How McFadden met Rockafellar and learned to do more with less

Sørensen, Jesper R.-V.; Fosgerau, Mogens

Published in:
Journal of Mathematical Economics

DOI:
10.1016/j.jmateco.2021.102629

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
How McFadden met Rockafellar and learned to do more with less

Jesper R.-V. Sørensen, Mogens Fosgerau *

Department of Economics, University of Copenhagen, Denmark

A R T I C L E   I N F O

Article history:
Received 10 June 2021
Received in revised form 4 November 2021
Accepted 22 December 2021
Available online 31 December 2021
Manuscript handled by Editor John K.H. Quah

Keywords:
Additive random utility model
Discrete choice
Convex duality
Demand inversion
Partial identification

A B S T R A C T

We exploit the power of convex analysis to synthesize and extend a range of important results concerning the additive random utility model of discrete choice. With no restrictions on the joint distribution of random utility components or the functional form of systematic utility components, we formulate general versions of the Williams–Daly–Zachary theorem for demand and the Hotz–Miller demand inversion theorem. Based on these theorems, we provide necessary and sufficient conditions for demand and its inverse to reduce to functions. These conditions jointly imply that demand is a continuous function with a continuous inverse.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

In this paper, we systematically apply the power of convex analysis (Rockafellar, 1970) to a range of fundamental results concerning the additive random utility model (ARUM) of discrete choice (McFadden, 1974). We provide a unified treatment with minimal assumptions, and obtain several complete results with not only sufficient but also necessary conditions. In addition, we provide relatively simple proofs that rely on well established mathematical results from standard convex analysis, thus avoiding the re-invention and re-verification of mathematical arguments.

The ARUM describes utility maximizing choice among a set of mutually exclusive alternatives, each of which is characterized by a utility that is the sum of a systematic and a random component. We allow maximal generality: The systematic component is an arbitrary nonparametric function of conditioning variables, and there are no restrictions on the random component (other than measurability). Utility ties may occur with positive probability, and any tie-breaking rule is allowed. Consequently, the ARUM demand is a conditional choice probability (CCP) correspondence that is not in general single-valued.

Our first result is a general version of the Williams–Daly–Zachary (WDZ) theorem (McFadden, 1981), which is a discrete-choice analogue of Roy’s identity. We consider a normalized surplus function, which does not require integrability of the random utility component, and show that it is everywhere subdifferentiable with a subdifferential that is equal to the CCP correspondence.¹ The general WDZ theorem operationalizes the computation of the ARUM CCP correspondence based on a surplus function when the distribution of the random components is known. Moreover, as the subdifferential of a convex function, the CCP correspondence is cyclic monotone. The resulting family of inequalities, closely related to the Afriat (1967) conditions, can be exploited for semiparametric identification and estimation as in Shi, Shum, and Song (2018), but without the smoothness and regularity imposed on the random utility components by these authors.

Our second result extends the Hotz and Miller (1993) (HM) result on the invertibility of the CCP function to our general setting. We find that the inverse CCP correspondence is equal to the subdifferential of the conjugate surplus and that its domain contains the interior of the probability simplex. Thus, the HM and WDZ results are each other’s dual. Our general HM theorem may be viewed as a constructive identification result, which characterizes the identified set as the set of solutions to a convex minimization problem involving the surplus function. These results are closely related to Chiong, Galichon, and Shum (2016), who characterize the identified set based on the solution to an optimal transport problem.

We proceed to find necessary and sufficient conditions for the CCP correspondence and its inverse to be single-valued. We first find that the CCP correspondence is single-valued if and only if the surplus is differentiable, and if and only if the conjugate surplus is almost strictly convex. In this case, we obtain the differential WDZ result that the CCP correspondence reduces to the gradient of the surplus. These properties are shown to be

¹ See Appendix A for definitions and key results from convex analysis.
equivalent to a primitive condition on the distribution of the random utility components which requires that utility ties almost surely do not occur. Since our characterization of such distributions is exact, our statement of the differential WDZ theorem is maximally general.

We then establish the dual result, a differential HM theorem. We find that the inverse CCP correspondence is single-valued if and only if the surplus is strictly convex, and if and only if the conjugate surplus is almost differentiable. In this case, the inverse CCP correspondence reduces to the gradient of the conjugate surplus. These properties are again shown to be equivalent to a primitive condition on the distribution of the random utility components which amounts to it having sufficiently rich support. This result is therefore also maximally general.

When the conditions for both differential theorems hold, we find that the CCP correspondence reduces to a one-to-one function from $\mathbb{R}$ onto the interior of the probability simplex that is continuous in both directions. Continuity of the inverse function is useful for establishing asymptotic properties of estimators based on HM inversion. Surjectivity of this CCP function guarantees that its inverse is defined everywhere, and was first established by Norets and Takahashi (2013) under stronger conditions.

1.1. Literature review

In the classical ARUM formulation, random utility components are taken to be absolutely continuous with full support (see, e.g., Anderson, De Palma, and Thissie, 1992). According to our differential theorems, these classical assumptions are stronger than required for the WDZ and HM theorems.

The WDZ theorem, concluding that choice probabilities are given by the gradient of a surplus function, is attributed to Williams (1977) and Daly and Zachary (1978), see McFadden (1978). McFadden (1981, Section 5.8) and Rust (1994, Theorem 3.1).

Using the classical ARUM formulation within a dynamic discrete-choice (DDC) setup, Hotz and Miller (1993, Proposition 1) show that the CCP function is invertible. When the distribution of the random utility components is known, the HM inversion result implies that the vector of systematic utility components is identified (up to location) from the choice probability vector. This invertibility is often employed as a first step in two- (or multi-) step procedures for estimation of structural model parameters, notably Berry (1994) and Berry, Levinsohn, and Pakes (1995).

Convex analysis, and convex duality in particular, is a very natural approach to the problem of identification in the ARUM of discrete choice. The idea of convex conjugation – also known as Fenchel transformation – of the surplus was introduced to the discrete-choice literature by Hofbauer and Sandholm (2002) under classical assumptions. Galichon and Salanie (2020) used convex conjugation to identify a class of two-sided matching models. Building on a (now obsolete) 2010 version of Galichon and Salanie (2020), Chiong et al. (2016) used duality to capture the empirical content of the (single-agent) DDC model. Chiong et al. (2016) also give a computationally efficient method for estimation of their DDC model by means of solving an optimal transport problem. Their optimal transport approach to the first stage of a two-stage estimation method facilitates HM inversion in cases where the conjugate surplus is not available in closed form.

Most existing demand estimation techniques fail in the presence of zero demands, but zero demands are common, especially in “big data” (e.g., Nurski and Verboven, 2016; Quan and Williams, 2018). We establish that the inverse CCP correspondence is always well defined. Our general partial identification result therefore readily allows for zero probabilities. Moreover, given the constructive nature of this identification, our result may therefore be used as a basis for estimation even if one encounters zero probabilities empirically.

Our identification of the deterministic utility components parallels the analyses in Hofbauer and Sandholm (2002), Galichon and Salanie (2020) and Chiong et al. (2016) in treating the distribution of the random utility components as known. Other authors rely on different model assumptions in order to obtain identification. Notably, in pioneering work, Rosa Matzkin establishes nonparametric identification of both the systematic utility components and the distribution of random utility components under various independence and/or shape restrictions motivated by economic theory (see, e.g., Matzkin, 1992, 1993, 1994). Using a special regressor, Fosgerau and Kristensen (2020) show identification in a class of index models that comprise the ARUM with very little structure imposed and without continuity of the deterministic utility component. Allen and Rehbeck (2019) show identification of the systematic utility components in a wider class of models with additively separable unobservable heterogeneity assumed to be independent of conditioning variables. In contrast, our identification results rely on neither independence assumptions nor the presence of a special regressor. The present paper should thus be viewed as complementary to the work of the above-mentioned authors.

We formulate the general ARUM in Section 2 and the general WDZ theorem in Section 3. Section 4 dualizes the general WDZ theorem to produce a general HM inversion result, which, in turn, produces partial identification of systematic utilities. Section 5 presents assumptions that just suffice for the CCP correspondence and its inverse to be single-valued, thus yielding point identification. Section 6 concludes. Appendix A gives a summary of elements of convex analysis used throughout the paper. Appendix B contains all proofs, and Appendix C includes a supplementary example.

Notation. Vectors $x$ and $y$ are understood as columns. $(x, y)$ denotes the usual scalar/inner product. $\| \cdot \|$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean and supremum norms, respectively. The interior and closure of a set $S$ are denoted $\text{int} S$ and $\text{cl} S$, respectively. The subdifferential of a function with arguments such as $f(\cdot|\cdot)$ is understood as the subdifferential with respect to the first argument. The same convention applies to gradients. We write $\text{conv} S$ for the convex hull of a set $S$, i.e. all convex combinations of elements of $S$.

2. Discrete choice model setup

In this section we set up a general version of the additive random utility model (ARUM) of discrete choice. In this model a

---

2 See also Lindberg (2012) for a proof of the WDZ theorem under the condition that choice probabilities are continuous functions.

3 Fosgerau, Melo, Shum, and Sørensen (2021a) suggest an alternative computational approach based on convex optimization.

4 Aggregating data over markets or products may smooth over the heterogeneity of interest (Quan and Williams, 2018). Omitting products without sales implicitly assumes that there is no demand for these products and may create a selection bias (Berry, Levinsohn, and Pakes, 2004).

5 An exception to standard techniques can be found in Gandhi, Lu, and Shi (2019), who propose a solution to the zero-demands problem based on constructing bounds for the conditional expectation of inverse demand. However, in their paper, zero market shares may only occur in the sample—not the population. Hence, these authors rationalize zeros in demand in a manner quite different from ours. We thank an anonymous referee for pointing out this difference. Fosgerau, Paulsen, and Rasmussen (2021b) present a discrete choice model in which most choice alternatives have a true probability of zero.
single agent chooses from \( J + 1 \) mutually exclusive alternatives labelled \([0, 1, \ldots, J]\) := \( J \). The utility the agent derives from choosing alternative \( j \) is

\[
\pi_j(X) + \varepsilon_j, \quad j \in J,
\]

where \( \pi_j(X) \) denotes the systematic components of utility, and \((\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_J)\) the random components (random elements of \( \mathbb{R}^{J+1} \)). Each function \( \pi_j \) translates conditioning variables \( X \), having support \( \mathcal{X} \), into “utils.” Conditioning variables \( X \) are observable to both the agent and the researcher. While in empirical applications, the functions \( \pi_j : \mathcal{X} \to \mathbb{R}, j \in J \), are typically parametrically specified, our treatment of these will be nonparametric. The \( \varepsilon_j \)'s are random variables observed only by the agent, defined on a given probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \).

The agent chooses an alternative \( j \) that maximizes \( \pi_j(X) + \varepsilon_j \). Given that the solution to this optimization problem only depends on utility differences (calculated with respect to any reference alternative), we may normalize \( \pi_0(X) = 0 \) for all \( X \in \mathcal{X} \) and \( \varepsilon_0 \equiv 0 \) and, thus, treat alternative 0 as the “outside option.”

This normalization allows us to formulate the agent’s problem in a notationally convenient manner. For this purpose, let \( \{u_j\}_{j \in J} \) be the canonical basis for \( \mathbb{R}^J \) and write \( u_0 := 0 \) for the origin in \( \mathbb{R}^J \). The latter vector “points” towards the outside option. The agent’s problem may then be reformulated as

\[
\max_{j \in J} \left\{ \pi_j(X) + \varepsilon_j \right\} = \max_{j \in J} \left\{ \pi(X) + e, u_j \right\},
\]

where \( \pi(X) := (\pi_1(X), \ldots, \pi_J(X)) \) collects the (normalized) systematic components and \( e := (\varepsilon_1, \ldots, \varepsilon_J) \). We define the argmax correspondence \( \mathcal{A} : \mathbb{R}^J \Rightarrow \mathcal{J} \) by

\[
\mathcal{A}(v) := \text{argmax}_{j \in J} \{v, u_j\},
\]

such that the solution to the agent’s problem is \( \mathcal{A}(\pi(X) + e) \), a random subset of \( \mathcal{J} \).

We denote the cumulative distribution function (CDF) of \( e \) conditional on \( X \) by \( F_e|X \).\(^6\) Initially impose no conditions on this CDF. Our analysis thus differs from most of the literature, which assumes the conditional distribution absolutely continuous (with respect to Lebesgue measure) with finite means (i.e., \( \mathbb{E}[|e_j|] < +\infty \) for all \( j \)). When discussing identification we follow Chong et al. (2016) among others (see the Introduction) and treat \( F_e|X \) as known to the reader.\(^7\)

A fundamental entity in our analysis is the surplus function \( W(\cdot|X) : \mathbb{R}^J \to \mathbb{R} \) (McFadden, 1978, 1981), defined for each fixed \( X \in \mathcal{X} \) by

\[
W(v|X) := \mathbb{E} \left[ \max_{j \in J} \{v + e, u_j\} \mid X = x \right].
\]  

(1)

The function \( W(\cdot|X) \) may be interpreted as the expected surplus of the agent when faced with utilities \( v \) relative to zero utilities. Note here that the vector \( v \) takes the place of \( \pi(X) \), which provides us with the ability to vary \( v \), e.g., for the purpose of computing derivatives, while holding \( X \) constant. In the definition of \( W(\cdot|X) \), we follow McFadden (1981) and subtract the maximum of the random utility components. This subtraction ensures that the expectation exists, even if \( e \) is not integrable (in the sense that \( \mathbb{E}[|e_j|] = +\infty \) for some \( j \)). Indeed, the function \( \max_{j \in J} \{v, u_j\} \) is Lipschitz continuous with respect to \( \|\cdot\|_{\infty} \) with Lipschitz constant 1, which implies that for all \( v \in \mathbb{R}^J \),

\[
\mathbb{E} \left[ \max_{j \in J} \{v + e, u_j\} - \max_{j \in J} \{e, u_j\} \right] \leq \max_{1 \leq i \leq J} |v_i| < +\infty.
\]

It follows from the Radon–Nikodym theorem that the surplus function in (1) is always well defined; no conditions need to be placed on the distribution of \( (X, e) \). Beyond ensuring the well-definedness of the surplus, subtracting the maximum of the random utility components has no substantial effect on its shape.\(^8\)

In contrast to most of the existing literature, we do not specify how the agent breaks utility ties and allow any rule consistent with utility maximization. Consequently, our approach allows for incomplete models.\(^9\) For each \( x \in \mathcal{X} \), we define the CCP correspondence \( \mathcal{P}(\cdot|X) : \mathbb{R}^J \Rightarrow \mathbb{R}^J \) as the set of all CCP vectors that are consistent with utility-maximizing choices. This is the set expectation

\[
\mathcal{P}(v|X) := \mathbb{E} \left[ \text{conv} \left\{ \{u_j\}_{j \in \mathcal{J}(v + e)} \right\} \mid X = x \right],
\]

where the right-hand side is defined as

\[
\{ \mu \in \mathbb{R}^J ; \mu = \mathbb{E}[\xi|X = x] \text{ for } \xi : \Omega \to \mathbb{R} \text{ measurable} \}

satisfying \( \mathcal{P}(\xi|X) = \{x \} \Rightarrow \mathbb{R}^J \text{ measurable} \).

In this definition, we may think of different \( \xi \)'s as choice rules consistent with utility maximization but employing different tie-breaking rules. An element of \( \mathcal{P}(\cdot|X) \) is a choice probability vector, which is the expected value of the choice rule given \( X = x \).\(^{10}\) In Section 5, we provide necessary and sufficient conditions for \( \mathcal{P}(\cdot|X) \) and its inverse to reduce to functions (see Theorems 4 and 5, respectively).

Furthermore, given the specific choice rule \( \xi^* \) actually employed by the utility maximizing agent, the conditional choice probability (CCP) function \( \mathcal{P} : \mathcal{X} \to \mathbb{R}^J \) is defined by

\[
\mathcal{P}(\cdot|X) := \mathbb{E}[\xi^*|X = x].
\]

Then in particular, \( \mathcal{P}(\pi(X)|X) \) for each \( X \in \mathcal{X} \). We note that \( \mathcal{P}(\pi(X)|X) \in \Delta, X \in \mathcal{X} \), where \( \Delta \) is the unit simplex,

\[
\Delta := \text{conv} \{ u_j \}_{j \in J} = \left\{ q \in \mathbb{R}^J ; q_j \geq 0, \sum_{j=1}^J q_j \leq 1 \right\}
\]

Thus, the probabilities in \( \mathcal{P}(\pi(X)|X) \) may sum to less than one, and the probability of the outside option is the residual \( \mathcal{P}(\xi^* = u_0|X = x) = 1 - \sum_{j=1}^J p_j(X) \).

3. The Williams–Daly–Zachary Theorem

Denote the domain of a correspondence \( \rho : \mathbb{R}^m \Rightarrow \mathbb{R}^m \) by \( \text{dom} \rho := \{x \in \mathbb{R}^m ; \rho(x) \neq \emptyset \} \).

Our first result is a generalization of the famous Williams–Daly–Zachary (WDZ) Theorem to the present general case where the conditional distribution of the random utility components is unrestricted.

Theorem 1 (General Williams–Daly–Zachary). For any fixed \( x \in \mathcal{X} \), the surplus function \( W(\cdot|X) \) is everywhere subdifferentiable, \( \text{dom} \partial W (\cdot|X) = \mathbb{R}^J \), and its subdifferential coincides with the CCP correspondence, \( \partial W (\cdot|X) = \mathcal{P}(\cdot|X) \). In addition,

\footnote{In the literature it is quite standard to define the surplus by \( W(v|X) := \mathbb{E}[\max_{j \in \mathcal{J}(v + e)}(v + e, u_j)|X = x] \). However, this definition requires the \( e_j \)'s to be integrable, which rules out fat-tailed distributions, such as the Cauchy family.}

\footnote{Koning and Ridder (2003) refer to incomplete models as incoherent.}

\footnote{Note that each \( \xi \) is almost surely confined to a bounded set and, therefore, automatically integrable, in the sense that each coordinate of \( \xi \) is integrable. Therefore \( \mathcal{P}(\cdot|X) \) is a well-defined correspondence.}
1. The surplus function \( W(\cdot | x) \) is finite, convex and Lipschitz continuous with respect to \( \| \cdot \|_\infty \) with Lipschitz constant 1.

2. The CCP correspondence \( \mathcal{P}(\cdot | x) \) is cyclic monotone with domain \( \text{dom} \mathcal{P}(\cdot | x) = \mathbb{R}^J \).

Recall the classical WDZ finding that if the conditional distribution \( F_{\mathcal{X}} \) of \( e \) given \( X = x \) is absolutely continuous, then the surplus \( W(\cdot | x) \) is everywhere differentiable with gradient coinciding with the CCP function \( \mathbf{v} \mapsto E[U_{\mathcal{A}(\mathbf{v}^* \mathcal{F})}|X = x] \). Theorem 1 generalizes this result to the case where the random utility components may follow a completely arbitrary distribution and where the utility maximizing agent may use any tie-breaking rule.

Shi et al. (2018) innovatively combine the classical (differential) WDZ theorem with cyclic monotonicity of the surplus gradient to derive identifying inequalities for the systematic utilities \( \pi(x) \). They also construct an estimator (of the structural parameters) based on the sample analogue of these inequalities upon replacing the actual CCPs \( p(x) \) with an estimator thereof. Theorem 1 shows that the CCP correspondence is cyclic monotone, such that the inequality
\[
\langle q_0, v_0 - v_1 \rangle + \langle q_1, v_1 - v_2 \rangle + \cdots + \langle q_n, v_n - v_0 \rangle \geq 0
\]
holds for any integer \( n \geq 1 \) and any sequence \( \{(v_i, q_i)\}_{i=0}^n \) of pairs such that \( q_i \in \mathcal{P}(v_i|x) \). The set \( \mathcal{P}(\pi(x)|x) \) contains all CCP vectors consistent with utility-maximizing behaviour, including \( p(x) \). The property of cyclic monotonicity is thus informative for identifying utilities regardless of how ties are broken. Hence, Theorem 1 allows us to extend the reasoning of Shi et al. (2018) to allow for arbitrary distributions. Note that their approach to identification and estimation does not require knowledge of the (conditional) distribution of random utility components. Our extension to their work thus illustrates an application of our results in which we may dispose of this assumption.

### 4. Hotz-Miller inversion

The general WDZ theorem provides a possibly set-valued mapping from utilities to CCPs consistent with utility maximization. The issue of identification is an inverse problem: Given the CCPs \( p(x) \), can one recover the systematic utilities \( \pi(x) \)? Appealing to Theorem 1, in our search for the correspondence inverse \( \mathcal{P}^{-1}(\cdot | x) \) at a vector \( q \in \mathbb{R}^J \), we are asking for the set of vectors \( \mathbf{v} \) satisfying \( q \in \partial W(\cdot | x) \). The correspondence inverse is conveniently expressed by means of Fenchel conjugation.

Given an \( x \in \mathcal{X} \), we define the conjugate surplus function \( \widetilde{W}^*(\cdot | x) : \mathbb{R}^J \rightarrow (-\infty, +\infty] \) as the convex conjugate of \( W(\cdot | x) \):
\[
\widetilde{W}^*(q|x) := \sup_{\mathbf{v} \in \mathbb{R}^J} \{ \langle \mathbf{v}, q \rangle - W(\mathbf{v}|x) \}, \quad q \in \mathbb{R}^J.
\]
Finiteness of the surplus (Theorem 1.1) ensures that its conjugate is never \(-\infty\). When the ARUM is a multinomial logit model, the conjugate surplus function is the (negative) Shannon entropy (Anderson et al., 1988). For any distribution of random utility, the (negative) conjugate surplus function may thus be viewed as a generalized entropy (Galichon and Salanie, 2020; Fosgerau, Melo, de Palma, and Shum, 2020).

Denote the image of a correspondence \( \rho : \mathbb{R}^m \Rightarrow \mathbb{R}^m \) by
\[
\text{Im} \rho := \{ y \in \mathbb{R}^m; \mathbf{y} \in \rho(x) \text{ for some } x \in \mathbb{R}^m \} = \bigcup_{x \in \mathbb{R}^m} \rho(x).
\]

Our next result is a generalization of the Hotz and Miller (1993) inversion result to the present general case where the conditional distribution of the random utility components is unrestricted.

**Theorem 2 (General Hotz–Miller).** For any \( x \in \mathcal{X} \), the conjugate surplus \( \widetilde{W}^*(\cdot | x) \) is subdifferentiable on \( \text{int} \Delta \subseteq \partial \mathcal{W}^*(\cdot | x) \subseteq \Delta \), and its subdifferential coincides with the inverse CCP correspondence, \( \mathcal{P}^{-1}(\cdot | x) \) is continuous on \( \text{int} \Delta \).

1. The conjugate surplus \( \widetilde{W}^*(\cdot | x) \) is lower semi-continuous and proper convex with effective domain \( \text{int} \Delta \subseteq \text{dom} \mathcal{W}^*(\cdot | x) \subseteq \Delta \), in particular, \( \mathcal{W}^*(\cdot | x) \) is continuous on \( \text{int} \Delta \).

2. The inverse CCP correspondence \( \mathcal{P}^{-1}(\cdot | x) \) is cyclic monotone with domain \( \text{int} \Delta \subseteq \text{dom} \mathcal{P}^{-1}(\cdot | x) \subseteq \text{Im} \mathcal{P}(\cdot | x) \subseteq \Delta \).

The original Hotz and Miller (1993, Proposition 1) states that if the conditional distribution \( F_{\mathcal{X}} \) of \( e \) given \( X = x \) is absolutely continuous with full support, then the Hotz and Miller (1993) (or CCP) function
\[
\mathbf{v} \mapsto \phi(\mathbf{v}|x) := E[U_{\mathcal{A}(\mathbf{v}^* \mathcal{F})}|X = x]
\]
is injective, such that one may speak of an (Hotz–Miller) inverse function. In contrast, when utility ties have positive probability, both \( \phi(\cdot | x) \) and otherwise equivalent definitions are ill-defined as functions mapping into the probability simplex \( \Delta \). For this reason, one cannot in general speak of “Hotz–Miller inversion” in the classical sense. However, the inverse CCP correspondence \( \mathcal{P}^{-1}(\cdot | x) \) is always well defined and nonempty-valued everywhere on (at least) int \( \Delta \) (cf. Theorem 2).

The properties stated in Theorem 2 are derived using ‘zero-order’ convex analysis alone and do not rely on any differential theory. They correspond via convex duality to the properties in Theorem 1. Lower semi-continuity, properness and convexity are generic to convex conjugates of proper convex functions (Rockafellar, 1970, Theorem 12.2). A convex function is continuous on any open subset of its effective domain (Rockafellar, 1970, Theorem 10.1). Theorem 2 shows that the effective domain of every conjugate surplus function contains the simplex interior. One implication of the theorem is therefore that continuity on the simplex interior is a property common to all conjugate surplus functions.

The effective domain \( \text{dom} \mathcal{W}^*(\cdot | x) \) does not include the simplex boundary in general. In the context of binary choice \( J = 1 \), in Appendix C we show that whether or not the boundary is included boils down to whether or not the (scalar) random utility is integrable. More broadly, one may show that \( \text{dom} \mathcal{W}^*(\cdot | x) \) is the entire simplex whenever all the \( s_j \)'s are integrable. Continuity
(relative to the closed set $\Delta$) then follows from Rockafellar (1970, Theorem 10.2).

From Theorem 2 and the variational representation of subdifferentials of proper conjugate pairs [see (A.4) in Appendix A], we see that the inverse CCP correspondence $\mathcal{P}^{-1}(\cdot|\cdot)$ may be evaluated as

$$\mathcal{P}^{-1}(q|\cdot) = \partial W^*(q|\cdot) = \text{argmax}_{v \in \mathbb{R}} \{\langle v, q \rangle - W(v|x)\}.$$ 

This observation leads to a partial identification result.

**Theorem 3 (Partial Identification).** Fix $x \in \mathcal{X}$. Then the utility values $\pi(x)$ are partially identified with identified set being the closed convex set

$$\mathcal{P}(x) := \partial W^*(p(x)|x) = \text{argmax}_{v \in \mathbb{R}} \{\langle v, p(x) \rangle - W(v|x)\}. \quad (6)$$

The identified set is both nonempty and bounded (hence compact) if and only if $p(x) \in \text{int} \Delta$.

Theorem 3 provides constructive partial identification of the utility values $\pi(x)$. The theorem also shows that if the true CCPs lie on the boundary of the probability simplex, then systematic utilities cannot be point identified as the identified set must be either empty or unbounded. The possibility of unboundedness is intuitive. For example, if we consider the extreme case where one alternative attracts all the demand ($p(x) = u_1$), then provided that some utility vector $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$ is able to rationalize the distribution of choices, we can assign any utility vector $v' := (v_1', v_2', \ldots, v_n')$ and the alternative while leaving the remaining elements unchanged. Note that the potential unboundedness does not result from a lack of normalization as we have already taken $\pi_0 = 0$.

While a formal treatment of the topic of estimation lies outside the scope of this paper, Eq. (6) is suggestive of multiple estimation strategies. First, when an analytic expression for $\partial W^*(\cdot|x)$ is available, then one may immediately obtain an estimator $\hat{\mathcal{P}}(x) := \partial W^*(\hat{p}(x)|x)$ of the identified set $\mathcal{P}(x)$ by substituting the true CCPs with an estimator $\hat{p}(x)$. Second, when such an analytic expression is unavailable or difficult to come by, one may still use the “argmax” characterization in (6) to construct a set estimator based on an approximation $W(\cdot|x)$ of the surplus $W(\cdot|x)$. In the latter case, the set-estimated estimator takes the form $\partial W^*(\hat{p}(x)|x)$, with $W^*(\cdot|x)$ denoting the convex conjugate of the approximate surplus (which we here presume to be finite convex).

Note that if no one chooses a particular alternative in the data, then the empirical CCPs lie on the boundary of the probability simplex. The resulting estimator of the identified set remains well defined. In this sense, our approach to identification is robust to the “zeros-in-demand problem” mentioned in the introduction. Zero empirical probabilities imply that the resulting utility set estimate is either empty or unbounded. Either finding is informative in the sense of narrowing down the candidate utility space ($\mathbb{R}^n$). Theorem 3 is closely related to Chiong et al. (2016, Theorem 4), who express the identified set as part of the solution to the dual of a possibly infinite-dimensional optimal assignment problem. Their result similarly does not rest on the distribution of the random utility components being smooth or having full support. The main difference between the two findings thus lies in the resulting characterizations. Specifically, since their optimal assignment problem amounts to (possibly infinite-dimensional) linear programming, their characterization of the identified set amounts to finding the optimal solutions from a system of (possibly infinitely many) linear inequalities. In contrast, Theorem 3 characterizes the identified set using convex duality. Both characterizations are constructive.

In Section 5 we fully characterize the distributions of random utility components resulting in a differentiable surplus in terms of a continuity condition [see Theorem 4 and Condition C(x), in particular]. In addition, we show that a condition of “sufficiently rich” support controls the strictness of convexity of the surplus and also dictates the differentiability of its conjugate [see Theorem 5 and Condition S(x), in particular].

**Example 1 (Two-Point Support and Partial Identification).** Consider the case of binary choice ($J = \{0, 1\}$), and drop the “1” subscripts throughout. Let conditioning variables be deterministic, such that $X = x$ for some fixed vector $x$. Then the systematic utility function boils down to the scalar $\pi := \pi(x)$. Let $e$ be Rademacher distributed, such that $P(e = -1) = P(e = 1) = \frac{1}{2}$. Then

$$W(v) = \begin{cases} \frac{1}{2}v, & v \in [-1, 1], \\ v - \frac{1}{2}, & v > 1, \end{cases}$$

which is piecewise affine and therefore everywhere differentiable except at line-segment endpoints. Differentiating on the interior of each line segment and closing the gaps, we see that the surplus subdifferentiable (and thus CCP correspondence) takes the form:

$$\mathcal{P}(v) = \partial W(v) = \begin{cases} [0], & v < -1, \\ [0, 1] & v = -1, \\ \{\frac{1}{2}\}, & v \in (-1, 1), \\ [\frac{1}{2}, 1], & v = 1, \\ [1], & v > 1. \end{cases}$$

These are the probability levels (for alternative 1) consistent with utility maximization as a function of the utility level. For example, if $\pi \in (-1, 1)$, then $\mathcal{P}(\pi) = \{\frac{1}{2}\}$, reflecting that each alternative is chosen with equal probability, and if $\pi > 1$, then $\mathcal{P}(\pi) = \{1\}$, reflecting that alternative 1 is always optimal. In the (boundary) case $\pi = 1$, the agent is indifferent whenever $e = -1$. Since the latter occurs with probability $\frac{1}{2}$, depending on the tie-breaking involved, we may observe any probability level $[\frac{1}{2}, 1]$.

Inversion of the CCP correspondence and evaluation of the inverse at the true choice probability $p$ produces the identified set. The desired inverse may here be obtained by mirroring the graph of the CCP correspondence in the 45 degree line. Observing $p = \frac{1}{2}$ leads to the identified set $\mathcal{P}^{-1}(\frac{1}{2}) = [-1, 1]$, as illustrated in Fig. 1. This set is compact and convex and contains $(-1, 1)$. Observing $p = 1$ leads to $\mathcal{P}^{-1}(1) = [1, +\infty)$, which is closed and convex but unbounded. Observing any $p$ strictly in between leads to point identification. In this example, the image of the CCP correspondence is here the entire unit simplex.

5. Differential theory and point identification

Up to this point, we have derived generic properties of the surplus function, its conjugate and the CCP correspondence, and translated these properties into a partial identification result. We next provide conditions under which stronger properties may be obtained and translate these properties into a point identification result.

A proper convex function $f$ is called almost differentiable if it is differentiable on the interior of its effective domain, but subdifferentiable nowhere else. Intuitively, the latter qualification means that the graph of $f$ becomes infinitely steep as we approach the domain boundary from within. For finite functions, the boundary qualification is vacuous, and almost differentiability reduces to differentiability.

The dual notion to almost differentiability is almost strict convexity. A proper convex function $f$ is almost strictly convex if it is strictly convex on every convex subset of dom $\partial f$. 
To control the smoothness/convexity of the surplus and its conjugate, we introduce two conditions, the first being

\textbf{Condition C}(x). Fix \( x \in X \). For all \( v, h \in \mathbb{R}^J \),

\[ P\left( \max_{j \in J} \langle h, u_j \rangle \geq \min_{j \in J} \langle h, u_j \rangle \big| X = x \right) = 0. \] (7)

To interpret this condition, recall that a function \( f \) finite at \( x \) is one-sided directionally differentiable at \( x \) in the direction \( h \) if the limit

\[ \lim_{\lambda \to 0^+} \frac{f(x + \lambda h) - f(x)}{\lambda} \]

exists (in \( \mathbb{R} \)), where \( \lambda \to 0_+ \) is short for “as \( \lambda \geq 0 \) approaches zero.” It is directionally differentiable at \( x \) in the direction \( h \) if limits exists for both \( h \) and \( -h \) and these limits coincide, and (fully) differentiable at \( x \) if this conclusion holds true for all directions \( h \). The event involved in (7) is equivalent to the max function \( \max_{j \in J} \langle \cdot + e, u_j \rangle \) failing to be directionally differentiable at \( v \) in the direction \( h \). Eq. (7) makes it explicit that nondifferentiability may occur only in the event of utility ties, as otherwise \( \mathcal{A}(v + e) \) is a singleton and the involved “max” and “min” would coincide. Condition C(\( x \)) forces the conditional distribution of \( e \) given \( X = x \) to view points leading to ties as negligible. It is therefore a continuity condition.

More precisely, Condition C(x) translates into the following properties and relations.

\textbf{Theorem 4 (Differential Williams–Daly–Zachary).} Fix \( x \in X \). Then Condition C(\( x \)) is equivalent to each of the following:

1. \( P(\cdot | x) \) is single-valued.
2. \( W(\cdot | x) \) is differentiable.
3. \( W^\ast(\cdot | x) \) is almost strictly convex.

In this case, \( P(\cdot | x) \) reduces to the gradient mapping \( \nabla W(\cdot | x) \), i.e. for all \( v \in \mathbb{R}^J \), \( P(v | x) \) consists of the vector \( \nabla W(v | x) \) alone.

Theorem 4 shows that the surplus function is differentiable if and only if Condition C(\( x \)) holds, which, in turn, is equivalent to \( P(\cdot | x) \) reducing to a function. Condition C(\( x \)) thus constitutes a necessary and sufficient condition for the Hotz and Miller (1993) function \( \phi(\cdot | x) \) in (5) to be well defined. Letting \( v = \pi(\cdot | x) \), Theorem 4 provides a modest extension to the classical, differential WDZ theorem by not requiring the existence of a Lebesgue density. That \textit{absolute} continuity is not necessary for the conclusion of the differential WDZ theorem has been remarked by Norets and Takahashi (2013). However, the necessity of Condition C(\( x \)) appears to be new.

For ease of reference, we here provide a simple condition sufficient for Condition C(\( x \)).

\textbf{Proposition 1.} Fix \( x \in X \) and suppose that the conditional distribution of \( e \) given \( X = x \) assigns zero measure to every hyperplane in \( \mathbb{R}^J \). Then Condition C(\( x \)) holds.

A hyperplane has zero Lebesgue measure, so the (classical) assumption of absolute continuity of the conditional distribution of \( e \) given \( X = x \) suffices for Condition C(\( x \)).

To further characterize the smoothness/convexity of the surplus and its conjugate, we also introduce
Condition $S(x)$. Fix $x \in \mathcal{X}$. For all $v, h \in \mathbb{R}^J$, $h \neq 0$,
\[
P\left(\max_{j \in J} \langle v + h + \epsilon, u_j \rangle > \max_{j \in J} \langle v + \epsilon, u_j \rangle + \max_{j \in J, \Delta} \langle h, u_j \rangle | X = x \right) > 0.
\]
(8)

One may show that the event in (8) is equivalent to $\max_{j \in J} \langle v + h + \epsilon, u_j \rangle$ being nonconstant on the line segment connecting $v$ and $v + h$, $h \neq 0$. Condition $S(x)$ is therefore a condition of sufficiently rich support. Like Condition $C(x)$, Condition $S(x)$ does not require the existence of a Lebesgue density. Moreover, Condition $S(x)$ may be satisfied with less than full support. For example, the support may have holes of finite diameter.\(^\text{17}\)

Condition $S(x)$ corresponds to the following properties and relations.

**Theorem 5** (Differential Hotz–Miller). Fix $x \in \mathcal{X}$. Then Condition $S(x)$ is equivalent to each of the following:

1. $\mathcal{P}^{-1} (\cdot | x)$ is single-valued.
2. $W (\cdot | x)$ is strictly convex.
3. $W^* (\cdot | x)$ is almost differentiable.

In this case, $\mathcal{P}^{-1} (\cdot | x)$ reduces to the gradient mapping $\nabla W^* (\cdot | x)$, i.e. $\mathcal{P}^{-1} (q | x)$ consists of the vector $\nabla W^* (q | x)$ alone for $q \in \text{int } \Delta$, while $\mathcal{P} (q | x) = \emptyset$ for $q \notin \text{int } \Delta$. In particular, dom $\partial W^* (\cdot | x) = \text{int } \Delta$.

While the Hotz and Miller (1993) function $\phi (\cdot | x)$ given in (5) need not be defined as a map from utilities to choice probabilities (see the discussion following **Theorem 2**), **Theorem 5.1** shows that our support condition, Condition $S(x)$, is both necessary and sufficient for our inverse CCP correspondence $\mathcal{P}^{-1} (\cdot | x)$ to reduce to a function.

We also provide a simple condition sufficient for Condition $S(x)$.

**Proposition 2.** Fix $x \in \mathcal{X}$ and suppose that the conditional distribution of $\epsilon$ given $X = x$ assigns positive measure to every nonempty open subset of $\mathbb{R}^J$. Then Condition $S(x)$ holds.

Every nonempty open set has positive Lebesgue measure, so the (classical) assumption of full support of the conditional distribution of $\epsilon$ given $X = x$ suffices for Condition $S(x)$.

**Theorems 4 and 5** are dual to each other. We have stated them as separate theorems to highlight the importance of Conditions $C(x)$ and $S(x)$ in turn. Combining the two theorems yields the following strong result.

**Theorem 6** (Homeomorphic Demand). Fix $x \in \mathcal{X}$. Then Conditions $C(x)$ and $S(x)$ are jointly equivalent to each of the following:

1. $\mathcal{P} (\cdot | x)$ is one-to-one.
2. $W (\cdot | x)$ is differentiable and strictly convex.
3. $W^* (\cdot | x)$ is almost differentiable and almost strictly convex.

In this case, $\mathcal{P} (\cdot | x)$ reduces to a one-to-one function from $\mathbb{R}^J$ onto $\text{int } \Delta$, continuous in both directions, with $\mathcal{P} (\cdot | x) = \{\nabla W (\cdot | x)\}$ on $\mathbb{R}^J$ and $\mathcal{P}^{-1} (\cdot | x) = \{\nabla W^* (\cdot | x)\}$ on $\text{int } \Delta$.

**Theorem 6** has several implications. First, under the conditions of the theorem, the CCP correspondence $\mathcal{P} (\cdot | x)$ reduces to the Hotz and Miller (1993) function $\phi (\cdot | x)$, such that **Theorem 6** extends Hotz and Miller (1993, Proposition 1) to allow for not necessarily absolutely continuous stochastic utility distributions with less than full support. Second, this function is a bijection between $\mathbb{R}^J$ and $\text{int } \Delta$, such that the theorem reproduces Norets and Takahashi (2013, Corollary 1) under weaker assumptions. Finally, **Theorem 6** reveals an added benefit to our convex analysis approach: $\mathcal{P} (\cdot | x)$ reduces to a bijection, which is continuous in both directions, i.e. a homeomorphism.\(^\text{18}\) Consequently, under Conditions $C(x)$ and $S(x)$, the Hotz and Miller (1993) inverse function $\phi^{-1} (\cdot | x)$ is not only well defined—it is also well behaved.

Arguing as in the lead-up to **Theorem 3**, evaluating the inverse CCP correspondence at $v = \pi (x)$ now produces a result which concerns point identification.

**Theorem 7** (Point Identification). Fix $x \in \mathcal{X}$ and let Conditions $C(x)$ and $S(x)$ hold. Then the utility values $\pi (x)$ are point identified by

$$
\pi (x) = \nabla W^* (\langle p (x) | x \rangle) = \arg\max_{v \in \mathbb{R}^J} \{\langle v, p (x) \rangle - W (v | x)\}.
$$

(9)

Given that the differential WDG theorem is a discrete-choice analogue of Roy’s identity, **Theorem 7** may be interpreted as a dual version of Roy’s identity. Being a special case of our constructive set-identification result (**Theorem 3**), our point-identification result is also constructive. As mentioned in the Introduction, closely related duality results have been used for the purpose of analysing the convergence properties of stochastic fictitious play in certain classes of games (Holbauer and Sandholm, 2002, Theorem 2.1), as well as for identification in two-sided matching models (Galichon and Salanie, 2020, Proposition 2) and (single-agent) dynamic discrete choice (Chiong et al., 2016, Theorem 3), all of which treat the distribution(s) of unobserved heterogeneity as known.\(^\text{19}\) Since static discrete choice may be viewed as a special case of dynamic discrete choice or as one-sided matching, the main contribution of **Theorem 7** lies in providing conditions that are not only sufficient for point identification, but also necessary.

For ease of reference, we end this section by restating the conclusions from **Theorem 6** under classical assumptions.

**Corollary 1.** Fix $x \in \mathcal{X}$. If $F_x (\cdot | x)$ is absolutely continuous with full support on $\mathbb{R}^J$, then the following statements hold:

1. $\mathcal{P} (\cdot | x)$ is one-to-one.
2. $W (\cdot | x)$ is everywhere differentiable and strictly convex.
3. $W^* (\cdot | x)$ is almost differentiable and almost strictly convex.

Consequently, $\mathcal{P} (\cdot | x)$ reduces to a one-to-one function from $\mathbb{R}^J$ onto $\text{int } \Delta$, continuous in both directions, with $\mathcal{P} (\cdot | x) = \{\nabla W (\cdot | x)\}$ on $\mathbb{R}^J$ and $\mathcal{P}^{-1} (\cdot | x) = \{\nabla W^* (\cdot | x)\}$ on $\text{int } \Delta$.

6. Conclusion

In this paper we have synthesized and extended a range of classical results for the additive random utility model (ARUM) of discrete choice, utilizing the power of convex analysis. The generalization of the Williams–Daly–Zachary Theorem employs no assumptions on the structure of utility other than additivity, which means no further generalization is possible. A general Hotz–Miller inverse always exists in the form of an inverse conditional choice probability (CCP) correspondence, mapping positive probability vectors to compact and convex sets of utilities. When the distribution of random utility components is known, this inverse

\(^{17}\) Note that Condition $S(x)$ does not require the support to be a connected set. Connectedness is part of the premise of the bijectivity result of Norets and Takahashi (2013, Corollary 1).

\(^{18}\) A classical result shows that the inverse $f^{-1}$ of a one-to-one, continuous mapping $f$ from $X \subseteq \mathbb{R}^n$ onto $Y \subseteq \mathbb{R}^m$ is continuous if $X$ is compact (Nikaido, 1968, Theorem 1.4). However, continuity of the inverse function need not hold in the absence of a compact domain, see Nikaido (1968, p. 9) for an example.

\(^{19}\) See also Allen and Rehbeck (2019, Lemma 3), who establish a dual version of “Roy’s Identity” for an entire class of latent utility models with additively separable heterogeneity under an independence assumption.
CCP correspondence provides constructive partial identification of utilities. The identified set is the solution to a convex optimization problem, defined in terms of the ARUM surplus function. While a formal treatment of the topic of estimation lies outside the scope of this paper, the constructive nature of our identification results suggests strategies for estimation. Resulting estimators remain well defined in the presence of zeros in empirical probabilities and are therefore robust to the “zeros-in-demand” problem commonly encountered in modern datasets.

Without any restrictions on the distribution of random utility components, utility maximization does not uniquely determine utilities, treating the distribution of random utility components, utility maximization does not uniquely determine common well defined in the presence of zeros in empirical probabilities. Weby and large adopt the terminology and notation of Rockafellar (1970), (A.2) for any integer \( n \geq 1 \) and any set \( \{(x_i, x'_i)\}_0^n \) of pairs such that \( x_i^c \in \rho (x_i) \).

The concept of cyclic monotonicity generalizes the notion of monotonicity for univariate functions to possibly multivariate functions. To establish this connection, recall that a correspondence \( \rho : \mathbb{R}^m \rightrightarrows \mathbb{R}^m \) is called single-valued if \( \partial \rho (x) \) contains at most one element for all \( x \). A single-valued correspondence \( \rho \) may be identified with a function \( f : \mathbb{R}^m \to \mathbb{R}^m \) with domain \( D := \text{dom} \rho = \{ x \in \mathbb{R}^m : \rho (x) \neq \emptyset \} \). Cyclic monotonicity of \( \rho \) then boils down to the requirement that

\[
\langle f (x_0) - x_1, x_0 - x_1 \rangle + \langle f (x_0) - x_2, x_1 - x_2 \rangle + \cdots + \langle f (x_n) - x_0, x_n - x_0 \rangle \geq 0 \tag{A.3}
\]

for any integer \( n \geq 1 \) and any set \( \{ x_i \}_0^n \) of points in \( D \). In accordance with the definition employed by Shi et al. (2018), we therefore take the previous display to be the definition of cyclic monotonicity for functions.

We will make use of the following result linking cyclic monotonicity and convexity.

**Proposition A.1 (Cyclic Monotonicity of Subdifferentials).** Let \( f \) be proper convex. Then \( df \) is cyclic monotone.

The proposition is a consequence of the subgradient inequality (see, e.g., Rockafellar, 1970, p. 238). Any univariate differentiable convex function must admit a monotone increasing derivative. As shown in Proposition A.1, cyclic monotonicity extends the notion of a monotone increasing derivative to multivariate convex functions not necessarily differentiable.

Given a convex function \( f : \mathbb{R}^m \to (-\infty, +\infty) \), we define its convex (or Fenchel) conjugate \( f^* \) by

\[
f^* (x^*) := \sup_{x \in \mathbb{R}^m} \left\{ \langle x, x^* \rangle - f(x) \right\}, \quad x^* \in \mathbb{R}^m.
\]

The convex conjugate \( f^* \) is proper if and only if \( f \) itself is proper. When the latter holds, \( f^* \) may thus be viewed as a (convex) function from \( \mathbb{R}^m \) to \(( -\infty, +\infty) \). The subdifferential \( df^* \):

\[
\text{dom} \; df^* := \{ x \in \mathbb{R}^m : df (x) \neq \emptyset \}.
\]

is called the domain of \( df \). Unlike the effective domain of a convex function, the domain of its subdifferential need not be convex.

A finite convex function is everywhere subdifferentiable, but not necessarily differentiable. If \( f \) is in fact differentiable at a point \( x \), then \( df (x) \) consists of the single vector given by the gradient \( \nabla f (x) \) of \( f \) at \( x \). Conversely, if \( df (x) \) is the singleton \( \{ x^* \} \), then \( f \) is differentiable at \( x \) with \( \nabla f (x) = x^* \). The notion of subdifferentiability therefore extends the notion of differentiability for convex functions.

Subdifferentiability is tightly connected to the notion of cyclic monotonicity. A correspondence \( \rho : \mathbb{R}^m \rightrightarrows \mathbb{R}^m \) is called cyclic monotone if

\[
\{ x_0^c, x_0 - x_1 \} + \{ x_1^c, x_1 - x_2 \} + \cdots + \{ x_n^c, x_n - x_0 \} \geq 0 \tag{A.2}
\]

We thank the editor and reviewers for very helpful and constructive comments. We thank Alfred Galichon, Emerson Melo, John Rust, Bernard Salanié, and Matt Shum for useful comments and discussion. M. Fosgerau and J. R.-V. Sørensen have received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 740369).

### Appendix A. Convex analysis

For easy reference, we here gather some key definitions and results from convex analysis.\footnote{We by and large adopt the terminology and notation of Rockafellar (1970), which we find to be an excellent reference on the topic.} A function \( f : \mathbb{R}^m \to (-\infty, +\infty] \) is convex if

\[
f ((1 - \lambda) x + \lambda y) \leq (1 - \lambda) f (x) + \lambda f (y), \quad 0 < \lambda < 1, \tag{A.1}
\]

for every \( x \) and \( y \) in \( \mathbb{R}^m \), and strictly convex if the inequality in (A.1) is strict whenever \( x \) and \( y \) differ. Its effective domain, denoted \( \text{dom} f \), is the subset of \( \mathbb{R}^m \) where \( f \) is finite,\footnote{”Finite” is here synonymous with “real-valued”. We avoid the latter term to avoid confusing real-valued and extended real-valued functions.} a convex set. Call a function \( f : \mathbb{R}^m \to (-\infty, +\infty] \) proper provided that it is nowhere \(-\infty\) and not identically \(+\infty\). Call it finite if it is nowhere \( \pm \infty \), i.e., if \( f \) actually takes values in \( \mathbb{R} \).

A vector \( x^* \) in \( \mathbb{R}^m \) is called a subgradient of a convex function \( f \) at a point \( x \) if it satisfies the subgradient inequality

\[
f (z) \geq f (x) + \langle x^*, z - x \rangle \quad \text{for all } z \in \mathbb{R}^m.
\]
\[ f^* (x^*) = \arg \max_{x \in \mathbb{R}^m} \left\{ \langle x, x^* \rangle - f(x) \right\} \tag{A.4} \]

The inverse of a correspondence \( \rho : \mathbb{R}^m \rightrightarrows \mathbb{R}^m \) is the correspondence \( \rho^{-1} : \mathbb{R}^m \rightrightarrows \mathbb{R}^m \) defined by

\[ \rho^{-1}(y) := \{ x \in \mathbb{R}^m : y \in \rho(x) \} \]

Subdifferentials of proper conjugate pairs are inverse to each other in the correspondence sense, i.e., \( x \in \partial f^* (x^*) \) if and only if \( x^* \in \partial f(x) \).

We call a proper convex function \( f \) almost differentiable if it satisfies the following three conditions:

1. \( \text{int}(\text{dom} f) \) is nonempty.
2. \( f \) is differentiable throughout \( \text{int}(\text{dom} f) \).
3. \( \partial f(x) = \emptyset \) for all points \( x \in \text{dom} f \setminus \text{int}(\text{dom} f) \).

The dual notion to almost differentiability is almost strict convexity. A proper convex function \( f \) is almost strictly convex if it is strictly convex on every convex subset of \( \text{dom} \partial f \), which need not be convex. A proper convex function \( f \) is almost differentiable if and only if its conjugate \( f^* \) is almost strictly convex, and \( f^* \) is almost differentiable if and only if \( f \) is almost strictly convex (Rockafellar and Wets, 2009, Theorem 11.13).

Appendix B. Proofs

Throughout this section we let \( \omega \) denote a generic element of the sample space \( \Omega \).

Proof of Theorem 1. We first prove item 1 of the theorem, then the main statement, and finally item 2.

Item 1: The vector-max function is Lipschitz continuous with respect to \( \| \cdot \|_\infty \) with Lipschitz constant 1, which, in combination with (a conditional version of) Jensen’s inequality, establishes both existence and finiteness of the surplus function. The surplus then inherits the claimed Lipschitz property from the vector-max function by linearity of (conditional) expectations and again appealing to the conditional version of Jensen’s inequality. For fixed \( \omega \in \Omega \), the mapping

\[ v \mapsto f(v, e(\omega)) := \max_{j \in J} \{ v + e(\omega), u_j \} - \max_{j \in J} e(\omega), u_j \} \tag{B.1} \]

is an affine transformation of a pointwise maximum of affine functions and is therefore convex (Rockafellar, 1970, Theorem 5.5). Integration preserves inequalities, so convexity of the surplus follows.

To prove the main statement, first recall that a finite convex function is everywhere subdifferentiable (Rockafellar, 1970, Theorem 23.4). Now, observe that the subdifferential of the vector-max function \( f(\cdot, 0) \) is the convex hull of the directions \( u_i \) corresponding to coordinates achieving the maximum. Since \( f(v, \cdot) \) is integrable with respect to \( f_{\mathbb{R}^m} (\cdot, x) \) for any \( v \in \mathbb{R}^l \), the order of subdifferentiation and integration (expectation) may be interchanged (Bertsekas, 1973, Proposition 2.2), so

\[ \partial W(v|x) = \mathbb{E} \left[ \partial f(v; e) \right| X = x \right] = \mathbb{E} \left[ \text{conv} \left( \{ u_j \}_{j \in J, \Delta(x,w) \} \right) \right| X = x = \mathcal{P}(v|x) \right]. \]

We now turn to item 2. By 1, \( W(\cdot, x) \) is finite convex, hence proper convex. Cyclic monotonicity of \( \mathcal{P}(\cdot, x) \) now follows from the CCP correspondence coinciding with the subdifferential of the surplus, cf. Proposition A.1.

Proof of Theorem 2. We first prove item 1 of the theorem, then the main statement, and finally item 2.

Item 1: Lower-semi-continuity, properness and convexity of \( W^*(\cdot, x) \) follow from finiteness (hence properness) and convexity of \( W(\cdot, x) \) (Rockafellar, 1970, Theorem 12.2). A proper convex function is continuous on any open subset of its effective domain Rockafellar (1970, Theorem 10.1). It therefore only remains to show the claimed effective domain containments. Recall that the support function \( \sigma_c \) of a convex set \( C \) is defined by \( \sigma_c(x) := \sup_{x \in C} \langle x, x' \rangle \). To show the claimed containments, it suffices to show that \( (\text{dom} W^*(\cdot, x)) = \Delta \) (Rockafellar, 1970, Corollary 6.3.1), which, in turn, is equivalent to \( \text{dom} W^*(\cdot, x) \) and \( \Delta \) having the same support function (Rockafellar, 1970, Corollary 13.1.1).

Rockafellar (1970, Theorem 13.3) and Theorem 1 combine to show that the support function of \( W^*(\cdot, x) \) is the recession (or horizon) function of \( W(\cdot, x) \) [see Rockafellar (1970, Chapter 8] and Rockafellar and Wets (2009, Chapter 3)]. For any \( h \in \mathbb{R}^l \), it therefore follows from Rockafellar (1970, Corollary 8.5.2) and Theorem 1, that we may express the support function \( \sigma_{\text{dom} W^*(\cdot, x)}(h) \) of \( \text{dom} W^*(\cdot, x) \) as the limit

\[ \sigma_{\text{dom} W^*(\cdot, x)}(h) = \lim_{t \to \infty} \frac{W(t, h|x)}{t}, \quad h \in \mathbb{R}^l. \]

Write the surplus expression in terms of the support function of the simplex,

\[ W(v|x) = \mathbb{E} \left[ \max_{j \in J} \{ v + e, u_j \} - \max_{j \in J} e(\omega), u_j \} | X = x \right] = \mathbb{E} \left[ \sigma_{\Delta}(v + e - \sigma_{\Delta}(\omega)) | X = x \right], \]

fix \( h \in \mathbb{R}^l \), and let \( \{ t_m \}_m \subset \mathbb{R}_{++} \) be any monotone increasing sequence satisfying \( t_m \to \infty \). By convexity and continuity of the support function \( \sigma_{\Delta} \), the sequence \( \{ f_m \}_m \) of functions \( f_m : \mathbb{R}^l \to \mathbb{R} \)

\[ f_m(t) := \frac{1}{t_m} \left[ \sigma_{\Delta}(t_m h + t) - \sigma_{\Delta}(t) \right] \]

is monotone increasing (\( f_1 < f_2 < \cdots \)) with pointwise limit given by

\[ \lim_{m \to \infty} f_m(t) = \lim_{m \to \infty} \frac{1}{t_m} \sigma_{\Delta}(t_m h + t) = \sigma_{\Delta}(h), \]

which does not depend on \( t \). It now follows from a monotone convergence theorem argument that

\[ \sigma_{\text{dom} W^*(\cdot, x)}(h) = \lim_{m \to \infty} \mathbb{E} \left[ f_m(\omega) | X = x \right] = \mathbb{E} \left[ \sigma_{\Delta}(h) | X = x \right] = \sigma_{\Delta}(h). \]

Equality of support functions now follows from \( h \) being arbitrary.

To prove the main statement, first observe that \( \partial W^*(\cdot, x) = (\partial W)^{-1}(\cdot, x) \), which follows from finiteness (hence properness) of \( W(\cdot, x) \) (Theorem 1.1) and subdifferentials of conjugate pairs being inverse to each other in the correspondence sense (Rockafellar, 1970, Corollary 23.5.1). The claim that \( \partial W^*(\cdot, x) = \mathcal{P}^{-1}(\cdot, x) \) then follows from Theorem 1. The containments \( \Delta \subseteq \partial W^*(\cdot, x) \) follow from Theorem 2.1 and the fact that a proper convex function is subdifferentiable nowhere outside of its effective domain and everywhere on its (relative) interior (Rockafellar, 1970, Theorem 23.4).

Item 2: Cyclic monotonicity of \( \mathcal{P}^{-1}(\cdot, x) \) follows from it coinciding with the subdifferential of a proper convex function, cf. Proposition A.1. The claimed containments are immediate consequences of the main statement.

Proof of Theorem 3. Eq. (6) is immediate from Theorem 1, Theorem 2 and the variational characterization of subdifferentials of conjugate pairs in (A.4). That \( \Pi(x) \) is a nonempty bounded
set if and only if \( p(x) \in \text{int } A \) follows from (Rockafellar, 1970, Theorem 23.4) and \( \text{int } A = \text{int } (\text{dom } W^* (\cdot | x)) \) (Theorem 2. List 1).

**Proof of Theorem 4.** Theorem 1 yields the equivalence 1 \( \iff \) 2, and the equivalence 2 \( \iff \) 3 follows from almost differentiability and almost strict convexity being dual properties and \( W = \mathbb{R}^m \) (Theorem 1. 1), such that almost differentiability reduces to differentiability.

To show the remaining equivalence Condition \( C(x) \iff 2 \), fix \( v, h \in \mathbb{R}^l \), and suppress the “conditioning on \( X = x \)” in order to ease notation. We then want to show that \( W \) is directionally differentiable at \( v \) in the direction \( h \) if and only if (7) holds. Bertsekas (1973, Proposition 2.1) shows that \( W'(v) = \{ f' (v; e) : h \} \in \mathbb{R} \) where \( f' (v; e (\omega), h) \) denotes the one-sided directional derivative of \( f (: ; e (\omega)) \) [see (B.1)] at \( v \) in the direction \( h \). This one-sided directional derivative is two-sided if and only if the nonnegative \( W' (v) + W' (v) = \text{in fact zero}. That is, if and only if,

\[
W' (v; h) + W' (v; -h) = E [f'(v; e; h) + f'(v; e; -h)] = 0.
\]

The integrand is itself nonnegative, so the previous display holds if and only if

\[
P (f' (v; e; h) + f' (v; e; -h) > 0) = 0.
\]

A finite convex function is everywhere subdifferentiable with a compact-valued subdifferential (Rockafellar, 1970, Theorem 23.4). As established in the proof of Theorem 1, the finite convex function \( f (: ; e (\omega)) \) has subdifferential at \( v \) given by the nonempty compact set

\[
\partial f (v; e (\omega)) = \text{conv } \left\{ \left. (u_j) \in A(v + e (\omega)) \middle| u_j \right\} \right. .
\]

Its one-sided directional derivative at \( v \) in the direction \( h \) is therefore given by

\[
f' (v; e (\omega)) = \sup_{q \in \partial f (v; e (\omega))} \left. (h, q) \right| = \max_{j \in \text{int } (A(v + e (\omega)))} \left( h, u_j \right) = \max_{j \in \text{int } (A(v + e (\omega)))} \left( h, u_j \right) \quad \text{(B.2)}
\]

(Rockafellar, 1970, Theorem 23.4). The claim therefore follows from

\[
f' (v; e (\omega)) = \sup_{q \in \partial f (v; e (\omega))} \left. (h, q) \right| = \max_{j \in \text{int } (A(v + e (\omega)))} \left( h, u_j \right) \quad \text{(B.2)}
\]

The final statement follows from Rockafellar (1970, Theorem 26.1) and Theorem 1.

**Proof of Proposition 1.** We again suppress the “conditioning on \( X = x \)” throughout. A nonzero \( h \) implies that

\[
\left\{ t \in \mathbb{R}^l \mid \max_{j \in \text{int } A} (h, u_j) > \min_{j \in \text{int } A} (h, u_j) \right\} = \left\{ t \in \mathbb{R}^l \mid |A(t)| > 1 \right\},
\]

where \( |A(t)| \) denotes the cardinality of \( A(t) \). Introducing the hyperplanes

\[
H_{jk} := \left\{ t \in \mathbb{R}^l \mid \left( t, u_j - u_k \right) = 0 \right\}, \quad j, k \in J, \quad j \neq k,
\]

we may further contain the right-hand side in \( \bigcup_{j \neq k} H_{jk} \). Translating these sets by \( -v \) and invoking the union bound, we see that

\[
\text{P} \left( \max_{j \in \text{int } A(v + e)} (h, u_j) > \min_{j \in \text{int } A(v + e)} (h, u_j) \right) \leq \sum_{j \neq k} \mu (H_{jk} - v),
\]

where \( \mu \) denotes the (conditional) law of \( e \) (given \( X = x \)). Each \( H_{jk} - v \) is a hyperplane, so the right-hand side is zero by the assumption placed on \( \mu \).

**Proof of Theorem 5.** The equivalence 1 \( \iff \) 3 follows from Rockafellar (1970, Theorem 26.1) and \( \partial W^* = \mathcal{P}^{-1} \) (Theorem 2.2). The equivalence 2 \( \iff \) 3 follows from almost differentiability and almost strict convexity being dual to each other (Rockafellar and Wets, 2009, Theorem 11.13) and \( \text{dom } W = \mathbb{R}^l \) (Theorem 1. 1), such that almost strict convexity reduces to strict convexity on \( \mathbb{R}^l \).

To show the remaining equivalence Condition \( S(x) \iff 2 \), first observe that a finite convex function \( f \) is everywhere subdifferentiable with a compact-valued subdifferential (Rockafellar, 1970, Theorem 23.4). It is strictly convex if and only

\[
f(x + h) - f(x) + \langle x^*, h \rangle > 0 \quad \text{for all } x, h \in \mathbb{R}^m, x^* \neq 0, \text{ and all } x^* \in \partial f (x), \text{ i.e. if and only the graph of } f \text{ lies strictly above every tangent line (except for the point of tangency). Since each } \partial f (x) \text{ is compact, strict convexity may equivalently be expressed as}
\]

\[
\text{f(x + h) - f(x) + max}_{x^* \in \partial f (x)} \langle x^*, h \rangle = f(x) + f'(x; h) \quad \text{for all } x, h \in \mathbb{R}^m, x^* \neq 0, \text{ where we have again used Rockafellar (1970, Theorem 23.4)}.
\]

Next, fix \( v, h \in \mathbb{R}^m \), \( h \neq 0 \), and again suppress the “conditioning on \( X = x \)” of the supremum \( W \). We therefore want to show that \( W(v + h) \neq W(v) + W'(v; h) \) is equivalent to \( 8 \). Bertsekas (1973, Proposition 2.1) shows that the order of one-sided directional differentiation and integration (expectation) may be interchanged. As established in the proof of Theorem 4, the finite convex function \( f (: ; e (\omega)) \) [see (B.1)] has one-sided directional derivative at \( v \) in the direction \( h \) given by

\[
f'(v; e (\omega), h) = \max_{j \in \text{int } (A(v + e (\omega)))} (h, u_j) \quad \text{(cf. (B.2)). We may therefore express } W'(v; h) \text{ as}
\]

\[
W'(v; h) = E \left[ \max_{j \in \text{int } (A(v + e))} (h, u_j) \right].
\]

Combine the previous observations to see that

\[
W(v + h) - W(v) = E \left[ \max_{j \in J} (v + h + e, u_j) - \max_{j \in J} (v + e, u_j) \right].
\]

The integrand is itself nonnegative, so \( W(v + h) > W(v) + W'(v; h) \) if and only if

\[
P \left( \max_{j \in J} (v + h + e, u_j) > \max_{j \in J} (v + e, u_j) + \max_{j \in \text{int } A(v + e)} (h, u_j) \right) = 0.
\]

The final statement also follows from Rockafellar (1970, Theorem 26.1).

**Proof of Proposition 2.** We again suppress the “conditioning on \( X = x \)” throughout. Introducing the sets

\[
C_j := \left\{ t \in \mathbb{R}^l \mid j \in A(t) \right\}, \quad j \in J.
\]

the claim follows once we establish that

\[
\bigcup_{j \in J} \text{int } (C_j) \setminus (C_j - h) \subseteq \left\{ t \in \mathbb{R}^l \mid \max_{j \in J} (v + t, u_j) > \max_{j \in J} (v, u_j) \right\}.
\]

Indeed, suppose for the moment that the above inclusion holds. Then translating these two sets by \( -v \), we see that

\[
P \left( \max_{j \in J} (v + h + e, u_j) > \max_{j \in J} (v + e, u_j) + \max_{j \in \text{int } A(v + e)} (h, u_j) \right) \geq \mu \left( \bigcup_{j \in J} \text{int } (C_j) \setminus (C_j - h) - v \right).
\]
where $\mu$ denotes the (conditional) law of $\varepsilon$ (given $X = x$). Being the intersection of closed halfspaces, each $C_j$ is closed and, thus, each $\text{int}(C_j) \setminus \{C_j = h\}$ open. Moreover, given $h \neq 0$, the set $\text{int}(C_j) \setminus \{C_j = h\}$ must be nonempty for at least one $j$. A union of open sets is open, so the desired conclusion follows by the assumption placed on $\mu$.

It remains to show (B.3). To this end, suppose that $t$ belongs to the left-hand side of (B.3) but not the right-hand side. Then there exists $j \in J$ such that

$$\langle t, u_j \rangle = \max_{k \in J} \langle t, u_k \rangle, \quad (h + t, u_j) < \max_{k \in J} (h + t, u_k) \quad \text{and} \quad \max_{k \in J} (h + t, u_k) \leq \max_{k \in J} \langle t, u_k \rangle + \max_{A \in \mathcal{A}(t)} \langle h, u_k \rangle.$$ 

Putting these pieces together and rearranging, we must therefore have

$$\langle h, u_j \rangle < \max_{A \in \mathcal{A}(t)} \langle h, u_k \rangle.$$ 

However, $t$ is interior to $C_j$, so $\mathcal{A}(t) = \{j\}$, which is the desired contradiction. 

**Proof of Theorem 6.** The claimed equivalences follow from combining Theorems 4 and 5. The final statement then follow from Rockafellar (1970, Theorem 26.5). 

**Proof of Theorem 7.** By construction, $p(x) \in \mathcal{P}(\pi(x)) \cup \{0\}$. Under the stated assumptions, $\mathcal{P}(\cdot|x)$ reduces to a one-to-one function from $\mathbb{R}^m$ onto $\Delta$ given by $\mathcal{P}^{-1}(\cdot|x) = \{W^+(\cdot|x)\}$ on $\Delta$ (Theorem 6), so $p(x) \in \Delta \text{ and } \pi(x) = W^+(p(x))$. The last equality now follows from the variational characterization of subdifferentials of conjugate pairs in (A.4).

**Proof of Corollary 1.** The assumption of absolute continuity implies Condition $C(x)$ via Proposition 1. Condition $S(x)$ follows from the assumption of full support and Proposition 2. The claims now follow from Theorem 6. 

**Appendix C. On the domain of the conjugate surplus**

In this appendix we illustrate by example that the inclusions of Theorem 2.1 cannot be improved in general. To see that the inclusion $\text{dom} W^+(\cdot|x) \subseteq \Delta$ may be strict, let $f = 1$ (binary choice), and let $\varepsilon_1$ be distributed independently of $X$ according to a (for simplicity) continuous CDF $F$. Then $W$ does not depend on $X$ and may be calculated to be

$$W(v) = \begin{cases} v(1 - F(v)) + \int_0^v t df(t), & v \geq 0, \\ v(1 - F(v)) - \int_v^\infty t df(t), & v < 0. \end{cases}$$

It follows that

$$W^+(1) = \sup_{v \in \mathbb{R}} \{v \cdot 1 - W(v)\} \geq \sup_{v \geq 0} \left\{ vF(v) - \int_v^\infty t df(t) \right\} \geq \sup_{v \geq 0} \left\{ \int_v^\infty (-t) df(t) \right\} = \int_0^\infty (-t) df(t).$$

Similarly, $W^+(0) \geq \int_0^\infty t df(t)$. By definition, the random variable $\varepsilon_1$ is integrable if and only if $E[|\varepsilon_1|] < +\infty$, which, in turn, is equivalent to both $\int_0^\infty (-t) df(t) < +\infty$ and $\int_0^\infty t df(t) < +\infty$. Hence, if $\varepsilon_1$ is not integrable, then we must have $W^+(0) = +\infty$ and/or $W^+(1) = +\infty$ and therefore $\Delta = \{0, 1\} \not\subseteq \text{dom } W$. For example, if $\varepsilon_1 \sim \text{Logistic}(0, 1)$, a straightforward calculation shows that $\text{dom } W^* = \{0, 1\} = \text{ int } \Delta$. In contrast, for $\varepsilon_1 \sim \text{Logistic}(0, 1)$, a limiting argument shows that both interval endpoints are included in the conjugate-surplus domain.