiv preprint arXiv:2503.13432*.

Gu, J., T. M. Russell, and T. Stringham (2025): "Counterfactual identification and latent space enumeration in discrete outcome models," *Review of Economic Studies*, rdaf058.

Hanemann, W. M. (1984): "Discrete/continuous models of consumer demand," *Econometrica: Journal of the Econometric Society*, 541–561.

Hauser, J. R. and B. Wernerfelt (1990): "An evaluation cost model of consideration sets," *Journal of consumer research*, 16, 393–408.

He, Y., S. Sinha, and X. Sun (2024): "Identification and Estimation in Many-to-One Two-Sided Matching Without Transfers," *Econometrica*, 92, 749–774.

Hofbauer, J. and W. H. Sandholm (2002): "On the global convergence of stochastic fictitious play," *Econometrica*, 70, 2265–2294.

Honka, E. (2014): "Quantifying search and switching costs in the US auto insurance industry," *The RAND Journal of Economics*, 45, 847–884.

Honka, E., A. Hortaçsu, and M. Wildenbeest (2019): "Empirical search and consideration sets," in *Handbook of the Economics of Marketing*, Elsevier, vol. 1, 193–257.

Honoré, B. E. and E. Tamer (2006): "Bounds on parameters in panel dynamic discrete choice models," *Econometrica*, 74, 611–629.

Hotz, V. J. and R. A. Miller (1993): "Conditional choice probabilities and the estimation of dynamic models," *The Review of Economic Studies*, 60, 497–529.

Howell, J. S. (2010): "Assessing the impact of eliminating affirmative action in higher education," *Journal of Labor Economics*, 28, 113–166.

Hu, Y. and M. Shum (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

Kitamura, Y. and J. Stoye (2018): "Nonparametric analysis of random utility models," *Econometrica*, 86, 1883–1909.

Kleinberg, R., B. Waggoner, and E. G. Weyl (2016): "Descending price optimally coordinates search," *arXiv preprint arXiv:1603.07682*.

Koning, R. H. and G. Ridder (2003): "Discrete choice and stochastic utility maximization," *The Econometrics Journal*, 6, 1–27.

Kono, H., K. Saito, and A. Sandroni (2025): "Random Utility with Unobservable Alternatives," .

Koppelman, F. S. and C.-H. Wen (2000): "The paired combinatorial logit model: properties, estimation and application," *Transportation Research Part B: Methodological*, 34, 75–89.

Kovach, M. and G. Tserenjigmid (2022): "Behavioral foundations of nested stochastic choice and nested logit," *Journal of Political Economy*, 130, 2411–2461.

Machina, M. J. (1982): "" Expected utility" analysis without the independence axiom," *Econometrica: Journal of the Econometric Society*, 277–323.

——— (1985): "Stochastic choice functions generated from deterministic preferences over lotteries," *The economic journal*, 95, 575–594.

Magnac, T. and D. Thesmar (2002): "Identifying dynamic discrete decision processes," *Econometrica*, 70, 801–816.

Manski, C. F. (1977): "The structure of random utility models," *Theory and decision*, 8, 229.

——— (1989): "Anatomy of the selection problem," *Journal of Human resources*, 343–360.

——— (2007): "Partial identification of counterfactual choice probabilities," *International Economic Review*, 48, 1393–1410.

Matějka, F. and A. McKay (2015): "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *American Economic Review*, 105, 272–298.

Mattsson, L.-G. and J. W. Weibull (2002): "Probabilistic choice and procedurally bounded rationality," *Games and Economic Behavior*, 41, 61–78.

Mbakop, E. (2023): "Identification in Some Discrete Choice Models: A Computational Approach," *arXiv preprint arXiv:2305.15691*.

McFadden, D. (1978): "Modeling the Choice of Residential Location," *Transportation Research Record*, 72–77.

——— (1981): "Econometric models of probabilistic choice," *Structural analysis of discrete data with econometric applications*, 198272.

McFadden, D. and M. K. Richter (1990): "Stochastic rationality and revealed stochastic preference," *Preferences, Uncertainty, and Optimality, Essays in Honor of Leo Hurwicz, Westview Press: Boulder, CO*, 161–186.

McFadden, D. L. and M. Fosgerau (2012): "A theory of the perturbed consumer with general budgets," Tech. rep., National Bureau of Economic Research.

Monardo, J. (2025): "Functional Form and Shape Restrictions in Discrete Choice Models," *Available at SSRN 4681913*.

Müller, D., E. Melo, and R. Schlotter (2023): "A Distributionally Robust Random Utility Model," *arXiv preprint arXiv:2303.05888*.

Natarajan, K., M. Song, and C.-P. Teo (2009): "Persistency model and its applications in choice modeling," *Management Science*, 55, 453–469.

Nguyen, H. (2025): "Neural Networks for Efficient Estimation of High-Dimensional Dynamic Discrete Choice Models," .

Norets, A. and S. Takahashi (2013): "On the surjectivity of the mapping between utilities and choice probabilities," *Quantitative Economics*, 4, 149–155.

Pakes, A. and J. Porter (2024): "Moment inequalities for multinomial choice with fixed effects," *Quantitative Economics*, 15, 1–25.

Rehbeck, J. N. (2017): "Essays on Revealed Preference and Decisions," Ph.D. thesis, University of California, San Diego.

Roberts, J. H. and J. M. Lattin (1991): "Development and testing of a model of consideration set composition," *Journal of Marketing Research*, 28, 429–440.

Rockafellar, R. T. (2015): *Convex analysis*, Princeton university press.

Rockafellar, R. T. and R. J. Wets (1998): *Variational analysis*, Springer.

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

——— (1988): "Maximum likelihood estimation of discrete control processes," *SIAM journal on control and optimization*, 26, 1006–1024.

——— (1994): "Structural estimation of Markov decision processes," *Handbook of econometrics*, 4, 3081–3143.

Shi, X., M. Shum, and W. Song (2018): "Estimating semi-parametric panel multinomial choice models using cyclic monotonicity," *Econometrica*, 86, 737–761.

Sørensen, J. R.-V. and M. Fosgerau (2022): "How McFadden met Rockafellar and learned to do more with less," *Journal of Mathematical Economics*, 100, 102629.

Tebaldi, P., A. Torgovitsky, and H. Yang (2023): "Nonparametric estimates of demand in the california health insurance exchange," *Econometrica*, 91, 107–146.

Thompson, T. S. (1989): "Identification of semiparametric discrete choice models," .

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

Varian, H. R. (1982): "The nonparametric approach to demand analysis," *Econometrica: Journal of the Econometric Society*, 945–973.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341.

Walters, C. R. (2018): "The demand for effective charter schools," *Journal of Political Economy*, 126, 2179–2223.

Weitzman, M. L. (1979): "Optimal Search for the Best Alternative," *Econometrica: Journal of the Econometric Society*, 641–654.

Yang, E. and I. Kopylov (2023): "Random quasi-linear utility," *Journal of Economic Theory*, 209, 105650.