A Distributionally Robust Random Utility Model

Emerson Melo[§] David Müller[*] Ruben Schlotter[‡]

December 18, 2023

Abstract

This paper introduces a Distributionally Robust Random Utility Model (DRO-RUM), allowing the preference shock (unobserved heterogeneity) distribution to be misspecified or unknown. Leveraging tools from the literature on robust optimization, we contribute in several ways. First, we show that the DRO-RUM inherently generates a shock distribution that incorporates a correlation between the utilities of different alternatives. Second, we establish that the distributionally robust social surplus gradient yields the choice probability vector. This result extends the well-known William-Daly-Zachary theorem to environments where the shock distribution can be misspecified. Third, we establish a robust version of the Fenchel equality. Fourth, we show how the DRO-RUM approach helps study the demand inversion problem, the random coefficients RUM, and the problem of limited consideration in discrete choice models. Finally, we illustrate how our results apply under two different notions of statistical distances: the ϕ-divergence and Sinkhorn distance approaches.

Keywords: Discrete choice, Random utility, Convex analysis, Distributionally robust optimization.

JEL classification: C35, C61, D90.

[§]Department of Economics, Indiana University, Bloomington, IN 47408, USA. Email: emelo@iu.edu.

[*]Department of Mathematics, TU Chemnitz, 09126 Chemnitz, Germany. Email: david.mueller@mathematik.tu-chemnitz.de

[‡]Hannover Re, Karl-Wiechert-Allee 50, 30625 Hannover, Germany. Email: ruben.schlotter@hannover-re.com

1 Introduction

The Random Utility Model (RUM), introduced by Marschak (1959), Block and Marschak (1959), and Becker et al. (1963), has become the standard approach for modeling stochastic choice problems. The seminal work of McFadden (1978a,c, 1981) contributes significantly to making RUM an empirically tractable method applicable in various areas of applied microeconometrics, including labor markets, industrial organization, health economics, transportation, and operations management. McFadden's work, in particular, establishes an economic foundation and econometric framework that connects observable choices to stochastic choice behavior. This distinctive feature renders RUM well-suited for addressing complex choice environments and conducting welfare analysis (McFadden (2001) and Train (2009)).

In a RUM a decision maker (DM) faces a discrete choice set of alternatives in which each option is associated with a *random* utility. Then the DM chooses a particular option with a probability equal to the event that such alternative yields the highest utility among all available alternatives. Most of the applied literature models the random utility associated with each alternative as the sum of an *observable* and *deterministic* component and a *random preference shock*. Under this additive specification, different distributional assumptions on the random preference shock will generate different RUMs. Thus, all the effort is to provide conditions on the distribution of the preference shock such that the choice probabilities are consistent with the random utility maximization hypothesis (McFadden (1981)).

More importantly, assuming that the shock distribution is known to the analyst, we can estimate the parameters describing the deterministic utility associated with each alternative, carry out counterfactual welfare analysis, and predict future choice behavior. From a modeling standpoint, this assumption means that the analyst can *correctly* specify the shock distribution that describes the *unobserved* heterogeneity in DM's behavior.

In this paper, we develop a RUM framework that allows for the possibility that the analyst (or the DM) does not know the true shock distribution. In doing so, we propose a distributional robust framework that relaxes the assumption that the shock distribution is known in advance. In particular, we develop a RUM framework that allows for misspecification in the shock distribution. By modeling the uncertainty regarding the true distribution, we follow the distributionally robust optimization literature and consider an environment where the analyst has access to a reference distribution F. This distribution corresponds to an approximation of the true statistical law generating the realizations of preference shocks. We refer to F as the nominal distribution. Accordingly, we model uncertainty distribution in terms of an *uncertainty set*, which consists of all probability distributions that are *close* to F. We rely on the concept of *statistical* divergences to measure the distance between probability distributions. More precisely, we use the notions of ϕ-divergences (Csiszar (1967); Liese and Vajda (1987)) and Sinkhorn distances (Cuturi (2013), Wang et al. (2023)). Ex-

amples of ϕ-divergences include the Kullback-Leibler, Renyi, and Cressie-Read distances, among many others. In the case of the Sinkhorn distance, we use the notion of cost function, where typical examples are the Mahalanobis and Euclidean distances. Thus, the uncertainty set contains the nominal F and all feasible distributions within a certain radius as measured by the ϕ-divergence or the Sinkhorn distance.

Based on the uncertainty set, we introduce the robust social surplus function, corresponding to the maximum social surplus achievable over all feasible distributions. Like the traditional RUM, the robust social surplus is a convex function that contains all the relevant information to study and understand our distributionally robust RUM (DRO-RUM).

1.1 Contributions

We make four contributions. First, we show that the analysis of the DRO-RUM corresponds to the study of the properties of a strictly convex finite dimensional stochastic optimization program. This characterization directly implies that the *endogenous* robust distribution associated with the DRO-RUM introduces correlation between the preference shocks, even when the nominal F may assume independence.

Second, we show that the gradient of the robust social surplus function yields the choice probability vector. The latter result is a nontrivial generalization of the celebrated Williams-Daly-Zachary (WDZ) theorem to environments where the true shock distribution is unknown. Furthermore, we show that the DRO-RUM preserves the convex structure of the traditional RUM. In particular, we derive a robust Fenchel duality framework that connects the robust social surplus and its convex conjugate.

In our third contribution, we apply our framework to address three distinct economic problems. Firstly, we characterize the empirical implications of the DRO-RUM. More precisely, we demonstrate that for an observed choice probability vector, there exists a *unique* mean utility vector that can rationalize the observed data within the context of a DRO-RUM. Notably, we establish that this mean utility vector corresponds to the gradient of the convex conjugate of the robust social surplus function. From a mathematical standpoint, we show that the mean utility vector is derived as the unique solution of a strictly convex stochastic programming problem. In our second application, we showcase how our framework can effectively tackle the robust demand inversion problem when random coefficients are present. Finally, we explore the adaptation of the DRO-RUM to incorporate limited consideration. Specifically, we delve into how our model can be integrated with the recently proposed limited consideration RUM by Aguiar et al. (2023).

In our fourth contribution, we discuss the extension of the DRO-RUM beyond the realm of ϕ-divergences. Specifically, we introduce the Sinkhorn DRO-RUM distance, a concept recently introduced in the distributionally robust optimization literature by Wang et al. (2023). The Sinkhorn distance formalizes the

discrepancy between probability measures as the solution to a regularized optimal transport problem. In this context, we demonstrate that Sinkhorn DRO-RUM also preserves the convex analytic properties of the RUM. Notably, the Sinkhorn DRO-RUM can be approached through a finite-dimensional optimization program, adding to its tractability. Leveraging this feature, we establish the WDZ theorem, the concept of Robust Fenchel equality, and address the demand inversion problem in this more general setting.

We close the paper by conducting a series of numerical simulations to elucidate the properties of our framework. Specifically, we compare the choice behavior of the DRO-RUM with the multinomial logit (MNL) and multinomial probit (MNP) models. Our primary focus centers on analyzing the influence of the *robustness parameter*, which governs the size of the feasible set, thus affecting both choice probabilities and the surplus function.

1.2 Related literature

Our paper is related to several strands of literature. First, our paper relates to the literature on RUMs and convex analysis. The closest articles to ours are the works by Chiong et al. (2016), Galichon and Salanié (2021), and Fosgerau et al. (2021). Similar to us, these papers exploit the convex structure of the RUM to study the nonparametric identification of the mean utility vector when aggregate market data is available (observed choice probabilities). Our paper and results differ substantially from their work by allowing a more flexible framework regarding distributional assumptions.

Second, our paper relates to the semiparametric choice model (SCM) literature. The work by Natarajan et al. (2009) introduces the SCM in an environment where the true *joint* distribution is unknown, but the analyst has access to the set of marginal distributions associated with each alternative. This particular instance of the SCM is known as the marginal distribution model (MDM). Mishra et al. (2014) studies the MDM approach's theoretical and empirical performance. Mishra et al. (2012) study a second instance of the SCM, which exploits cross moments constraints. In particular, they assume that the true distribution is unknown but the analyst has access to the true variance-covariance matrix that captures the correlation structure across the set of discrete alternatives.

At first glance, our approach is similar to SCM. As discussed by Feng et al. (2017), the latter are generally defined by a supremum over a set of distributions. We adapt to this definition by introducing the DRO-RUM, where the true distribution is unknown, and the analyst, therefore, considers all distributions in an uncertainty set. Despite the similarity between the general SCM and our approach, both frameworks have important differences. First, our approach requires no assumption on the marginal distributions or variance-covariance matrix. Instead, our model only requires knowledge of a nominal distribution. Second, using the concept of ϕ-divergence(or Sinkhorn distance) enables the researcher to incorporate robustness, where she can control the uncertainty

concerning the shock distribution by selecting the *robustness parameter*. Hence, the feasible set is not determined explicitly by fixing some moments or marginal distributions but is rather implicitly constructed by choosing the nominal distribution and the magnitude of the *robustness parameter*. Moreover, our approach can generate different models by allowing the choice of several ϕ-divergence functions and different nominal distributions. Third, we show that the DRO-RUM preserves the convex structure (and duality) of the traditional RUM approach. In particular, we generalize the WDZ and provide a robust Fenchel duality analysis. Fourth, we apply our approach to study the demand inversion problem, the random coefficient RUM, and the problem of limited consideration in discrete choice models.

Our paper also connects to the literature on robustness in macroeconomics (Hansen and Sargent (2001, 2008)). However, this literature predominantly addresses recursive problems utilizing the Kullback-Leibler distance. A recent contribution by Christensen and Connault (2023) introduces robustness techniques to analyze the sensitivity of counterfactuals to parametric assumptions about the distribution of latent variables in structural models. While related, their focus differs from the problem addressed in our paper. Notably, they do not explore the convex analytic properties of the DRO-RUM model. Additionally, Christensen and Connault (2023) do not delve into the robustness problem concerning the Sinkhorn distance.

Similarly, our work is also related to the decision theory literature on ambiguity and model uncertainty. The seminal papers by Gilboa and Schmeidler (1989) and Maccheroni et al. (2006) provide axiomatic foundations to represent DM's preferences in environments where she faces model ambiguity. Strzalecki (2011) studies from an axiomatic standpoint the connection between multiplier preferences and robustness.[1] However, none of these papers study the problem of model uncertainty in the context of RUMs.

Finally, our paper is closely related to the literature on distributionally robust optimization. Shapiro (2017) and Kuhn et al. (2019) provide an up-to-date treatment of the subject. Applications vary from inventory management to regularization in machine learning.[2]However, to our knowledge, this literature has not studied the problem of the distributional robustness of the RUM.

The rest of the paper is organized as follows. Section 2 reviews the traditional RUM approach and introduces the problem of robustness. Section 3 presents the DRO-RUM model and discusses its main properties. Section 4 discusses some applications. Section 5 contents several numerical experiments comparing the outcome of the DRO-RUM with respect to MNL and MNP. Section 6 discuss the DRO-RUM in the context of Sinkhorn distances. Finally, section 7 concludes the paper by providing an overview of possible extensions.

[1]For an excellent survey of the literature on ambiguity and model uncertainty, we refer the reader to Hansen (2014) and Marinacci (2015).

[2]In Economics, one of the first papers studying robust optimization problems is Scarf (1958)

Notation. Throughout the paper we use the following notation and definitions. Let us denote $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and consider extended real-valued functions

$$f : \mathbb{V} \to \bar{\mathbb{R}},$$

where $\mathbb{V}$ is a finite dimensional real vector space. Consequently, we denote by $\mathbb{V}^*$ its dual space consisting of all linear functionals. In particular, we often work with subspaces of $\mathbb{R}^n$. The set defined by

$$\mathrm{dom} f = \{x \in \mathbb{V} : f(x) < +\infty\}$$

is called the *(effective) domain* of f. A function is said to be *proper* if it takes nowhere the value $-\infty$ and $\mathrm{dom} f \neq \emptyset$. For a proper function $f : \mathbb{V} \to \bar{\mathbb{R}}$ the set $\partial f(x)$ represents its *subdifferential* at $x \in \mathrm{dom} f$, i.e.

$$\partial f(x) = \{g \in \mathbb{V}^* : f(y) \geq f(x) + \langle g, y - x \rangle, \quad \text{for all} \quad y \in \mathbb{R}^n\},$$

where $g \in \mathbb{V}^*$ is said to be a *subgradient*. If the subdifferential set is a singleton, i. e. the subgradient g is unique, we denote by $\nabla f(x)$ the *gradient* of the function f at $x \in \mathrm{int}\,(\mathrm{dom} f)$. The *convex conjugate* of a proper function $f : \mathbb{V} \to \bar{\mathbb{R}}$ is

$$f^*(g) = \sup_{x \in \mathbb{V}} \{\langle x, g \rangle - f(x)\}, \quad g \in \mathbb{V}^*.$$

$\mathbb{E}_F(\cdot)$ denotes the expectation operator with respect to a distribution F.

2 The Random Utility Model

Consider a decision maker (DM) making a utility-maximizing discrete choice among alternatives $j \in \mathcal{J} = \{0, 1, \ldots, J\}$. The utility of option j is

$$\tilde{u}_j = u_j + \varepsilon_j, \tag{1}$$

where $u = (u_0, u_1, \ldots, u_J)^T$ is deterministic and $\varepsilon = (\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_J)^T$ is a vector of random utility shocks. The alternative 0 has the interpretation of an outside option. We assume that the set of potential deterministic utility vectors is given by the set $\mathcal{U} \triangleq \{u \in \mathbb{R}^{J+1} : u_0 = 0\}$. In other words, $\mathcal{U}$ is the set of mean utility vectors with the normalization $u_0 = 0$ for the outside option.

Following McFadden (1978a, 1981), the previous description corresponds to the classic additive random utility model (RUM). Our presentation of the RUM framework here will emphasize convex-analytic properties.

Assumption 1 *The random vector ε follows a distribution F that is absolutely continuous with finite means, independent of u, and fully supported on $\mathbb{R}^{J+1}$.*

Assumption 1 leaves the distribution of ε unspecified, thus allowing for a wide range of choice probability systems far beyond the often-used logit model. The assumption allows arbitrary correlation between the ε_j's may be important

in applications. As a direct consequence of Assumption 1, the DM's choice probabilities correspond to:

$$p_j(u) \equiv \mathbb{P}\left(u_j + \varepsilon_j = \max_{j' \in \mathcal{J}} \{u_{j'} + \varepsilon_{j'}\}\right), \quad j = 0, 1, \ldots, J.$$

A fundamental object in the RUM framework is the *surplus function* of the discrete choice model (so named by McFadden (1981)). It is given by the *expected indirect utility* defined as:

$$W(u) = \mathbb{E}_F\left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\}\right]. \tag{2}$$

Under Assumption 1, W is convex and differentiable and the choice probability vector $p(u)$ coincides with the gradient of W[3]

$$\frac{\partial}{\partial u_k} W(u) = p_k(u) \text{ for } k = 0, 1, \ldots, J$$

or, using vector notation, $p(u) = \nabla W(u)$. The previous result is the celebrated Williams-Daly-Zachary (henceforth, WDZ) theorem, famous in the discrete choice literature (McFadden (1978a, 1981)).

2.1 Examples of RUM

One of the most widely used RUMs is the multinomial logit (MNL) model, which assumes that the entries of $(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_J)^T$ follow i.i.d. Gumbel distributions with scale parameter η. Given this assumption, we can write the social surplus function in closed form:

$$W(u) = \eta \log\left(\sum_{j=0}^{J} e^{u_j/\eta}\right) + \eta\gamma, \tag{3}$$

where γ is the Euler-Mascheroni constant. It follows from (3) that the WDZ theorem implies that $p_j(u)$ is given by:

$$\frac{\partial W(u)}{\partial u_j} = \frac{e^{u_j/\eta}}{\sum_{l=0}^{J} e^{u_l/\eta}} \quad \text{for } j \in \mathcal{J}. \tag{4}$$

The MNL model belongs to a broader class of RUM models called generalized extreme value (GEV) models introduced by McFadden (1978b). This class of models is defined via a generating function $G : \mathbb{R}_+^{J+1} \to \mathbb{R}_+$, which has to satisfy the following properties:

(G1) G is homogeneous of degree $\frac{1}{\eta} > 0$.

(G2) $G(x_0, x_1, \ldots, x_j, \ldots, x_J) \to \infty$ as $x_j \to \infty$, $j = 0, 1, \ldots, J$.

[3]The convexity of W follows from the convexity of the max function. Differentiability follows from the absolute continuity of ε. See Shi et al. (2018), Chiong et al. (2016), and Melo et al. (2019) for semiparametric econometric approaches based on these convex-analytic properties of discrete-choice models.

(G3) For the partial derivatives of G w.r.t. k distinct variables it holds:

$$\frac{\partial^{k+1} G\left(x_0,\ldots,x_J\right)}{\partial x_{j_0},\partial x_{j_1}\cdots\partial x_{j_k}} \geq 0 \text{ if } k+1 \text{ is odd}, \quad \frac{\partial^{k+1} G\left(x_0,\ldots,x_J\right)}{\partial x_{j_0},\partial x_{j_1}\cdots\partial x_{j_k}} \leq 0 \text{ if } k+1 \text{ is even}.$$

McFadden (1978b, 1981) show that a function G satisfying conditions (G1)-(G3) implies that the joint distribution of the random vector ε corresponds to the following probability density function:

$$f_\epsilon\left(y_0,y_1,\ldots,y_J\right) = \frac{\partial^{J+1}\exp\left(-G\left(e^{-y_0},\ldots,e^{-y_J}\right)\right)}{\partial y_0\cdots\partial y_J},$$

An essential property of the GEV class is that the social surplus function corresponds to (McFadden, 1978b)

$$W(u) = \eta \ln G\left(e^u\right) + \eta\gamma,$$

where γ is the Euler-Mascheroni constant. From the WDZ theorem it follows that the choice probability of the j-th alternative corresponds to:

$$p_j(u) = \frac{\partial W(u)}{\partial u_j} = \eta\frac{\partial G\left(e^u\right)}{\partial e^{u_j}}\cdot\frac{e^{u_j}}{G\left(e^u\right)} \quad \forall j \in \mathcal{J}.$$

It is easy to see that the generating function

$$G(e^u) = \sum_{j=0}^{J} e^{u_j/\eta} = 1 + \sum_{j=1}^{J} e^{u_j/\eta}$$

leads to the MNL model.

The main advantage of the GEV class is its flexibility to capture complex patterns correlation across the random variables ε_j's. Examples of this are the Nested Logit (NL), the Paired Combinatorial Logit (PCL), the Ordered GEV (OGEV), and the Generalized Nested Logit (GNL) model, which are particular instances of the GEV family.

2.2 A robust framework for the RUM

A fundamental assumption in the RUM is that the shock distribution is known to the researcher (and the DM). This means that the distribution of ε is correctly specified. Our main goal in this paper is to relax this condition by allowing the distribution of ε to be unknown. Instead, the distribution of ε is an argument in an optimization problem that corresponds to the definition of the social surplus function. We formalize this idea by replacing expression (2) with the *robust social surplus* function:

$$\underline{W}^{RO}(u) = \sup_{G\in\mathcal{M}(F)} \mathbb{E}_G\left[\max_{j\in\mathcal{J}}\left\{u_j+\varepsilon_j\right\}\right], \tag{5}$$

where $\mathcal{M}(F)$ is a set of probability distributions that are close to a predetermined distribution F which satisfies Assumption 1.

The Definition (5) warrants some important remarks. Firstly, it is crucial to recognize that expression (5) signifies a scenario in which the decision-maker (or the analyst) faces uncertainty regarding the true distribution that generates the vector ε. Consequently, (5) incorporates the selection of a distribution G as an integral part of the concept of robust social surplus.

Secondly, it's worth noting that, for a fixed distribution F, $W(u)$ quantifies the indirect expected utility. Therefore, Definition (5) characterizes an environment in which the decision-maker makes choices based on the most favorable error distribution. In essence, Definition (5) represents an optimistic DRO problem.

However, from an economic standpoint, the notion of DRO encodes the idea of being robust to worst-case scenarios (Hansen and Sargent (2001, 2008), Maccheroni et al. (2006), Gilboa and Schmeidler (1989)). In order to incorporate this idea, we modify the expression (5) as follows:

$$\underline{W}^{RO}(u) = \inf_{G \in \mathcal{M}(F)} \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right], \tag{6}$$

The expression (6) defines the robust social surplus by considering a distribution G that minimizes the value of $\mathbb{E}_G(\max_{j\in\mathcal{J}}\{u_j + \varepsilon_j\})$. In particular, the expression (6) formalizes a situation where the DM (or the analyst) adopts a pessimistic attitude due to the lack of knowledge about the true shock distribution. In other words, it represents a conservative approach that takes into account the worst-case scenario regarding the distribution of shocks. Throughout the paper, we will focus on both cases.

It is worth pointing ou that, from an economic standpoint, there is an important interpretation of the definitions (5) and (6). The *robust*-RUM considers a situation where a DM faces preference shocks but has some flexibility concerning the distribution generating those errors. According to this interpretation, the set $\mathcal{M}$ determines how much flexibility the DM has while making a decision. Similarly, the set $\mathcal{M}$ also influences how optimism and pessimism are captured by (5) and (6). In later sections, we present some economic implications of this point of view. More precisely, we will incorporate the robust surplus approach into the costly attention allocation model by Aguiar et al. (2023) and analyze the implications of the our approach.

In the uncertainty set $\mathcal{M}(F)$, the distribution F can be seen as a *best guess* of the joint distribution of error terms. In order to be robust against misspecification the DM (or the analyst) takes into account all possible distributions that are close to the nominal distribution F. In fact, a key aspect of our approach is related to the structure of the set $\mathcal{M}(F)$. In Section 3 we specify the $\mathcal{M}(F)$ in terms of ϕ-divergence functions, which enables us to use the notion of statistical divergences between probability distributions (Csiszar (1967), Liese and Vajda (1987), and Pardo (2005)). Hence, we will refer to this as the *distributionally robust*-RUM (DRO-RUM). As we shall see, by doing this we are able to characterize the resulting DRO-RUM surplus function in terms of a convex finite

dimensional optimization program. This characterization is key in studying the properties of the DRO-RUM approach.

Finally, we point out that the decisions of the robust surplus approach inherit the randomness of RUM. Let $G^\star$ denote the distribution (or a limit of a sequence of distributions) that attains the optimal value in (5) (or (6)). The choice probability for alternative j under this model is given by (provided that it is well defined):

$$p_j^{RO}(u) = \mathbb{P}_{G^\star}\left(j = \arg\max_{j' \in \mathcal{J}}\{u_{j'} + \varepsilon_{j'}\}\right) \tag{7}$$

2.3 Connection with the semiparametric choice model approach

It is worth pointing out that the definition of the RO-RUM is similar to the *semiparametric choice model* (SCM), which has been recently introduced in the operation research literature (Natarajan et al. (2009)). Formally, in the SCM the surplus function is defined as the supremum over distributions. In particular, the definition (5) is a particular instance of the SCM.

By doing so, the SCM can capture complex substitution patterns and correlation between the different alternatives in the choice set $\mathcal{J}$. Feng et al. (2017) provide a detailed overview of several discrete choice models, where the authors refer to SCM as a supremum over a general set of distributions. Thus, the *robust-RUM* could be seen as an instance of a semi-parametric choice model. There are some existing instances of SCM in the literature. In their original paper, Natarajan et al. (2009) restrict the feasible set to joint distributions with given information on the marginal distributions. This particular instance of the SCM is known as the *marginal distribution model* (MDM).[4] A second class of SCMs exploits cross-moment constraints. In particular, Mishra et al. (2012) study the *cross-moment model* (CMM), which considers the set $\mathcal{M}$ to be the set of distributions consistent with a *known* variance-covariance matrix.[5]

Despite the apparent similarity, our approach differs from the existing SCM in several key aspects. First, as we will demonstrate, our framework diverges from SCM in how we specify the set of distributions. In traditional SCMs, analysts are typically required to explicitly construct a feasible set, often by fixing the marginal distributions or the variance-covariance matrix. In contrast, in our robust approach, the analyst defines the feasible set implicitly by specifying the nominal distribution F and then upper bounding the distance of other distributions from F. As a result, the DRO-RUM approach does not require knowledge of the marginals or variance-covariance matrix. In Section 3, we will

[4] In MDM, the marginal distributions of the random vector ε are fixed. Formally, we write $\varepsilon_i \sim F_i$, where F_i is the marginal distribution function of the i-th error, $i = 1, \ldots, J$. In this case, we define $\mathcal{M} \triangleq \textsc{Mar} = \{F : \varepsilon_i \sim F_i \quad \forall i \in \mathcal{J}\}$.

[5] Formally, the CMM considers the set of distributions $\mathcal{M}(0, \Sigma) = \{G : \mathbb{E}_F(\varepsilon) = 0, \quad \mathbb{E}_G(\varepsilon\varepsilon^\top) = \Sigma\}$. In the definition of $\mathcal{M}(0, \Sigma)$, the variance-covariance matrix Σ is assumed to be known.

explore how in the DRO-RUM, the researcher has control over the distance parameter by selecting a robustness parameter's magnitude. Thus, our approach fundamentally operates on a different principle compared to existing SCMs.

Secondly, we delve into the analysis of both the optimistic and pessimistic versions of the robust social surplus. In contrast, the SCM approach exclusively considers the optimistic case. This mathematical distinction holds economic significance, as modeling robustness from a solely optimistic perspective may prove inadequate in certain decision-making scenarios. For example, in Section 4.3, we demonstrate that within the framework of a Random Utility Model (RUM) with limited consideration, an optimistic RUM (a specific instance of the SCM) implies that a decision-maker will prefer choice sets with greater uncertainty. In this specific case, we uncover an intriguing observation: the robust surplus shows a positive correlation with the variance of preference shocks among different alternatives within a particular choice set. In essence, in this case, we can characterize the decision-maker as a risk-loving agent. However, this seemingly counterintuitive behavior can be mitigated when we consider the pessimistic version of the social surplus function. Therefore, the DRO-RUM framework offers the flexibility to capture a range of behavioral patterns associated with non-standard decision-making scenarios.

Additionally, we show that the DRO-RUM corresponds to the solutions of finite dimensional optimization problem. This latter fact allows us to extend the WDZ theorem to environments where the shock distribution is misspecified. Finally, Section 4 shows how the DRO-RUM enables us to recover the mean utility vector u.

3 A Distributionally Robust - RUM model

In this section, we formally introduce the DRO-RUM approach. Following the distributionally robust optimization literature, we consider an environment where the researcher (or the DM) has access to a reference distribution F, which may be an approximation (or estimate) of the true statistical law governing the realizations of ε. We refer to F as the *nominal* distribution. Then, we define a set of probability distributions that are *close* to F. We rely on statistical distances to formalize the notion of distance between probability distributions.

3.1 ϕ-divergences

We measure the distance between two probability distributions by the so-called ϕ-divergence.
Let $\phi : \mathbb{R} \to (-\infty, +\infty]$ be a proper closed convex function such that dom ϕ is an interval with endpoints $\alpha < \beta$, so, int (dom ϕ) $= (\alpha, \beta)$. Since ϕ is closed, we have $\lim_{t\to\alpha_+} \phi(t) = \phi(\alpha)$, if α is finite and $\lim_{t\to\beta_-} \phi(t) = \phi(\beta)$, if β is finite.

Throughout the paper we assume that ϕ is nonnegative and attains its minimum at the point $1 \in$ int (dom ϕ), i.e. $\phi(1) = 0$. The class of such functions is denoted by Φ.

Definition 1 *Given $\phi \in \Phi$, the ϕ-divergence of the probability measure G with respect to F is*

$$D_\phi(G\|F) = \begin{cases} \int_{\mathbb{R}^{J+1}} \phi\left(\frac{g(\varepsilon)}{f(\varepsilon)}\right) f(\varepsilon) d\varepsilon & \text{if } \mathrm{G} \ll \mathrm{F} \\ +\infty & \text{otherwise} \end{cases} \tag{8}$$

where f and g are the associated densities of F and G respectively.

To avoid pathological cases, throughout the paper, we assume the following:

$$\phi(0) < \infty, 0 \cdot \phi\left(\frac{0}{0}\right) \equiv 0, 0 \cdot \phi\left(\frac{s}{0}\right) = \lim_{\varepsilon \to 0} \varepsilon \cdot \phi\left(\frac{s}{\varepsilon}\right) = s \lim_{t \to \infty} \frac{\phi(t)}{t}, \quad s > 0. \tag{9}$$

If the measure G is absolutely continuous w.r.t. F, i. e. $G \ll F$, the ϕ-divergence can be conveniently written as:

$$D_\phi(G\|F) = \mathbb{E}_F\left(\phi(L(\varepsilon))\right), \tag{10}$$

where $L(\varepsilon) \triangleq g(\varepsilon)/f(\varepsilon)$ is the likelihood ratio between the densities g and f, also known as Radon-Nikodym derivative of the two measures. Using the expression (10) combined with the convexity of ϕ, Jensen's inequality implies that

$$D_\phi(G\|F) \geq \phi\left(\mathbb{E}_F\left(L(\varepsilon)\right)\right) = \phi(1) = 0 \tag{11}$$

with equality if $G = F$, so that $D_\phi(G\|F)$ is a measure of distance of G from F.[6] Furthermore, the ϕ-divergence functional is convex in both of its arguments. The following proposition summarizes these key properties.

Proposition 1 *The ϕ-divergence functional (8) is well-defined and nonnegative. It is equal to zero if and only if $f(t) = g(t)$ a.e. Furthermore, D_ϕ is convex on each of its arguments.*

In our analysis, a key element will be the convex conjugate of ϕ. For $\phi \in \Phi$ its conjugate denoted by ϕ^* is:

$$\phi^*(s) = \sup_{t \in \mathbb{R}}\{st - \phi(t)\} = \sup_{t \in \operatorname{dom}\phi}\{st - \phi(t)\} = \sup_{t \in \operatorname{int}\operatorname{dom}\phi}\{st - \phi(t)\}, \tag{12}$$

where the last equality follows from Corollary 12.2.2 in Rockafellar (1970). The conjugate ϕ^* is a closed proper convex function, with int dom $\phi^* = (a, b)$, where

$$a = \lim_{t \to -\infty} t^{-1}\phi(t) \in [-\infty, +\infty); b = \lim_{t \to +\infty} t^{-1}\phi(t) \in (-\infty, +\infty].$$

Moreover, since ϕ is convex and closed, we have for its bi-conjugate $\phi^{**} = \phi$, (Rockafellar (1970)). It is worth noting that using the fact that 1 is the minimizer of ϕ and it is in the interior of its domain, so $\phi'(1) = 0$ holds. In addition, using the property that ϕ is convex and closed, we have by Fenchel equality $y = \phi'(x)$ iff $x = \phi^{*'}(y)$. Applying this latter observation to $x = 1$ and $y = 0$ we obtain $\phi^{*\prime}(0) = 1$.

[6] We recall that $1 \in$ int dom ϕ is the point where ϕ attains its minimum 0.

3.2 The DRO-RUM framework

The main idea is to consider an environment where the analyst (or a DM) does not know the true distribution governing realizations of the shock vector ε. In this environment, the role of F is an approximation or some best guess of the "true" unknown distribution. Recognizing this ambiguity or potential misspecification of the distribution F, we make use of the ϕ-divergence to define the *uncertainty set* $\mathcal{M}_\phi(F)$ as:

$$\mathcal{M}_\phi(F) = \{G \ll F : D_\phi(G||F) \leq \rho\}, \tag{13}$$

Formally, $\mathcal{M}_\phi(F)$ is the set of all probability measures G that are absolutely continuous w.r.t F, whose distance from F, as measured by the ϕ-divergence, is at most ρ. The hyperparameter ρ is the radius of $\mathcal{M}_\phi(F)$, which reflects how uncertain is the researcher (or the DM) about the plausibility of F being correct. Let us further elaborate on this interpretation. Following Hansen and Sargent (2001, 2008), **?**, and **?**, we interpret the set (13) as an environment in which the analyst (or the DM) has some best guess F of the true *unknown* probability distribution, but does not fully trust it. For instance, the researcher may consider that the nominal distribution F corresponds to the Gumbel distribution. In this case, $\mathcal{M}_\phi(F)$ accounts for many other probability distributions G to be feasible, where ρ determines the size of the feasible set.

Endowed with the set $\mathcal{M}_\phi(F)$, we can modify expressions (5) and (6) to obtain a *distributionally robust* surplus function. Thus, the surplus function of the DRO-RUM corresponds to the following optimization problems:

$$\overline{W}(u) = \sup_{G \in \mathcal{M}_\phi(F)} \left\{ \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right] \right\} \tag{14}$$

and

$$\underline{W}(u) = \inf_{G \in \mathcal{M}_\phi(F)} \left\{ \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right] \right\} \tag{15}$$

Some remarks are in order. First, a fundamental aspect of programs (14) (15) is the role of the parameter ρ which controls the size of $\mathcal{M}_\phi(F)$. Because of this, we can interpret ρ as an *index of robustness*. More precisely, when $\rho = 0$ we get $\mathcal{M}_\phi(F) = \{F\}$, which means that we recover the RUM under the distribution F.[7] On the other hand, when $\rho \longrightarrow \infty$ the uncertainty set $\mathcal{M}_\phi(F)$ admits a much larger set of possible distributions, including those that may not satisfy Assumption 1.[8] The DRO-RUM aims to set ρ to reflect the perceived

[7] We note that $\rho = 0$ implies that $D_\phi(G\|F) = 0$. Then by Proposition 1, we know that this latter equality holds if and only if $F = G$.

[8] To see this, we note that when $\rho \longrightarrow \infty$ the ϕ-divergence is unbounded. This latter fact implies that the set $\mathcal{M}_\phi(F)$ consists of all distributions which are absolute continuous w.r.t. to F. As F is fully supported, this only implies that the distributions in $\mathcal{M}_\phi(F)$ must be continuous but certainly not fully supported on $\mathbb{R}^{J+1}$. In fact, $\mathcal{M}_\phi(F)$ may consist of distributions that are absolutely continuous w.r.t Lebesgue measure but without finite means. For instance, the Pareto distribution with shape parameter $\alpha = 1$ is absolutely continuous but fails to have a finite mean.

uncertainty that the researcher (or a DM) experiences about the distributional assumption for ε.

The following lemma establishes some elementary properties of $\overline{W}(u)$.[9]

Lemma 1 *For the DRO-RUM the surplus function $\overline{W}(u)$ satisfies:*

(i) $\overline{W}(u + c \cdot e) = \overline{W}(u) + c$ *for all* $c \in \mathbb{R}, u \in \mathbb{R}^{J+1}$.

(ii) $\overline{W}(u) \geq \overline{W}(v)$ *for all* $u, v \in \mathbb{R}^{J+1}$ *with* $u \geq v$.

(iii) $\overline{W}(u) \geq \max_{j \in \mathcal{J}} u_j + \min_{j \in \mathcal{J}} \mathbb{E}_F [\varepsilon_j]$.

Similarly, $\underline{W}(u)$ satisfies properties (i), (ii), and the property:

(iv) $\underline{W}(u) \leq \max_{j \in \mathcal{J}} u_j + \min_{j \in \mathcal{J}} \mathbb{E}_F [\varepsilon_j]$.

The following result characterizes $\overline{W}(u)$ and $\underline{W}(u)$.

Proposition 2 *Let Assumption 1 hold and define the random variable $H(u, \varepsilon) \triangleq \max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\}$. Then the following statements hold:*

(i) The problem (14) *is equivalent to solving the following finite-dimensional convex program:*

$$\overline{W}(u) = \inf_{\lambda \geq 0, \mu \in \mathbb{R}} \left\{ \lambda\rho + \mu + \lambda \mathbb{E}_F \left[\phi^* \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \right] \right\}, \tag{16}$$

where λ is the Lagrange multiplier associated to the uncertainty set $\mathcal{M}_\phi(F)$ and μ the multiplier associated to G being a probability measure. Furthermore, the program (16) is convex in μ and λ.

(ii) The problem (15) *is equivalent to solving the following finite-dimensional convex program:*

$$\underline{W}(u) = \sup_{\lambda \geq 0, \mu \in \mathbb{R}} \left\{ -\lambda\rho - \mu - \lambda \mathbb{E}_F \left[\phi^* \left(\frac{-H(u, \varepsilon) - \mu}{\lambda} \right) \right] \right\}, \tag{17}$$

where λ is the Lagrange multiplier associated to the uncertainty set $\mathcal{M}_\phi(F)$ and μ the multiplier associated to G being a probability measure. Furthermore, the program (17) is concave in μ and λ.

Some remarks are in order. First, it is worth pointing out that the result in Proposition 2 follows from standard arguments in the literature on distributionally robust optimization problems (e.g. Ben-Tal et al. (2013); Bayraksan and Love (2015); **?**). However, the contribution of Proposition 2 is connect the RUM with the literature on DRO problems. Second, the result in Proposition 2 characterizes the functions $\overline{W}(u)$ and $\underline{W}(u)$ as the solution of a *finite-dimensional*

[9] We must mention that the result in Lemma 1 resembles the properties of the welfare-choice based models in Feng et al. (2017). In fact, they show that the the social surplus function associated with a RUM corresponds to a welfare-choice function. We use the result in Lemma 1 to understand properties of the associated robust optimization with $\overline{W}(u)$.

stochastic convex optimization problem. The efficiency in solving programs (16) and (17) strongly depends on expectation w.r.t. the nominal distribution F and the properties of the convex conjugate ϕ^* (see, for instance, Ruszczyński and Shapiro (2021)).

The next corollary is a straightforward consequence of Proposition 2 formalizes the connection between $W(u)$ and $W^{DRO}(u)$ when $\rho = 0$.

Corollary 1 *Let Assumption 1 hold. Then for $\rho = 0$ we get $\overline{W}(u) = \underline{W}(u) = W(u)$.*

3.3 A robust WDZ theorem

A fundamental aspect of RUMs is the possibility of characterizing choice probabilities under specific distributional assumptions on ε. Formally, and as a consequence of Assumption 1, the WDZ theorem establishes that the gradient of $W(u)$ yields the choice probability vector $p(u)$. In this section, we show that in the DRO-RUM, a similar result holds. In particular, we show that $\nabla\overline{W}(u) = \overline{p}^\star(u)$ where $\overline{p}^\star(u)$ corresponds to the choice probability vector generated by the optimal solution to (16) approach. Similarly, for the pessimistic case, we show that $\nabla\underline{W}(u) = \underline{p}^\star(u)$ where $\underline{p}^\star(u)$ is the choice probability vector associated with (17).

To establish this result, we need the following assumption.

Assumption 2 *$\phi^*(s)$ is strictly convex and differentiable with $\phi^{*\prime}(s) \geq 0$ for all s.*

We point out that many ϕ-divergence functions satisfy Assumption 2. Table 1 overviews three popular ϕ-divergences satisfying this assumption.

Divergence	$\phi(t)$	$\phi^*(s)$	Domain	$\phi^{*'}$	$\phi^{*''}$
Kullback-Leibler	$t\log t$	e^{s-1}	$\mathbb{R}$	e^{s-1}	e^{s-1}
Reverse Kullback-Leibler	$-\log(t)$	$-1-\log(-s)$	$\mathbb{R}_{--}$	$-\frac{1}{s}$	$\frac{1}{s^2}$
Hellinger Distance	$(\sqrt{t}-1)^2$	$\frac{s}{1-s}$	$s<1$	$\frac{1}{(1-s)^2}$	$-\frac{2}{(s-1)^3}$

Table 1: ϕ-divergences with their convex conjugates and first and second derivatives.

As a direct implication of the Assumption 2 we can establish the strict convexity and uniqueness of an optimal solution to (16).

Lemma 2 *Let Assumptions 1 and 2 hold. Then program* (16) *(respectively* (17)*) is strictly convex (concave) and has a unique optimal solution $\overline{\lambda}^\star$ and $\overline{\mu}^\star$.*

A second important implication of Assumption 2 is the possibility of characterizing the robust density associated to the optimal solution of the program (16).

Lemma 3 *Let Assumptions 1 and 2 hold. For a fixed $u \in \mathcal{U}$, let $\overline{\lambda}^\star > 0$ and $\overline{\mu}^\star \in \mathbb{R}$ be the unique solution to problem (16). Then the unique robust density $\overline{g}^\star(\varepsilon)$ corresponds to:*

$$\overline{g}^\star(\varepsilon) = \phi^{*\prime}\left(\frac{H(u,\varepsilon) - \overline{\mu}^\star}{\overline{\lambda}^\star}\right) f(\varepsilon) \quad \forall \varepsilon \in \mathbb{R}^{J+1}. \tag{18}$$

Similarly, let $\underline{\lambda}^\star > 0$ and $\underline{\mu}^\star \in \mathbb{R}$ be the unique solution to problem (17). Then the unique robust (pessimistic) density $\underline{g}^\star(\varepsilon)$ corresponds to:

$$\underline{g}^\star(\varepsilon) = \phi^{*\prime}\left(\frac{-H(u,\varepsilon) - \underline{\mu}^\star}{\underline{\lambda}^\star}\right) f(\varepsilon) \quad \forall \varepsilon \in \mathbb{R}^{J+1}. \tag{19}$$

Proof. Define $\Psi(\lambda, \mu) := \lambda\rho + \mu + \lambda \mathbb{E}_F\left(\phi^*\left(\frac{H(u,\varepsilon)-\mu}{\lambda}\right)\right)$. Optimizing $\Psi(\lambda,\mu)$ w.r.t λ and μ, the first order conditions combined with Assumption 2 yield that the optimal solution $\overline{\lambda}^\star$ and $\overline{\mu}^\star$ must satisfy

$$\begin{aligned} \mathbb{E}_F\left(\phi^{*\prime}\left(\frac{H(u,\varepsilon) - \overline{\mu}^\star}{\overline{\lambda}^\star}\right)\right) &= 1 \\ \int_{\mathbb{R}^{J+1}} \phi^{*\prime}\left(\frac{H(u,\varepsilon) - \overline{\mu}^\star}{\overline{\lambda}^\star}\right) f(\varepsilon) d\varepsilon &= 1 \end{aligned}$$

Define $\overline{g}^\star(\varepsilon) \triangleq \phi^{*\prime}\left(\frac{H(u,\varepsilon)-\overline{\mu}^\star}{\overline{\lambda}^\star}\right) f(\varepsilon)$. It follows that $\int_{\mathbb{R}^{J+1}} \overline{g}^\star(\varepsilon) d\varepsilon = 1$. Furthermore, by Assumption 2, it follows that $\overline{g}^\star(\varepsilon) \geq 0$ for all $\varepsilon \in \mathbb{R}^{J+1}$. Hence, we conclude that $\overline{g}^\star(\varepsilon)$ is indeed a probability density, and we call it the robust density associated with the problem (16). A similar argument shows (19). □

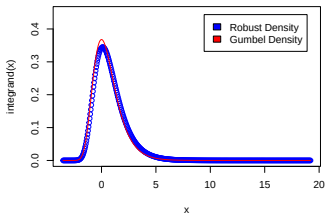
Some remarks are in order. First, the robust density $\overline{g}^\star$ (respectively $\underline{g}^\star$) depends on the choice of the ϕ-divergence through its conjugate ϕ^*. Moreover, the robust density depends on the deterministic utility vector via $H(u,\varepsilon)$, even though the nominal distribution F does not depend on u due to Assumption 1. In addition, $g^\star$ allows us to define the robust distribution function $\overline{G}^\star$ (respectively $\underline{G}^\star$), which, as we shall see, plays a key role in providing an explicit form for $\overline{W}(u)$ (respectively $\underline{W}(u)$). Second, Lemma 3 establishes that the robust density $\overline{g}^\star(\varepsilon)$ incorporates correlation in the elements of the random vector ε through the factor $\phi^{*\prime}((H(u,\varepsilon) - \overline{\mu}^\star)/\overline{\lambda}^\star)$. Thus, even though the nominal distribution F may assume that $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_J$ are independent, the DRO-RUM approach introduces correlation of these terms.

Example 1 *[KL-Divergence] We now consider the case of the Kullback-Leibler divergence. In doing so, we define ϕ as follows:*

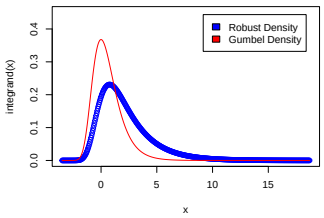
$$\phi(t) \triangleq t \log t, \; t \geq 0 \tag{20}$$

We note that in the previous expression, $0 \log 0 = 0$. Here

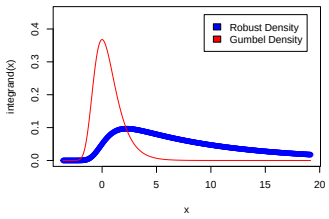
$$\int_{\mathbb{R}^J} \phi(L(\varepsilon)) dF(\varepsilon) \tag{21}$$



(a) $\rho = 0.01$



(b) $\rho = 0.5$



(c) $\rho = 3.5$

Figure 1: Three instances of ρ

defines the Kullback-Leibler divergence, denoted $D_{KL}(G\|F)$. *For* $\lambda > 0$ *the conjugate of* $\lambda\phi$ *is* $(\lambda\phi)^*(y) = \lambda\left(e^{y/\lambda} - 1\right)$. *From Proposition 2 we know that*

$$\overline{W}(u) = \inf_{\lambda \geq 0, \mu} \left\{ \lambda\rho + \mu + \lambda e^{-\mu/\lambda} \mathbb{E}_F \left[e^{H(u,\varepsilon)/\lambda} \right] - \lambda \right\} \tag{22}$$

In (22) minimizing with respect to μ *yields* $\mu^\star = \lambda \ln \mathbb{E}_F \left[e^{H(u,\varepsilon)/\lambda} \right]$. *Plugging* $\mu^\star$ *in (22) we obtain* $\lambda^\star$ *as the solution to*

$$\overline{W}(u) = \inf_{\lambda > 0} \left\{ \lambda\rho + \lambda \ln \mathbb{E}_F \left[e^{H(u,\varepsilon)/\lambda} \right] \right\}. \tag{23}$$

It is well-known that in the case of the KL divergence (e.g., Hu and Hong (2012) and Hansen and Sargent (2001)), the "robust" density is given by:

$$g^\star(\varepsilon) = f(\varepsilon) \frac{e^{H(u,\varepsilon)/\lambda^\star}}{\mathbb{E}_F(e^{H(u,\varepsilon)/\lambda^\star})} \tag{24}$$

where $f(\varepsilon)$ *is the density associated to a nominal distribution,* $H(u,\varepsilon) = \max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\}$ *and* $\overline{\lambda}^\star$ *is the unique optimal solution to (23). To see how the optimal density* (24) *compares to the case where the nominal is a Gumbel distribution, in Figure 1.*

The result in Lemma 3 enables us to characterize the choice probability vector $p^\star(u)$ similarly to the celebrated WDZ theorem. For the sake of exposition, we establish the result only in terms of $\overline{W}(u)$.

Theorem 1 *Let Assumptions 1 and 2 hold. Let* $(\overline{\lambda}^\star, \overline{\mu}^\star)$ *and* $(\underline{\lambda}^\star, \underline{\mu}^\star)$ *be the unique optimal solutions to programs* (16) *and* (17)*, which induce* $\overline{g}^\star, \overline{G}^\star, \underline{g}^\star$*, and* $\underline{G}^\star$*. Then the following statements hold:*

(i) The robust social surplus functions are given by:

$$\overline{W}(u) = \mathbb{E}_{\overline{G}^\star}\left(\max_{j\in\mathcal{J}}\{u_j + \varepsilon_j\}\right) \quad \textit{and} \quad \underline{W}(u) = \mathbb{E}_{\underline{G}^\star}\left(\max_{j\in\mathcal{J}}\{u_j + \varepsilon_j\}\right)$$

(ii) The choice probability vectors $\overline{p}^\star(u)$ *and* $\underline{p}^\star(u)$ *are given by:*

$$\nabla\overline{W}(u) = \overline{p}^\star(u) \quad \textit{and} \quad \nabla\underline{W}(u) = \underline{p}^\star(u).$$

Part (i) of the theorem establishes that given the optimal solutions $\overline{\lambda}^\star$ and $\overline{\mu}^\star$, $\overline{W}(u)$ takes the familiar expected maximum form that characterizes the RUM (see Eq.(2)). The main difference between the characterization in part (i) and the surplus functions from RUM is that expression (1) corresponds to the expectation with respect to the distribution $\overline{G}^\star$. Part (ii) shows that the gradient of $\overline{W}(u)$ yields the choice probability vector $\overline{p}^\star(u)$.

It is worth pointing out that the result in Theorem 1 also holds for the case of the program (17), where we replace $\overline{W}(u)$ and $\overline{p}^\star$ by $\underline{W}(u)$ and $\underline{p}^\star$ respectively. Thus, Theorem 1 generalizes the WDZ to environments where the nominal distribution F may be misspecified or incorrect. In other words, Theorem 1 shows that the DRO-RUM preserves the expected maximum form and the gradient structure of the popular RUM.

3.4 A Robust Fenchel equality

In this section we derive a distributionally robust version of the Fenchel equality for discrete choice models. In order to establish this result, we recall that $\mathcal{U}$ is the set of mean utility vectors with the normalization $u_0 = 0$ for the outside option.

Our first step is to understand the properties of the convex conjugate of $\overline{W}(u)$:

$$\overline{W}^*(p) = \sup_{u\in\mathcal{U}}\left\{\langle u, p\rangle - \overline{W}(u)\right\}. \tag{25}$$

In particular, we are interested in understanding the behavior of $\overline{W}^*(p)$ on its effective domain of:

$$\operatorname{dom}\overline{W}^* = \left\{p \in \mathbb{R}^{J+1} \,|\, \overline{W}^*(p) < \infty\right\}.$$

Similarly, we are interested in the case of the concave conjugate of $\underline{W}(u)$, which is defined as

$$\underline{W}^*(p) = \inf_{u\in\mathcal{U}}\left\{\langle u, p\rangle - \underline{W}(u)\right\}. \tag{26}$$

We are interested in understanding the behavior of $\underline{W}^*(p)$ on its effective domain of:

$$\operatorname{dom}\underline{W}^* = \left\{p \in \mathbb{R}^{J+1} \,|\, \underline{W}^*(p) < \infty\right\}.$$

The following lemma plays a key role in our analysis.

Lemma 4 *Let Assumptions 1 and 2 hold. Then $\overline{W}(u)$ is strictly convex in u. Similarly, $\underline{W}(u)$ is strictly concave in u.*

The following theorem establishes the continuity and smoothness of $\overline{W}^*$.

Theorem 2 *Let Assumptions 1 and 2 hold. The convex conjugate $\overline{W}^*$ is continuous on its domain $\operatorname{dom}\overline{W}^*$ which coincides with the probability simplex Δ_{J+1}. Furthermore, $\overline{W}^*$ is continuously differentiable on int $\operatorname{dom}\overline{W}^*$.*

Now we are ready to establish a robust version of the Fenchel equality

Theorem 3 *Let Assumptions 1 and 2 hold. Then*

(i) For $\overline{W}(u)$ and $\overline{p} \in \Delta_{J+1}, u \in \mathcal{U}$ the following holds:

$$\overline{p} = \nabla\overline{W}(u) \Leftrightarrow u = \nabla\overline{W}^*(\overline{p}). \tag{27}$$

(ii) For $\underline{W}(u)$ and $\underline{p} \in \Delta_{J+1}, u \in \mathcal{U}$ the following holds:

$$\underline{p} = \nabla\underline{W}(u) \Leftrightarrow u = \nabla\underline{W}^*(p). \tag{28}$$

Some remarks are in order. Firstly, parts (i) and (ii) extend Fenchel duality in RUMs to environments where the shock distribution can be misspecified. To illustrate this, it is worth noting that when $\rho = 0$, Theorem 3 reduces to the traditional Fenchel equality in discrete choice models (e.g., Chiong et al. (2016)).

Secondly, a crucial aspect of the result presented in Theorem 6 is the fixed nature of the deterministic mean utility vector u. Specifically, parts (i) and (ii) establish that $\overline{p}$ and $\underline{p}$ must exhibit *consistency* with u. This implies that Theorem 3 offers a method to determine $\overline{p}$ and $\underline{p}$ for a given vector u As we will discover, this fact plays a pivotal role in our analysis of demand inversion.

4 Applications

In this section, we delve into three distinct economic applications of the DRO-RUM approach. First, we explore how our approach can effectively address the demand inversion problem, as originally proposed by Berry (1994). In our second application, we investigate a robust inversion problem involving random coefficients. Finally, in our last application, we examine the concept of limited consideration within the context of the DRO-RUM framework.

4.1 Robust demand inversion

In this section, we discuss the empirical content of the DRO-RUM. In particular, we show how our approach is suitable to recover the mean utility vector allowing for uncertainty about the true distribution generating ε.

To gain some intuition, consider a situation where the choice probability vector $\hat{p}$ is observed from market data. Then the analyst's goal is to find a

vector u that rationalizes the observed $\hat{p}$. Following Berry (1994), this problem is known as the *demand inversion*. In particular, Berry (1994) shows that in the case of the MNL u satisfy the following

$$\hat{p}_j = \frac{e^{u_j}}{1 + \sum_{j'=1}^{J} e^{u_{j'}}} \quad \text{for } j = 1, \ldots, J.$$

and

$$\hat{p}_0 = \frac{1}{1 + \sum_{j'=1}^{J} e^{u_{j'}}}$$

Then using the previous expressions, we can solve for the mean utility vector u as a function of p:

$$\log(\hat{p}_j / \hat{p}_0) = u_j \quad \text{for } j = 1, \ldots, J.$$

In other words, we can express u in terms of the observed $\hat{p}$.

We can use a similar argument to find the vector u in the case of the nested logit, the random coefficient MNL model (Berry (1994); Berry et al. (1995)), and in the case of the inverse product differentiation logit model of Fosgerau et al. (2022). For general RUMs beyond the MNL and its variants, Galichon and Salanié (2021) develops a general approach based on convex duality and mass transportation techniques. They show that for any *fixed* distribution of ε the mean utility vector u is identified from the observed choice probability p.

This section aims to show that the DRO-RUM can be used to study the demand inversion problem in environments where the analyst does not know the true distribution of ε. Thus, our approach allows us to identify u under misspecification of the distribution governing the realizations of ε. In doing so, we use Theorem 1 which implies that:

$$\hat{p}_j = \frac{\partial \overline{W}(u)}{\partial u_j} \quad \forall j \in \mathcal{J}.$$

Furthermore, from Theorem 6 we get:

$$\overline{u}_j = \frac{\partial \overline{W}^*(\hat{p})}{\partial p_j} \quad \text{and} \quad \underline{u}_j = \frac{\partial \overline{W}^*(\hat{p})}{\partial p_j} \; \forall j \in \mathcal{J},$$

where $\overline{u}$ achieves the maximum in (25) and $\underline{u}$ obtains the minimum in (26). Then, given the *robust* distributions $\overline{G}^\star$ and $\underline{G}^\star$, we conclude that $\overline{u}$ and $\underline{u}$ are identified from the observed $\hat{p}$. In other words, we can find the vectors $\overline{u}$ and $\underline{u}$ that rationalize the observed choice probability vector $\hat{p}$.

The following result establishes the empirical content of the DRO-RUM.

Proposition 3 *Let Assumptions 1 and 2 hold. Let $\hat{p}$ be an observed (market data) choice probability vector. Then*

(i) $(\overline{u}^\star, \overline{\lambda}^\star, \overline{\mu}^\star)$ is the unique solution to the strictly convex optimization problem:

$$-\overline{W}^*(\hat{p}) = \inf_{u \in \mathcal{U}, \lambda \in \mathbb{R}_+, \mu \in \mathbb{R}} \left\{ \lambda\rho + \mu + \lambda \mathbb{E}_F \left(\phi^* \left(\frac{H(u,\varepsilon) - \mu}{\lambda} \right) \right) - \langle \hat{p}, u \rangle \right\}. \tag{29}$$

(ii) $(\underline{u}^\star, \underline{\lambda}^\star, \underline{\mu}^\star)$ *is the unique solution to the strictly concave optimization problem:*

$$-\underline{W}^*(\hat{p}) = \sup_{u\in\mathcal{U},\lambda\in\mathbb{R}_+,\mu\in\mathbb{R}} \left\{ -\lambda\rho - \mu - \lambda\mathbb{E}_F\left(\phi^*\left(\frac{-H(u,\varepsilon)-\mu}{\lambda}\right)\right) - \langle \hat{p}, u\rangle \right\}. \tag{30}$$

where $\underline{W}^\star(\hat{p})$ *is the concave conjugate of* $\underline{W}(u)$.

As we discussed in the introduction of this section, for a fixed distribution of ε, parts (i) and (ii) have been established in the Galichon and Salanié (2021). Our result differs from theirs in a fundamental aspect; we achieve the identification of the mean utility vectors $\overline{u}$ and $\underline{u}$, relaxing the assumption that the distribution of ε is known. In other words, our result shows how the robust Fenchel equality yields the nonparametric identification of $\underline{u}$ and $\overline{u}$ under (potential) misspecification of the shock distribution. Similarly, our result relates to dynamic discrete choice models' "inversion" approach(e.g. Chiong et al. (2016)) For instance the papers by Hotz and Miller (1993) and Arcidiacono and Miller (2011) establish that the mean utility vector u can be recovered as $\nabla^{-1}W(\hat{p}) = u$. Their approach only applies to the case of the MNL and GEV models. By exploiting convex optimization techniques, Fosgerau et al. (2021) extends Hotz and Miller (1993) and Arcidiacono and Miller (2011)'s inversion approach to models far beyond the GEV class. Similarly, Li (2018) considers a convex minimization algorithm to solve the demand inversion problem. He illustrates his method in the case of both the Berry et al. (1995) random coefficient logit demand model and the Berry and Pakes (2007) pure characteristics model. However, Fosgerau et al. (2021) and Li (2018)'s results only apply under the assumption that the distribution of ε is known. In contrast, Proposition 3 establishes that given a choice probability vector $\hat{p}$, we can identify $\overline{u}$ and $\underline{u}$ as the unique solutions of the optimization programs (29) and (30) respectively. This latter characterization captures the role of misspecification through the value of the Lagrange multipliers $(\overline{\lambda}^\star, \overline{\mu}^\star)$ and $(\underline{\lambda}^\star, \underline{\mu}^\star)$. Thus, Proposition 3 provides a distributionally robust nonparametric identification result.

4.2 A robust random coefficient model

In this section we show how our Theorem 3 can be applied to study a robust inversion problem with random coefficients. Formally, we analyze the random coefficient model assuming that the ϕ-divergence corresponds to the Kullback-Leibler distance. Following Berry et al. (1995) and Galichon and Salanié (2021), we consider a random coefficient model with $\varepsilon = Ze + T\eta$, where e is a random vector on $\mathbb{R}^k$ with distribution F_e , Z is a $|\mathcal{J}| \times k$ matrix, $T > 0$ is a scalar parameter, and η is a vector of $|\mathcal{J}|$ Gumbel random variables, whose distribution function is F_η. Assume that e and η are statistically independent. Fixing the distributions F_e and F_η, we can use the iterated expectation, combined with the independence of e and η (Eqs. B.6-B.7 in Galichon and Salanié (2021)) we get that

$$\begin{aligned} W(u) &= \mathbb{E}_{F_e}\left(\mathbb{E}_\eta\left(\max_{j\in\mathcal{J}}\{u_j+(Ze)_j+T\eta_j\}\right)|e\right), \\ &= \mathbb{E}_{F_e}\left(W(u+Ze)\right), \end{aligned}$$

where $W(u+Ze)=\int_{\mathbb{R}^{J+1}}\max_{j\in\mathcal{J}}\{u_j+(Ze)_j+T\eta_j\}f_\eta d\eta$. Using the fact that η follows a Gumbel distribution, we find that

$$W(u+Ze)=T\log\left(\sum_{j\in\mathcal{J}}e^{\frac{u_j+(Ze)_j}{T}}\right).$$

Let us assume that F_e approximates the true distribution generating e. Then we can define $\overline{W}(u)$ as follows:

$$\overline{W}(u)=\sup_{G_e\in\mathcal{M}_\phi(F_e)}\mathbb{E}_{G_e}\left(T\log\left(\sum_{j\in\mathcal{J}}e^{\frac{u_j+(Ze)_j}{T}}\right)\right)$$

To apply Theorem 3, we note that $H(u,\varepsilon)=T\log\left(\sum_{j\in\mathcal{J}}e^{\frac{u_j+(Ze)_j}{T}}\right)$. Then, using the Kullback-Leibler distance, we have that for an observable choice probability vector p, the identified mean utility vector $u^\star$ corresponds to the solution of the following program:

$$-\overline{W}^*(\hat{p})=\inf_{u\in\mathcal{U},\lambda>0}\left\{\rho\lambda+\lambda\ln\mathbb{E}_{F_e}\|e^{u+Ze}\|_{T^{-1}}-\langle\hat{p},u\rangle\right\}, \tag{31}$$

where $\|e^{u+Ze}\|_{T^{-1}}\triangleq\left(\sum_{j\in\mathcal{J}}e^{\frac{u_j+(Ze)_j}{T}}\right)^T$.

The program (31) allows us to identify the mean utility vector enabling some degree of misspecification in the distribution of e. It is worth remarking that program (31) is fairly tractable, so we can use traditional stochastic programming algorithms to find its unique solution. For the case of $\underline{W}(u)$ a similar argument can be applied.

4.3 The DRO-RUM and limited consideration

In this section we discuss how to incorporate limited consideration in the DRO-RUM. In particular, we adapt Aguiar et al. (2023)'s costly attention allocation model. Hence, we consider a situation where the DM faces a menu A and needs to allocate her attention, measured by $\pi_A\in\Delta(2^A)$, over all possible consideration sets (including the empty set). Aguiar et al. (2023) propose to measure the attractiveness of a set D by $\alpha(D)\triangleq\mathbb{E}_{F_D}(\max_{j\in D}\{u_j+\varepsilon_j\})$ where F_D is the distribution associated with $\varepsilon_D=(\varepsilon_j)_{j\in D}$ for all $D\in 2^A$. For the case of the empty set $\emptyset$ we can use the normalization $\alpha(\emptyset)=0$. The DM faces a cognitive cost associated with selecting a consideration set. This cost is captured by a function $C:[0,1]\longrightarrow\mathbb{R}\cup\{\infty\}$, which is assumed to be menu independent but depends on the allocated attention $\pi(D)$ as measured by $C(\pi(D))$. In

addition, following the literature on perturbed utility models (Fudenberg et al. (2015)), C is assumed to be convex.

Accordingly, the DM solves the following problem

$$\max_{\pi \in \Delta(2^A)} \left\{ \sum_{D \subseteq A} [\pi(D)\alpha(D) - C(\pi(D))] \right\} \tag{32}$$

In the previous expression, a key assumption is the fact that DM knows the distribution F_D associated with each set D. Following Aguiar et al. (2023), we assume that $K(t) = -(t \log t)/\theta$, where θ is the cost parameter. Accordingly, the unique solution to problem (32) is given by:

$$\pi(D) = \frac{\exp(\theta\alpha(D))}{\sum_{C \subseteq A} \exp(\theta\alpha(C))} \quad \forall\, D \subseteq A \tag{33}$$

Some remarks are in order. First, the allocation rule (33) tends to assign higher probability to sets with more alternatives. To see this, consider $D' \subset D$. In this case, it is easy to see that $\alpha(D') < \alpha(D)$ and $\pi(D') < \pi(D)$. Thus, everything else equals, rule (33) assigns larger probabilities to larger consideration sets.

The second remark about the solution (33) is the fact that it is derived under the assumption that the DM knows the distribution F_D for all $D \in 2^A$. However, we can use the DRO-RUM to build a robust, costly attention allocation framework. To see this, we replace $\alpha(D)$ by its distributionally robust counterpart $\overline{\alpha}(D) \triangleq \sup_{G_A \in \mathcal{M}_\phi(F_D)} \alpha(D)$ where F_D is the nominal distribution under the set $D \in 2^A$. Intuitively, $\overline{\alpha}(D)$ represents a *robust measure of attractiveness*.

To gain some intuition about how the DRO-RUM adds new insights, we consider the case of the χ^2-divergence $\phi(t) = \frac{1}{2}(t-1)^2$. In this case, it is straightforward to show that $\phi^*(s) = s + \frac{1}{2}s^2$. Defining the random variable $H_D \triangleq \max_{j \in D}\{u_j + \varepsilon_j\}$, we can use the expression (16) in Proposition 2 to show that:

$$\overline{\alpha}(D) = \alpha(D) + \sqrt{2\rho\mathbb{V}(H_D)} \quad \forall\, D \subseteq A \tag{34}$$

where $\alpha(D)$ is the original surplus associated with menu D, ρ is the robustness index, and $\mathbb{V}(H_D)$ is the variance of H_D. Then, replacing $\alpha(D)$ by $\overline{\alpha}(D)$ in the program (32) (or solution (33)), we find that a robust optimal attention allocation rule $\overline{\pi}_\rho$ is given by:

$$\overline{\pi}_\rho(D) = \frac{\exp(\theta(\alpha(D) + \sqrt{2\rho\mathbb{V}(H_D)}))}{\sum_{C \subseteq A} \exp(\theta(\alpha(C) + \sqrt{2\rho\mathbb{V}(H_C)}))} \quad \forall D \subseteq A. \tag{35}$$

Some remarks are in order. First, in the expression (35) is easy to see that when $\rho \longrightarrow 0$ we recover $\overline{\pi}_\rho \longrightarrow \pi$. Second, $\bar{\pi}_\rho$ shows that a robust DM will assign probabilities to different sets under limited consideration by considering the sum of $\alpha(D)$ and the standard deviation of H_D. In particular, the expression (35) establishes that the probability of selecting the choice set D is *increasing* in the variability of H_D. In other words, by using the optimistic case for $\alpha(D)$, we

model consideration sets where the DM prefers sets with more variability. This fact is counter intuitive, as it is reasonable to assume that the DM is risk averse and she may penalize the uncertainty associated to different sets. In order to capture this behavior, we consider the pessimistic version of $\alpha(D)$ denoted by $\underline{\alpha}(D)$. Using the χ^2-divergence, we get:

$$\bar{\alpha}(D) = \alpha(D) - \sqrt{2\rho\mathbb{V}(H_D)} \quad \forall D \subseteq A. \tag{36}$$

The expression (36) captures a situation where the DM dislikes the variability associated to H_D. Accordingly, a *pessimistic allocation rule* $\underline{\pi}_\rho(D)$ can be expressed as (35) by replacing $\overline{\alpha}(D)$ with $\underline{\alpha}(D)$.

5 Numerical Experiments

In this section, we discuss numerical simulations of our approach. We compare the DRO-RUM with the MNL and MNP models.[10]

For ease of exposition, in this section we focus in the DRO-RUM defined in (14).Accordingly, our main goal is to analyze the effect of the robustness index ρ on the choice probabilities. We consider a scenario with four alternatives where $\mathcal{J} = \{0, 1, 2, 3\}$. Our first parametrization of the utility vector u is $u = (0, 1, 2, 2.1)^T$. Based on this specification, we proceed to calculate the choice probabilities. In the case of the MNL, the choice probabilities are computed via Eq. (4) , where the scale parameter equals one ($\eta = 1$). In addition, we set the location parameter of each Gumbel error is assumed to zero. For the MNP, we consider two different parametrizations for the variance-covariance matrix of the random error vectors; $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ where

$$\Sigma_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 2 & -0.5 & 0.5 & 1.3 \\ -0.5 & 2 & 0 & 0.15 \\ 0.5 & 0 & 2 & 1 \\ 1.3 & 0.15 & 1 & 2 \end{pmatrix}.$$

We call the latter model MNP-dep and the former MNP-indep, as the random errors ε_j for $j = 0, 1, 2, 3$ are independent in the former model. We use 10,000,000 draws from the error vectors to stabilize the simulations to simulate the choice probabilities.
For the DRO-RUMs, we choose the Kulback-Leibler- divergence case presented in Example 1. We assume that the error terms of the nominal distribution are iid Gumbel distributed with location parameter zero and scale parameter one. This yields a way to examine the behavior and numerical stability of the DRO-RUM, and the impact of ρ on the choice probabilities.

The robust choice probabilities are simulated similarly to the MNP models. However, for the case of DRO-RUM we have to generate samples from the distribution defined by the density (24). First, the optimal $\lambda^\star$ in (23) as

[10]We recall that the MNP assumes that the error terms follow a normal distribution with a specific variance-covariance matrix.

well as $\mathbb{E}_F\left(e^{H(u,\varepsilon)/\lambda^\star}\right)$ are estimated using $50,000,000$ simulations from 4 iid Gumbel distributions. Based on the optimized parameters a higher dimensional acceptance-rejection algorithm provides an efficient sampling method. For performance, the code was written in Julia.[11]

We present the results in Table 2

	Alternative 1	Alternative 2	Alternative 3	Alternative 4
MNL	5.1885%	14.1037%	38.3379%	42.3699%
MNP-indep	1.6243%	10.2996%	41.443%	46.6331%
MNP-dep	1.7877%	19.359%	37.9993%	40.854%
$\rho = 0.1$	9.2391%	18.4783%	33.695%	38.587%
$\rho = 0.7$	13.2132%	19.7823%	31.6066%	35.3979%
$\rho = 1.3$	15.4725%	21.5501%	30.5269%	32.4505%
$\rho = 2.2$	18.1501%	21.46%	29.7536%	30.6363%
$\rho = 4.3$	21.5843%	23.2006%	27.1235%	28.0916%

Table 2: Choice Probabilities for utility vector $u = (0, 1, 2, 2.1)^T$

.

In the previous table, the first row displays the choice probability for the MNL. The second and third rows show the choice probabilities for the MNP-indep and MNP-dep. The fourth row shows the behavior of the DRO-RUM when $\rho = 0.1$. For this parametrization, the DRO-RUM yields choice probabilities that are similar (not equal) to the ones displayed by the MNP-dep. Similar behavior is observed for the case of $\rho = 0.7$.

Rows six to eight show the behavior of the DRO-RUM as we increase ρ. As expected, as the value of ρ increases, the choice probabilities look similar to the uniform choice between alternatives. In particular, for the case of $\rho = 4.3$ we note that DRO-RUM assigns probabilities similar to the uniform case. Intuitively, a large ρ , represents a situation where the analyst is highly uncertain about the true distribution. Thus, her behavior is overly cautious and considers a large set of possible (and feasible) distributions. Hence, when $\rho \longrightarrow \infty$, the analyst's best choice is to guess uniform probabilities.

Similarly, from the DM's perspective, large values of ρ indicate a cautious and flexible choice of the error term. Consequently, the random error term might follow a distribution that completely counteracts the deterministic utilities' effects and guarantees the same overall random utility for every alternative. Indeed, the robust surplus function (22) is strongly increasing with a larger index of robustness as shown in Figure 2, where we plot the surplus function evaluated at $u = (0, 1, 2, 2.1)^T$ for different values of ρ.

[11] The code can be found on Github under `https://github.com/rubsc/rejection_DRO_RUM`. We point out that in order to obtain the results for the DRO-RUM defined by (15), the code can be easily modified.

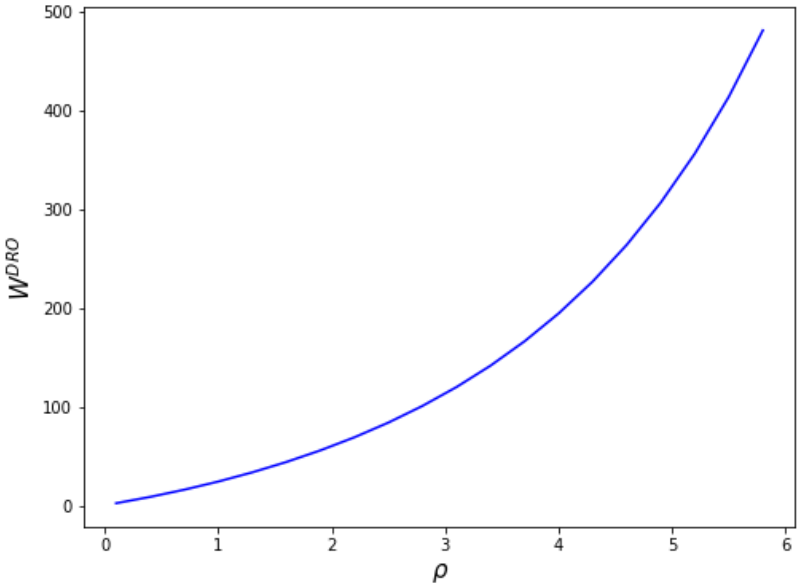


Figure 2: Robust surplus W^{DRO} for the utility vector u and different values of robustness index ρ.

A well-known pitfall of the MNL model is that it satisfies the independence of irrelevant alternatives (IIA) property. The IIA property establishes that the ratio between the probabilities of any two alternatives only depends on the differences between the utilities of these two alternatives. This property follows directly via formula (4). A direct implication of this fact is that when the deterministic utility of one alternative changes, the choice probabilities change proportionally so that the probability ratio between alternatives remains constant. In contrast, the DRO-RUM incorporates some dependence structure into the MNL.[12] Hence, it is interesting to simulate choice probabilities for a slight change in the deterministic utility vector. In Table 3, we summarize the choice the probabilities for the alternatives with utility vector $\tilde{u} = (0, 1, 2, 2.2)^T$.

[12]We recall that we are assuming that the nominal distribution is Gumbel.

	Alternative 1	Alternative 2	Alternative 3	Alternative 4
MNL	4.9671%	13.5021%	36.7024%	44.8284%
MNP-indep	1.4758%	9.5821%	39.2676%	49.6745%
MNP-dep	1.5544%	18.5851%	36.0749%	43.7856%
$\rho = 0.1$	6.8788%	15.0473%	36.5434%	41.5305%
$\rho = 0.7$	13.2076%	20.2437%	30.8569%	35.6918%
$\rho = 1.3$	15.2938%	21.5069%	30.7281%	32.4712%
$\rho = 2.2$	17.9704%	21.4406%	29.5254%	31.0636%
$\rho = 4.3$	21.5509%	23.264%	27.4421%	27.743%

Table 3: Choice Probabilities for utility vector $\tilde{u} = (0, 1, 2, 2.2)^T$

The violation of IIA, is visualized in Figure 3. Note that in the MNL, the decrease in choosing alternative 4 evenly increases the probability of choosing one of the alternatives $1 - 3$, indicated by the dotted line. At the same time, the substitution patterns for the robust models are way more flexible.

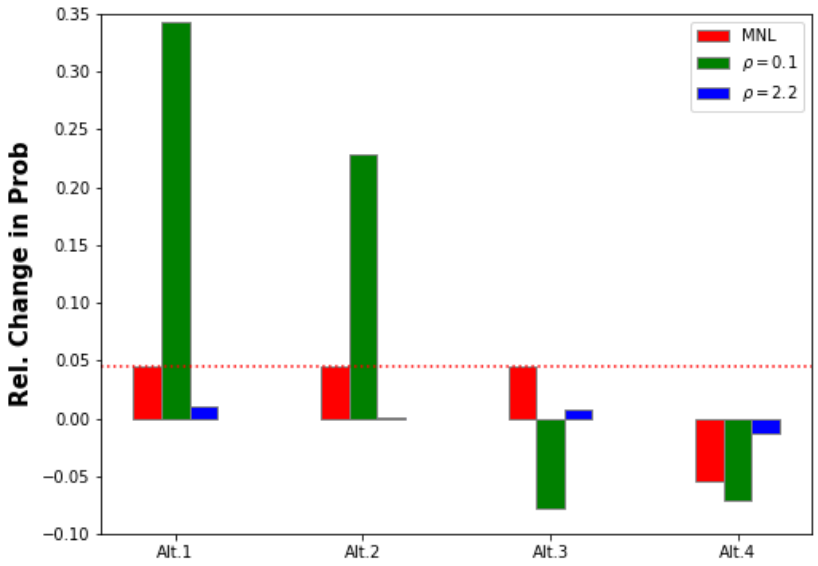


Figure 3: Relative change in probabilities if deterministic utility vector changes from $\tilde{u}$ to u.

6 Extension: A Sinkhorn DRO-RUM approach

In the previous sections, we study the DRO-RUM approach, where the uncertainty set $\mathcal{M}_\phi(F)$ was defined using ϕ-divergences. A natural and important

question arises regarding the generality of our results in terms of different ways of defining the uncertainty set.

The goal of this section is to extend our discussion to the scenario where the set $\mathcal{M}_\phi(F)$ is defined using the Sinkhorn distance. This approach, known as the Sinkhorn-DRO, is a recent development outlined by Wang et al. (2023) in the context of DRO problems.

We begin our analysis by formalizing the notion of Sinkhorn distance. In doing so, we introduce the following notation: for a measurable set $\mathcal{Z}$, $\mathcal{P}(\mathcal{Z})$ denote the set of probability distributions on $\mathcal{Z}$. Similarly, let $\mathcal{M}(\mathcal{Z})$ denote the set of measures in $\mathcal{Z}$.

Definition 2 (Sinkhorn Distance) *Let $\mathcal{Z}$ be a measurable set. Consider distributions $F, G \in \mathcal{P}(\mathcal{Z})$ where $\mathcal{P}(\mathcal{Z})$ is the set of probability distributions on $\mathcal{Z}$. Let $\mu, \nu \in \mathcal{M}(\mathcal{Z})$ be two reference measures such that $F \ll \mu, G \ll \nu$. For regularization parameter $\epsilon \geq 0$, the Sinkhorn distance between two distributions F and G is defined as*

$$\mathcal{W}_\delta(F, G) = \inf_{\gamma \in \Gamma(F,G)} \left\{ \mathbb{E}_{(x,y)\sim\gamma}[c(x, y)] + \delta H(\gamma \mid \mu \otimes \nu) \right\},$$

where $\Gamma(F, G)$ denotes the set of joint distributions whose first and second marginal distributions are F and G respectively, $c(x, y)$ denotes the transport cost, and $H(\gamma \mid \mu \otimes \nu)$ denotes the relative entropy of γ with respect to the product measure $\mu \otimes \nu$:

$$H(\gamma \mid \mu \otimes \nu) = \mathbb{E}_{(x,y)\sim\gamma} \left[\log \left(\frac{\mathrm{d}\gamma(x, y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)} \right) \right],$$

where $\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}$ stands for the density ratio of γ with respect to $\mu \otimes \nu$ evaluated at (x, y).

Some remarks are in order. First, we note that $\mathcal{W}_\epsilon(F, G)$ is defined in terms of an entropic-regularized optimal transport (OT) problem. The goal is to minimize the regularized expected cost by choosing an optimal join distribution γ with prescribed marginals F and G. The entropic regularization term $\delta H(\gamma \mid \mu \otimes \nu)$ is a crucial element in the definition. It distinguishes the Sinkhorn distance $\mathcal{W}_\delta$ from the Wasserstein distance approach (**?**). The introduction of this term, as studied in Wang et al. (2023), facilitates the derivation of strong duality results in the analysis of DRO problems. A second observation is the role of the cost function $c(x, y)$. It is defined for points $x \in \mathrm{supp}F$ and $y \in \mathrm{supp}G$. Different choices for c have been studied in the OT literature, including examples like the Mahalanobis distance $c(x, y) = (x - y)\Omega(x - y)$ where Ω is a positive definite matrix and the ℓ_2-norm $c(x, y) = \frac{1}{2}\|x-y\|_2^2$. The role of c at an intuitive level is similar to the role of the ϕ-function in the ϕ-divergence approach. A third crucial aspect concerns the selection of measures μ and ν. As outlined in Wang et al. (2023), we adopt $\mu = F$, and for ν, we opt for the Lebesgue measure. For an in-depth exploration of this issue, please refer to Remark 4 in Wang et al. (2023). Furthermore, it's noteworthy that in the case where $\delta = 0$,

the Sinkhorn distance coincides with the Wasserstein distance. This section exclusively considers situations where $\delta > 0$. Lastly, to keep the consistency with the DRO-RUM, we set $\mathcal{Z} = \mathbb{R}^{J+1}$.

Using the notion of Sinkhorn distance, we define the set of feasible (admissible) distributions as:

$$\mathcal{M}_{\rho,\delta}(F) = \{G : \mathcal{W}_\delta(F, G) \leq \rho\} \tag{37}$$

Accordingly, we define the optimistic DRO surplus as follows:

$$\widetilde{W}(u) \triangleq \sup_{G \in \mathcal{M}_{\rho,\delta}(F)} \mathbb{E}_G \left(\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right), \tag{38}$$

while the pessimistic case is defined as:

$$\underset{\sim}{W}(u) \triangleq \inf_{G \in \mathcal{M}_{\rho,\delta}(F)} \mathbb{E}_G \left(\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right) \tag{39}$$

It is easy to see that expressions (38) and (39 are similar to the definitions (5) and (6) respectively. The main difference is given by the substitution of the uncertainty set $\mathcal{M}_\phi(F)$ with $\mathcal{M}_{\rho,\delta}(F)$.

In line with Wang et al. (2023), in our analysis we assume the following technical conditions.[13]

Assumption 3 *Let $H(u, z) = \max_{j \in \mathcal{J}}\{u_j + z_j\}$ and $c(\varepsilon, z)$ be a cost function. We assume that the following hold.*

(i) $\nu\{z : 0 \leq c(\varepsilon, z) < \infty\} = 1$ *for μ-almost every ε;*

(ii) $\mathbb{E}_{z \sim \nu}\left[e^{-c(\varepsilon,z)/\delta}\right] < \infty$ *for μ-almost every ε;*

(iii) For every joint distribution γ on $\mathbb{R}^{J+1} \times \mathbb{R}^{J+1}$ with first marginal distribution μ, it has a regular conditional distribution γ_ε given the value of the first marginal equals ε.

(iv) There exists $\lambda > 0$ such that $\mathbb{E}_{z \sim \mathbb{Q}_{\varepsilon,\epsilon}}\left[e^{H(u,z)/(\lambda\delta)}\right] < \infty$ for μ-almost every ε.

We are ready to establish the following result:

Proposition 4 *Let Assumption 1 and 3 hold and define the random variable $H(u, z) \triangleq \max_{j \in \mathcal{J}}\{u_j + z_j\}$. Then the following statements hold:*

(i) The problem (38) *is equivalent to solving the following finite-dimensional convex program:*

$$\widetilde{W}(u) = \inf_{\lambda \geq 0} \left\{ \lambda\rho + \lambda\delta \mathbb{E}_{\varepsilon \sim F}\left[\log \mathbb{E}_{z \sim \nu}\left[e^{(H(u,z) - \lambda c(\varepsilon,z))/(\lambda\delta)} \right] \right] \right\} \tag{40}$$

where λ is the Lagrange multiplier associated to the uncertainty set $\mathcal{M}_{\rho,\delta}(F)$. Furthermore, the program (40) is jointly convex in (λ, u).

[13] We recall that we set $\mu = F$ and ν is the assumed to be the Lebesgue measure.

(ii) The problem (39) *is equivalent to solving the following finite-dimensional concave program:*

$$\underset{\sim}{W}(u) = \sup_{\lambda \geq 0} \left\{ -\lambda\rho - \lambda\delta \mathbb{E}_{\varepsilon \sim F} \left[\log \mathbb{E}_{z \sim \nu} \left[e^{(-H(u,z) - \lambda c(\varepsilon, z))/(\lambda\delta)} \right] \right] \right\} \quad (41)$$

where λ is the Lagrange multiplier associated to the uncertainty set $\mathcal{M}_{\rho,\delta}(F)$. Furthermore, the program (41) is concave in (λ, u).

The preceding outcome characterizes the robust social surplus functions (38) and (39) through a *one*-dimensional optimization problem. Notably, the variable λ corresponds to the Lagrange multiplier linked to the constraint imposed by $\mathcal{M}_{\rho,\delta}(F)$. This characterization bears similarity to the one in Proposition 2. However, there exist noteworthy distinctions between the two. Firstly, the characterization in Proposition 4 is contingent on the choice of the cost function c and the degree of robustness ρ, while the result in Proposition 2 relies on the choice of the ϕ-function. Secondly, in Proposition 4, there is no requirement to compute the conjugate of the cost function c. In the case of the ϕ-divergence function approach, the finite-dimensional characterization is contingent upon knowledge of the convex conjugate of ϕ^*.

For the sake of exposition, throughout this section, we primarily focus on the case of $\widetilde{W}(u)$. We are now ready to generalize the WDZ theorem in the context of the Sinkhorn distance.

Proposition 5 *Let Assumptions 1 and 3 hold. Let $\overline{\lambda}^\star > 0$ be an optimal solution to program (40). Then the following statements hold:*

(i) The optimal distribution is given by:

$$\frac{\mathrm{d}\widetilde{G}^\star(z)}{\mathrm{d}\nu(z)} = \mathbb{E}_{\varepsilon \sim F} \left[\alpha_\varepsilon \cdot \exp\left(\left(H(u,z) - \lambda^* c(\varepsilon, z) \right) / \left(\lambda^* \delta \right) \right) \right]$$

where $\alpha_\varepsilon \triangleq \left(\mathbb{E}_{z \sim \nu} \left[\exp\left(H(u,z) - \lambda^* c(\varepsilon, z) \right) / \left(\lambda^* \delta \right) \right] \right)^{-1}$

(ii) The robust social surplus corresponds to the following:

$$\widetilde{W}(u) = \mathbb{E}_{\widetilde{G}^\star} \left(\max_{j \in \mathcal{J}} \{ u_j + \varepsilon_j \} \right)$$

where $\widetilde{G}^\star$ is the distribution induced by (42).

(iii) The choice probability vector $\widetilde{p}^\star(u)$ is given by:

$$\nabla \widetilde{W}(u) = \widetilde{p}^\star(u).$$

In part (i) of the preceding result, we learn how to construct the optimal distribution $\widetilde{G}^\star$. It is evident that the construction of $\widetilde{G}^\star$ differs from the optimal robust distribution in the case of ϕ-divergences. Part (ii) establishes that the

surplus function $\widetilde{W}(u)$ takes the traditional form of an expected value over the maximum of random utilities. Importantly, part (iii) demonstrates that the WDZ holds in this case. This crucial observation plays a key role in establishing Robust Fenchel duality for situations where the uncertainty set is defined by $\mathcal{M}_{\rho,\delta}(F)$. The following proposition formalizes this result.

Proposition 6 *Let Assumptions 1 and 3 hold. Then:*

(i) For $\widetilde{W}(u)$ and $\widetilde{p} \in \Delta_{J+1}$, and $u \in \mathcal{U}$ the following holds:

$$\widetilde{p} = \nabla \widetilde{W}(u) \Leftrightarrow u = \nabla \widetilde{W}^*(\widetilde{p}). \tag{42}$$

(ii) For $\underset{\sim}{W}(u)$ and $\underset{\sim}{p} \in \Delta_{J+1}$, and $u \in \mathcal{U}$ the following holds:

$$\underset{\sim}{p} = \nabla \underset{\sim}{W}(u) \Leftrightarrow u = \nabla \underset{\sim}{W}^*(\underset{\sim}{p}). \tag{43}$$

The significance of the preceding result lies in its capability to extend the robust Fenchel duality analysis beyond the ϕ-divergence framework to include the Sinkhorn distance. Notably, a direct implication of Proposition 7 is the feasibility of addressing the robust inverse demand problem using the OT approach. The following result formalizes this fact.

Proposition 7 *Let Assumptions 1 and 3 hold. Let $\hat{p}$ be an observed (market data) choice probability vector. Then*

(i) $(\widetilde{u}^\star, \widetilde{\lambda}^\star)$ is the unique solution to the convex optimization problem:

$$-\widetilde{W}^*(\hat{p}) = \inf_{u \in \mathcal{U}, \lambda \in \mathbb{R}_+} \left\{ \lambda\rho + \lambda\delta \mathbb{E}_{\varepsilon \sim F} \left[\log \mathbb{E}_{z \sim \nu} \left[e^{(H(u,z) - \lambda c(\varepsilon, z))/(\lambda\delta)} \right] \right] - \langle \hat{p}, u \rangle \right\}. \tag{44}$$

where $\widetilde{W}^\star(\hat{p})$ is the convex conjugate of $\widetilde{W}(u)$.

(ii) $(\underset{\sim}{u}^\star, \underset{\sim}{\lambda}^\star)$ is the unique solution to the concave optimization problem:

$$-\underset{\sim}{W}^*(\hat{p}) = \sup_{u \in \mathcal{U}, \lambda \in \mathbb{R}_+} \left\{ -\lambda\rho - \lambda\delta \mathbb{E}_{\varepsilon \sim F} \left[\log \mathbb{E}_{z \sim \nu} \left[e^{(-H(u,z) - \lambda c(\varepsilon, z))/(\lambda\delta)} \right] \right] - \langle \hat{p}, u \rangle \right\}. \tag{45}$$

where $\underset{\sim}{W}^\star(\hat{p})$ is the concave conjugate of $\underset{\sim}{W}(u)$.

We conclude this section with a note on the computational complexity of the OT approach. As demonstrated in Proposition 4, the Sinkhorn Sinkhorn DRO-RUM can be formulated as a one-dimensional optimization program. This property allows the utilization of various stochastic optimization algorithms to find both $\widetilde{W}(u)$ and $\underset{\sim}{W}(u)$, respectively. Notably, Wang et al. (2023) propose an algorithm that can be implemented for the analysis of the Sinkhorn DRO-RUM. We defer the exploration of numerical properties and performance, as well as a comparative study with the ϕ-divergence approach, to future research.

7 Final remarks

In this paper, we introduced the DRO-RUM approach, allowing for unknown or misspecified shock distributions. We demonstrated that the DRO-RUM maintains the tractability and convex structure of the traditional Random Utility Model (RUM). Additionally, we characterized the robust Fenchel duality in the context of the DRO-RUM. Our results proved valuable in addressing the demand inversion problem, the DRO-RUM random coefficient model, and a model combining robustness with consideration sets. Furthermore, we established the stability and numerical properties of our approach. Finally, we discussed the straightforward extension of all our results when modeling the uncertainty set using the Sinkhorn distance.

Several potential extensions arise from our findings. For instance, the results presented in this paper could contribute to the study of two-sided matching markets with transferable utility. Additionally, our results are applicable to investigating robust identification in dynamic discrete choice models. The algorithmic aspects of the DRO-RUM also warrant analysis. Notably, a recent introduction by Müller et al. (2022) of a new family of prox-functions on the probability simplex based on discrete choice models raises the intriguing question of whether prox-functions can be generated from the DRO-RUM

References

Victor H. Aguiar, Maria Jose Boccardi, Nail Kashaev, and Jeongbin Kim. Random utility and limited consideration. *Quantitative Economics*, 14(1):71–116, 2023.

Peter Arcidiacono and Robert A. Miller. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867, 2011.

Güzin Bayraksan and David K. Love. *Data-Driven Stochastic Programming Using Phi-Divergences*, chapter 1, pages 1–19. 2015.

Gordon M. Becker, Morris H. Degroot, and Jacob Marschak. Stochastic models of choice behavior. *Behavioral Science*, 8(1):41–55, 1963.

A. Ben-Tal and M. Teboulle. Penalty functions and duality in stochastic programming via φ-divergence functionals. *Mathematics of Operations Research*, 12(2):224–240, 1987.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. ISSN 00251909, 15265501.

Steven Berry and Ariel Pakes. The pure characteristics demand model. *International Economic Review*, 48(4):1193–1225, 2007.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

Steven T. Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262, 1994. ISSN 07416261.

H.D. Block and Jacob Marschak. Random orderings and stochastic theories of response. Cowles Foundation Discussion Papers 66, Cowles Foundation for Research in Economics, Yale University, 1959.

Khai Xiang Chiong, Alfred Galichon, and Matt Shum. Duality in dynamic discrete-choice models. *Quantitative Economics*, 7(1):83–115, 2016.

Timothy Christensen and Benjamin Connault. Counterfactual sensitivity and robustness. *Econometrica*, 91(1):263–298, 2023.

I. Csiszar. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2: 229–318, 1967.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Bernard Dacorogna and Pierre Maréchal. The role of perspective functions in convexity, polyconvexity, rank-one convexity and separate convexity. *Journal of convex analysis*, 15(ARTICLE):271–284, 2008.

Guiyun Feng, Xiaobo Li, and Zizhuo Wang. Technical note—on the relation between several discrete choice models. *Operations Research*, 65(6):1516–1525, 2017.

Mogens Fosgerau, Emerson Melo, Matthew Shum, and Jesper R.-V. Sørensen. Some remarks on ccp-based estimators of dynamic models. *Economics Letters*, 204:109911, 2021. ISSN 0165-1765.

Mogens Fosgerau, Julien Monardo, and Andre de Palma. The inverse product differentiation logit model. *Working paper*, 2022.

Drew Fudenberg, Ryota Iijima, and Tomasz Strzalecki. Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409, 2015. ISSN 00129682, 14680262.

Alfred Galichon and Bernard Salanié. Cupid's Invisible Hand: Social Surplus and Identification in Matching Models. *The Review of Economic Studies*, 89 (5):2600–2629, 12 2021. ISSN 0034-6527.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989. ISSN 0304-4068.

Lars Peter Hansen. Nobel lecture: Uncertainty outside and inside economic models. *Journal of Political Economy*, 122(5):945–987, 2014. ISSN 00223808, 1537534X.

Lars Peter Hansen and Thomas J. Sargent. Robust control and model uncertainty. *The American Economic Review*, 91(2):60–66, 2001. ISSN 00028282.

Lars Peter Hansen and Thomas J. Sargent. *Robustness*. Princeton University Press, stu - student edition edition, 2008.

Jean-Baptiste Hiriart-Urruty and Claude Lemarechal. *Convex Analysis and Minimization Algorithms II*. Addison-Wesley Professional, 2 edition, 1993.

V. Joseph Hotz and Robert A. Miller. Conditional Choice Probabilities and the Estimation of Dynamic Models. *The Review of Economic Studies*, 60(3): 497–529, 07 1993. ISSN 0034-6527.

Zhaolin Hu and L. Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Working Paper*, 2012.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, chapter 6, pages 130–166. 2019.

Lixiong Li. A general method for demand inversion, 2018.

F. Liese and I. Vajda. Convex statistical distances. *Leipzig: Teubner-Texte zur Mathematik, Band 95.*, 1987.

Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006. ISSN 00129682, 14680262.

Massimo Marinacci. Model uncertainty. *Journal of the European Economic Association*, 13(6):1022–1100, 2015.

Jacob Marschak. Binary choice constraints on random utility indicators. Technical report, 1959.

D. McFadden. Modeling the choice of residential location. *in A. Karlqvis, A., Lundqvist, L., Snickars, L., Weibull, J. (eds.), Spatial Intearction Theory and Planning Models (North Holland, Amsterdam)*, pages 531–551, 1978a.

D. McFadden. Modeling the choice of residential location. *Transportation Research Record*, (673):72–77, 1978b.

D. McFadden. *Spatial interaction theory and residential location*, chapter : Modeling the choice of residential location, pages 75–96. North-Holland, Amsterdam, 1978c.

D. McFadden. *Structural Analysis of Discrete Data with Econometric Applications*, chapter : Econometric Models of Probabilistic Choice, pages 198–272. Cambridge: MIT, 1981.

Daniel McFadden. Economic choices. *American Economic Review*, 91(3):351–378, June 2001.

Emerson Melo, Kirill Pogorelskiy, and Matthew Shum. Testing the quantal response hypothesis. *International Economic Review*, 60(1):53–74, 2019.

Vinit Kumar Mishra, Karthik Natarajan, Hua Tao, and Chung-Piaw Teo. Choice prediction with semidefinite optimization when utilities are correlated. *IEEE Transactions on Automatic Control*, 57(10):2450–2463, 2012.

Vinit Kumar Mishra, Karthik Natarajan, Dhanesh Padmanabhan, Chung-Piaw Teo, and Xiaobo Li. On theoretical and empirical aspects of marginal distribution choice models. *Management Science*, 60(6):1511–1531, 2014.

David Müller, Yurii Nesterov, and Vladimir Shikhman. Discrete choice prox-functions on the simplex. *Mathematics of Operations Research*, 47(1):485–507, 2022.

Karthik Natarajan, Miao Song, and Chung-Piaw Teo. Persistency model and its applications in choice modeling. *Management Science*, 55(3):453–469, 2009.

L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC, 1 edition, 2005.

R. Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In Jean Pierre Gossez, Enrique José Lami Dozo, Jean Mawhin, and Lucien Waelbroeck, editors, *Nonlinear Operators and the Calculus of Variations*, pages 157–207, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg. ISBN 978-3-540-38075-7.

T. R. Rockafellar. *Convex Analysis*. 1970.

Andrzej Ruszczyński and Alexander Shapiro. *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*. 2021.

H. Scarf. A min-max solution of an inventory problem. *Studies in The Mathematical Theory of Inventory and Production*, 1958.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Xiaoxia Shi, Matthew Shum, and Wei Song. Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. *Econometrica*, 86(2): 737–761, 2018.

Tomasz Strzalecki. Axiomatic foundations of multiplier preferences. *Econometrica*, 79(1):47–73, 2011. ISSN 00129682, 14680262.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, December 2023.

A Proofs

A.1 Proof of Proposition 1

The proof that $D_\phi(G\|F)$ is well defined and nonnegative follows from Proposition 1 in Ben-Tal and Teboulle (1987). The convexity of D_ϕ follows from Proposition 2 in Ben-Tal and Teboulle (1987). □

A.2 Proof of Lemma 1

(i) The definition provides

$$\overline{W}(u + c \cdot e) = \sup_{G \in \mathcal{M}_\phi(F)} \left\{ \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j + c\} \right] \right\}$$

Due to the linearity of the expectation, it holds

$$c + \sup_{G \in \mathcal{M}_\phi(F)} \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right] = c + \overline{W}(u).$$

(ii) Take any $u, v \in \mathbb{R}^{J+1}$ with $u \geq v$. First we note that for any arbitrary feasible distribution $G \in \mathcal{M}_\phi(F)$ it holds

$$\overline{W}(u) \geq \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right] \overset{(*)}{\geq} \mathbb{E}_G \left[\max_{j \in \mathcal{J}} \{v_j + \varepsilon_j\} \right],$$

where $(*)$ holds due to the monotonicity of the expectation. Taking the supremum on the right-hand side, we conclude that $\overline{W}(u) \geq \overline{W}(v)$.

(iii) We deduce that for any $i \in \mathcal{J}$

$$\overline{W}(u) \geq \mathbb{E}_F \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right] \geq \mathbb{E}_F \left[u_i + \varepsilon_i \right] \geq u_i + \min_{j \in \mathcal{J}} \mathbb{E}_F \left[\varepsilon_j \right],$$

which is finite due to Assumption 1.

(iv) We deduce that for any $i \in \mathcal{J}$

$$\underline{W}(u) \leq \mathbb{E}_F \left[\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\} \right] \leq \max_{j \in \mathcal{J}} u_j + \mathbb{E}_F \left[\max_{j \in \mathcal{J}} \varepsilon_j \right],$$

which is finite due to Assumption 1.

□

A.3 Proof of Proposition 2

Part (i) follows from a direct application of Proposition 7.9 in Ruszczyński and Shapiro (2021). The proof of part (ii) follows from applying the same arguments. For completeness, we provide the details of the argument behind (i). First, we note that for a fixed utility vector u and using the likelihood ratio $L(\varepsilon) \triangleq dG(\varepsilon)/dF(\varepsilon)$, the DRO-RUM in (14) can be expressed as:

$$\begin{aligned}\overline{W}(u) &= \sup_G\{\mathbb{E}_G(H(u,\varepsilon)) : G \in \mathcal{M}_\phi(F)\} \\ &= \sup_{L\geq 0}\{\mathbb{E}_F[L(\varepsilon)H(u;\varepsilon)] \mid \mathbb{E}_F[\phi(L(\varepsilon))] \leq \rho, \mathbb{E}_F[L(\varepsilon)] = 1\} \quad (46)\end{aligned}$$

where the supremum is over a set of measurable functions.

The Lagrangian of problem (46) is :

$$\mathcal{L}(L,\lambda,\mu) = \int_{\mathbb{R}^{J+1}} [L(\varepsilon)H(u,\varepsilon) - \lambda\phi(L(\varepsilon)) - \mu L(\varepsilon)]dF(\varepsilon) + \lambda\rho + \mu. \quad (47)$$

The Lagrangian dual of problem (47) is the problem

$$\inf_{\lambda\geq 0,\mu\in\mathbb{R}} \sup_{L\geq 0} \mathcal{L}(L,\lambda,\mu) \quad (48)$$

Since Slater condition holds for problem (47)[14], there is no duality gap between (47) and its dual problem (48). Moreover, the dual problem has a nonempty and bounded set of optimal solutions.

By the interchangeability principle (Theorem 3A in Rockafellar (1976)), the maximum in (48) can be taken inside the integral, that is

$$\begin{aligned}&\sup_{L\geq 0} \int_{\mathbb{R}^{J+1}} [L(\varepsilon)H(u,\varepsilon) - \mu L(\varepsilon) - \lambda\phi(L(\varepsilon))]dF(\varepsilon) \\ &\quad = \int_{\mathbb{R}^{J+1}} \sup_{t\geq 0}\{t(H(u,\varepsilon)-\mu) - \lambda\phi(t)\}dF(\varepsilon),\end{aligned}$$

Noting that $(\lambda\phi)^*(H(u,\varepsilon)-\mu) = \sup_{t\geq 0}\{t(H(u,\varepsilon)-\mu)-\lambda\phi(t)\}$, then it follows that

$$\overline{W}(u) = \inf_{\lambda\geq 0,\mu\in\mathbb{R}} \left\{\lambda\rho + \mu + \mathbb{E}_F\left[(\lambda\phi)^*(H(u,\varepsilon)-\mu)\right]\right\}. \quad (49)$$

To show the convexity with respect to λ and μ we note that it suffices in (48) and (49) to take the inf with respect to $\lambda > 0$ rather than $\lambda \geq 0$, and that $(\lambda\phi)^*(y) = \lambda\phi^*(y/\lambda)$ for $\lambda > 0$. Therefore $W^{DRO}(u)$ is given by the optimal value of the following problem:

$$\inf_{\lambda>0,\mu\in\mathbb{R}} \{\lambda\rho + \mu + \lambda\mathbb{E}_F\left[\phi^*((H(u,\varepsilon)-\mu)/\lambda)\right]\} \quad (50)$$

Note that $\phi^*(\cdot)$ is convex. Hence, $\lambda\phi^*(y/\lambda)$ is jointly convex in y and $\lambda > 0$. It follows that the objective function of problem (50) is a convex function of $\lambda > 0$ and $\mu \in \mathbb{R}$ with $y = H(u,\varepsilon) - \mu$. Hence (50) is a convex problem. □

A.4 Proof of Corollary 1

Let us look at problem (46). If $\rho = 0$ we get from one constraint that

$$\mathbb{E}_F[\phi(L(\varepsilon))] \leq 0.$$

[14]For instance, we can take $L(\varepsilon) = 1$ for all $\varepsilon \in \mathbb{R}^{J+1}$.

Due to the definition of ϕ, this implies that $L(\varepsilon) = 1$. Hence, the Lagrangian simplifies since the supremum over the densities becomes trivial. Let us plug $L(\varepsilon) = 1$ into Equation (47):

$$\mathcal{L}(L, \lambda, \mu) = \int_{\mathbb{R}^{J+1}} H(u, \varepsilon) - \lambda \cdot \underbrace{\phi(1)}_{=0} - \mu \cdot 1] dF(\varepsilon) + \lambda \cdot 0 + \mu.$$

The latter is equivalent to

$$\mathbb{E}_F \left[H(u, \varepsilon) - \mu \right] + \mu = \mathbb{E}_F \left[H(u, \varepsilon) \right],$$

where the last equality holds due to the linearity of expectation. We indeed recover $W(u)$ for any distribution satisfying Assumption 1. A similar argument applies to $\underline{W}(u)$. Thus, we conclude that $\overline{W}(u) = \underline{W}(u) = W(u)$. □

A.5 Proof of Lemma 2

Due to Assumption 2, the function ϕ^* is strictly convex. Following similar steps as Dacorogna and Maréchal (2008), it follows that $\lambda \cdot \phi^*(\frac{s}{\lambda})$, $\lambda > 0$, is strictly convex. Further, the sum of a convex and strictly convex is strictly convex. This latter fact immediately implies strict convexity of the objective function in λ and μ. Given the strict convexity in λ and μ, it follows that program (16) has a unique solution. A similar argument holds for the program (17). □

A.6 Proof of Theorem 1

We only show part (i). The proof for (ii) is identical. To show the first part, let us define the function $\Psi(\cdot)$ as follows $\Psi(\lambda, \mu) \triangleq \lambda\rho + \mu + \lambda \mathbb{E}_F \left(\phi^* \left(\frac{H(u,\varepsilon) - \mu}{\lambda} \right) \right)$. Optimizing $\Psi(\lambda, \mu)$ with respect to λ and μ we get

$$\begin{aligned} \frac{\partial \Psi(\lambda, \mu)}{\partial \lambda} &= \rho + \mathbb{E}_F \left(\phi^* \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \right) \\ &+ \lambda \mathbb{E}_F \left(\phi^{*\prime} \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \left(\frac{H(u, \varepsilon) - \mu}{-\lambda^2} \right) \right) = 0 \\ \frac{\partial \Psi(\lambda, \mu)}{\partial \mu} &= 1 - \mathbb{E}_F \left(\phi^{*\prime} \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \right) = 0 \end{aligned} \tag{51}$$

Rearranging the first equation, we have:

$$\begin{aligned} &\lambda\rho + \lambda \mathbb{E}_F \left(\phi^* \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \right) + \mu \mathbb{E}_F \left(\phi^{*\prime} \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \right) \\ =&\mathbb{E}_F \left(\phi^{*\prime} \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) H(u, \varepsilon) \right). \end{aligned}$$

Similarly, in the second equation, we have:

$$\mathbb{E}_F \left(\phi^{*\prime} \left(\frac{H(u, \varepsilon) - \mu}{\lambda} \right) \right) = 1$$

Combining both expressions we find that the optimal $\lambda^\star$ and $\mu^\star$ must satisfy:

$$\lambda^\star \rho + \lambda^\star \mathbb{E}_F \left(\phi^* \left(\frac{H(u, \varepsilon) - \mu^\star}{\lambda^\star} \right) \right) + \mu^\star = \mathbb{E}_F \left(\phi^{*\prime} \left(\frac{H(u, \varepsilon) - \mu^\star}{\lambda^\star} \right) H(u, \varepsilon) \right).$$

Using expression (18) in Lemma 3, we obtain that the optimal solution $\overline{\lambda}^\star$ and $\overline{\mu}^\star$ satisfies:

$$\Psi(\overline{\lambda}^\star, \overline{\mu}^\star) = \mathbb{E}_F\left(\phi^{*\prime}\left(\frac{H(u,\varepsilon)-\overline{\mu}^\star}{\overline{\lambda}^\star}\right)H(u,\varepsilon)\right) = \mathbb{E}_{\overline{G}^\star}\left(\max_{j\in\mathcal{J}}\{u_j+\varepsilon_j\}\right).$$

Hence, we conclude that

$$\overline{W}(u) = \mathbb{E}_{\overline{G}^\star}\left(\max_{j\in J}\{u_j+\varepsilon_j\}\right).$$

(ii) To show that $\nabla\overline{W}(u) = \overline{p}^\star(u)$, we note that using the optimized value $\Psi(\overline{\lambda}^\star, \overline{\mu}^\star) = \overline{\lambda}^\star\rho + \overline{\mu}^\star + \overline{\lambda}^\star\mathbb{E}_F\left(\phi^*\left(\frac{H(u,\varepsilon)-\overline{\mu}^\star}{\overline{\lambda}^\star}\right)\right)$ we get:

$$\begin{aligned}
\frac{\partial\overline{W}(u)}{\partial u_j} &= \frac{\partial\Psi(\overline{\lambda}^\star,\overline{\mu}^\star)}{\partial u_j} \\
&= \int_{\varepsilon\in\mathbb{R}^{J+1}}\left(\phi^{*\prime}\left(\frac{H(u,\varepsilon)-\overline{\mu}^\star}{\overline{\lambda}^\star}\right)\frac{\partial H(u,\varepsilon)}{\partial u_j}\right)f(\varepsilon)d\varepsilon \\
&= \mathbb{E}_{\overline{G}^\star}\left(\frac{\partial H(u,\varepsilon)}{\partial u_j}\right) \\
&= \overline{p}_j^\star(u).
\end{aligned}$$

As previous result holds for all $j\in\mathcal{J}$, we get that $\nabla\overline{W}(u) = \overline{p}^\star(u)$. □

A.7 Proof of Lemma 4

For given λ and μ, for u_1, u_2 and $\alpha\in(0,1)$ we have

$$\begin{aligned}
&\lambda\mathbb{E}_F\left(\phi^*\left(\frac{H(\alpha u_1+(1-\alpha)\,u_2,\varepsilon)-\mu}{\lambda}\right)\right) \\
&\overset{(\star)}{\leq} \lambda\mathbb{E}_F\left(\phi^*\left(\frac{\alpha H(u_1,\varepsilon)+(1-\alpha)\,H(u_2,\varepsilon)-\mu}{\lambda}\right)\right),
\end{aligned}$$

where $(\star)$ holds due to the convexity of H and the monotonicity of ϕ^* due to Assumption 2. Exploiting the strict convexity of ϕ^* and the linearity and monotonicity of the expectation operator further yields:

$$\begin{aligned}
&\lambda\mathbb{E}_F\left(\phi^*\left(\frac{\alpha H(u_1,\varepsilon)+(1-\alpha)\,H(u_2,\varepsilon)-\mu}{\lambda}\right)\right) \\
&< \alpha\lambda\mathbb{E}_F\left(\phi^*\left(\frac{H(u_1,\varepsilon)-\mu}{\lambda}\right)\right)+(1-\alpha)\,\lambda\mathbb{E}_F\left(\phi^*\left(\frac{H(u_2,\varepsilon)-\mu}{\lambda}\right)\right).
\end{aligned}$$

Thus it follows that $\overline{W}(u)$ is strictly convex in u. To establish the concavity of $\underline{W}(u)$, we can apply the similar line of reasoning as before noting that $-H(u,\varepsilon)$ and $-\phi^*$ are concave functions. □

A.8 Proof of Theorem 2

Let us first show that $\operatorname{dom} \overline{W}^* \subseteq \Delta_{J+1}$. Fix a utility vector $\bar{u}$ and take any $p \in \mathbb{R}^{J+1}$ with $\langle p, e\rangle \neq 1$. Then, using Lemma 1 (iii) we have

$$\begin{aligned}\overline{W}^*(p) &\geq \sup_{\gamma\in\mathbb{R}} \langle p, \bar{u} + \gamma \cdot e\rangle - \overline{W}(\bar{u} + \gamma \cdot e) \\ &\overset{(iii)}{=} \langle p, \bar{u}\rangle - \overline{W}(\bar{u}) + \sup_{\gamma\in\mathbb{R}} \gamma\left(\langle e, p\rangle - 1\right) = \infty.\end{aligned}$$

Next, we take any vector $p \in \mathbb{R}^{J+1}$ with $p_i < 0$ for some $i \in \{0, 1, \ldots, J\}$. By Lemma 1 (ii), it follows that

$$\overline{W}^*(p) \geq \sup_{\gamma<0} \langle p, \gamma \cdot e_i\rangle - \overline{W}(e_i) \overset{(ii)}{\geq} \sup_{\gamma<0} \gamma \cdot p_i, -\overline{W}(0) = \infty.$$

Hence, it remains to prove the reverse implication, i. e. $\Delta_{J+1} \subseteq \operatorname{dom} \overline{W}^*$. Therefore, we derive an upper bound for the convex conjugate on the simplex:

$$\begin{aligned}\sup_{p\in\Delta_{J+1}} \overline{W}^*(p) &= \sup_{p\in\Delta_{J+1}} \left(\sup_{u\in\mathcal{U}} \langle p, u\rangle - \overline{W}(u)\right) \\ &= \sup_{u\in\mathcal{U}} \left(\sup_{p\in\Delta_{J+1}} \langle p, u\rangle - \overline{W}(u)\right).\end{aligned}$$

We apply (iii) from Lemma 1 which yields

$$\sup_{u\in\mathcal{U}} \left(\sup_{p\in\Delta_{J+1}} \langle p, u\rangle - \overline{W}(u)\right) = \sup_{u\in\mathcal{U}} \left(\max_{i\in\mathcal{J}} u_i - \overline{W}(u)\right) \leq -\min_{i\in\mathcal{J}} \mathbb{E}_F\left[\varepsilon_i\right].$$

Thus, the domain coincides with the simplex. For the continuity, we first observe that $\overline{W}^*$ is convex, and hence it is continuous on the relative interior of its domain. The Gale-Klee-Rockafellar theorem provides upper semi-continuity of $\overline{W}^*$ if the domain is polyhedral, which it is (Rockafellar, 1970). Furthermore, convex conjugates are always lower semi-continuous, and hence continuity follows. In order to establish that $\overline{W}^*$ is continuously differentiable, we note that Lemma 4 shows that $\overline{W}^*$ is strictly convex in u. Then by Theorem 4.1.1 in Hiriart-Urruty and Lemarechal (1993), we know that the strict convexity of $\overline{W}(u)$ implies that $\overline{W}^*(p)$ is continuously differentiable on int $\left(\operatorname{dom} \overline{W}^*\right)$. □

A.9 Proof of Theorem 3

The equivalence of parts (i) and (ii) follows from Theorems 1 and 2, which allows us to invoke Fenchel equality to conclude the result. □

A.10 Proof of Proposition 3

Proposition 2 implies that the previous expression corresponds to

$$\overline{W}^*(\hat{p}) = \sup_{u\in\mathcal{U}} \left\{\langle \hat{p}, u\rangle - \inf_{\lambda>0,\mu\in\mathbb{R}} \left\{\lambda\rho + \mu + \lambda\mathbb{E}_F\left(\phi^*\left(\frac{H(u,\varepsilon) - \mu}{\lambda}\right)\right)\right\}\right\}.$$

Equivalently, we have:

$$\overline{W}^*(\hat{p}) = -\inf_{u\in\mathcal{U},\lambda>0,\mu\in\mathbb{R}} \left\{ \lambda\rho + \mu + \lambda\mathbb{E}_F\left(\phi^*\left(\frac{H(u,\varepsilon)-\mu}{\lambda}\right)\right) - \langle \hat{p}, u\rangle \right\}.$$

Thus, we get:

$$-\overline{W}^*(\hat{p}) = \inf_{u\in\mathcal{U},\lambda>0,\mu\in\mathbb{R}} \left\{ \lambda\rho + \mu + \lambda\mathbb{E}_F\left(\phi^*\left(\frac{H(u,\varepsilon)-\mu}{\lambda}\right)\right) - \langle \hat{p}, u\rangle \right\}.$$

Combining Lemmas 2 and 4 we get that the (29) is strictly convex in u, λ, and μ. As a consequence, there exists a unique solution to the problem (29). □

A.11 Proof of Proposition 4

(i) Under Assumptions 1 and 3 allows us to apply Theorem 1[(iii)] in Wang et al. (2023) to establish that (39) is equivalent to solve the finite dimensional program (41). To establish the convexity in (λ, u), we note that $\lambda\delta\mathbb{E}_{\varepsilon\sim F}\left[\log\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda c(\varepsilon,z))/(\lambda\delta)}\right]\right]$ is convex in (λ, u). Then by adding this convex function to $\lambda\rho$, we conclude that program (41) is convex in (λ, u). (ii) To show this we note that

$$\underset{\sim}{W}(u) = -\sup_{G\in\mathcal{M}_{\rho,\delta}(F)} -W(u) = -\sup_{G\in\mathcal{M}_{\rho,\delta}(F)} \mathbb{E}_G\left(-\max_{j\in\mathcal{J}}\{u_j+\varepsilon_j\}\right).$$

Then, we can apply the argument in part (i) to

$$\sup_{G\in\mathcal{M}_{\rho,\delta}(F)} \mathbb{E}_G\left(-\max_{j\in\mathcal{J}}\{u_j+\varepsilon_j\}\right),$$

and the conclusion follows at once. □

A.12 Proof of Proposition 5

The proof of part (i) follows from Remark 4 in Wang et al. (2023). To show part (ii), we note that for $\lambda^\star$, we note that from the first order condition we get:

$$\begin{aligned}\lambda^*\rho + \lambda^*\delta\mathbb{E}_{\varepsilon\sim F}\left[\log\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda c(\varepsilon,z))/(\lambda^*\delta)}\right]\right] &= \mathbb{E}_{\varepsilon\sim F}\left[\frac{\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda^* c(\varepsilon,z))/(\lambda^*\delta)}H(u,z)\right]}{\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda^* c(x,z))/(\lambda^*\delta)}\right]}\right], \\ &= \mathbb{E}_{\underset{\sim}{G}^\star}\left(\max_{j\in\mathcal{J}}\{u_j+\varepsilon_j\}\right).\end{aligned}$$

To establish (iii), we note that from (ii) with $\lambda^\star > 0$, we can use the envelope theorem to obtain:

$$\frac{\partial}{\partial u_i}\left[\lambda^*\rho + \lambda^*\delta\mathbb{E}_{\varepsilon\sim F}\left[\log\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda c(\varepsilon,z))/(\lambda^*\delta)}\right]\right]\right] = \mathbb{E}_{\varepsilon\sim F}\left[\frac{\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda^* c(\varepsilon,z))/(\lambda^*\delta)}\frac{\partial H(u,z)}{\partial u_j}\right]}{\mathbb{E}_{z\sim\nu}\left[e^{(H(u,z)-\lambda^* c(x,z))/(\lambda^*\delta)}\right]}\right].$$

From part (ii) it follows that the last term can be expressed as

$$\frac{\partial\widetilde{W}(u)}{\partial u_j} = \mathbb{E}_{\underset{\sim}{G}^\star}\left(\frac{\partial H(u,\varepsilon)}{\partial u_j}\right) = \underset{\sim}{p}_j^\star(u).$$

The previous conclusion holds for all $j \in \mathcal{J}$. Thus, it follows that $\nabla\widetilde{W}(u) = \widetilde{p}^{\star}(u)$. □

A.13 Proof of Proposition 6

(i) In order to prove (i) we first note that the properties in Lemma 1 holds for the case $\widetilde{W}(u)$ and $\widetilde{W}(u)$ are satisfied. In particular, for the case $\widetilde{W}(u)$ we have that property (iii) in $\widetilde{W}(u)$ Lemma 1 holds. Then, the argument in Theorem 2 also applies to the case of $\widetilde{W}^{*}(p)$ and we get that $\nabla\widetilde{W}^{*}(p) = u$.The combination of this fact with Proposition 5(iii) allows us to invoke Fenchel equality to conclude (42). The equivalence of parts (i) and (ii) follows from Theorems 1 and 2, which allows us to invoke Fenchel equality to conclude the result. □

A.14 Proof of Proposition 7

The proof follows from applying the same argument used in th proving Proposition 3. □