DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

ESTIMATING DYNAMIC DISCRETE CHOICE MODELS VIA CONVEX ANALYSIS

Khai X. Chiong
California Institute of Technology

Alfred Galichon
Science Po (Paris, France)

Matthew Shum
California Institute of Technology

SOCIAL SCIENCE WORKING PAPER 1374

May 2013

ESTIMATING DYNAMIC DISCRETE CHOICE MODELS VIA CONVEX ANALYSIS

KHAI X. CHIONG[§], ALFRED GALICHON[†], AND MATT SHUM[♣]

Abstract. Using results from convex analysis, we characterize the identification and estimation of dynamic discrete-choice models based on the random utility framework. Based on these insights, we propose a new two-step estimator for these models, which is easily applicable to models in which the utility shocks may not derive from an extreme-value distribution, and may be mutually correlated with each other and with the state variables. Monte Carlo results demonstrate the good performance of this estimator, and we provide a short application using the dynamic bus engine replacement model in Rust (1987).

1. Introduction

Results on identification of dynamic discrete choice models (e.g. Magnac and Thesmar (2002)) allow for quite general specification of the additive choice-specific utility shocks, allowing for dependence and correlation of these shocks with the state variables. However, in practice, almost all applications of these models maintain the restriction assumption that the utility shocks are distributed i.i.d. type 1 extreme value, leading to choice probabilities taking the multinomial logit form. No doubt this is due to the computational convenience of the logit model, because in that case a number of structural components of the model have convenient, analytical closed forms.

The contribution of our paper is twofold. First, we show how the powerful tools of convex analysis can be used to describe the empirical content of dynamic discrete-choice models.

Date: April 2013. Preliminary version; comments welcome.

The authors thank Thierry Magnac for helpful comments, and John Rust for his data. Galichon's research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n°313699 and from FiME, Laboratoire de Finance des Marchés de l'Energie (www.fime-lab.org).

Based upon these insights, we exploit the convex nature of the problem to propose a new estimator for dynamic discrete-choice models which can accommodate any distribution of the utility shocks. Our findings expand the set of dynamic discrete-choice models suitable for applied work far beyond those with extreme-value distributed utility shocks.

2. Basic Model

2.1. **The framework.** In this section we review the basic dynamic discrete-choice setup, as encapsulated in Rust's (1987) seminal paper. The state variable is $X_t \in \mathcal{X}$ which, for convenience, we assume to be finite discrete-valued. Agents choose actions $Y_t \in \mathcal{Y}$ from a finite space $\mathcal{Y}$.

The single-period utility which an agent derives from choosing the action Y_t in period t is given by

$$\bar{u}\left(Y_t, X_t\right) + \epsilon_{Y_t}$$

where ϵ_{Y_t} denotes the utility shock pertaining to action Y_t, which differs across agents. Across agents and time periods, the set of utility shocks $\{\epsilon_y\}_{y\in\mathcal{Y}}$ is distributed according to a joint distribution function $Q_\epsilon(\cdots; X_t)$ which can depend on the current values of the state variables X_t.

Following Rust (1987), and most of the subsequent papers in this literature, we maintain the following conditional independence assumption (which rules out serially persistent forms of unobserved heterogeneity):

Assumption 1. *The transition probability $P\left(X_{t+1}|X_t, Y_t\right)$ is unaffected by ϵ.*

The discount rate is β. Agents are dynamic optimizers who solve

$$Y^* \in \arg\max_Y \left\{\bar{u}\left(Y, X\right) + \epsilon_Y + \beta\mathbb{E}\left[V\left(X', \epsilon'\right)|X, Y\right]\right\} \tag{1}$$

where the prime in X' denotes the realization at the next period, and the value function is recursively defined as

$$V\left(X, \epsilon\right) = \max_Y \left\{\bar{u}\left(Y, X\right) + \epsilon_Y + \beta\mathbb{E}\left[V\left(X', \epsilon'\right)|X, Y\right]\right\}.$$

We define the conditional choice probabilities (CCP's)

$$p(y|x) \equiv Prob(Y = y|X = x).$$

Defining

$$V(x) = \mathbb{E}\left[V(X, \epsilon) | X = x\right],$$

V solves the following equation

$$\begin{aligned} V(x) &= \sum_{y \in \mathcal{Y}} p(y|x) \left(\bar{u}(y, x) + \mathbb{E}[\epsilon_y | y, x] + \beta \sum_{x'} p\left(x'|x, y\right) V\left(x'\right) \right) \\ &= \sum_{y \in \mathcal{Y}} p(y|x) \left(\bar{u}(y, x) + \mathbb{E}[\epsilon_y | y, x] \right) + \beta \sum_{x'} p\left(x'|x\right) V\left(x'\right) \end{aligned} \tag{2}$$

found eg. in Pesendorfer and Schmidt-Dengler (2008), where

$$p\left(x'|x\right) = \sum_{y, x'} p(y|x)\, p\left(x'|x, y\right).$$

In the literature, $V(x)$ is called the *integrated* or *ex-ante value function*, because it measures the continuation value of the dynamic optimization problem before the agent observes his shocks ϵ, so that the optimal action is still stochastic from the agent's point of view.

2.2. **Convex analysis approach.** We now recast these results using convex analysis. This allows us to derive the relationships between the observables (which are the choice-probabilities $p(y|x)$), and the unobserved functions of interests which we want to identify and estimate (which are ultimately the per-period utilities $\bar{u}(y, x)$, and along the way, the integrated value function $V(x)$).

First, we introduce the indirect expected utility of a decision maker facing systematic utility w_y for alternative y

$$\mathcal{G}(w) = \mathbb{E}\left[\max_{y \in \mathcal{Y}} (w_y + \epsilon_y)\right]$$

which is called the "social surplus function" in McFadden's (1978) random utility framework, and can be interpreted as the expected welfare of a representative agent in the dynamic

discrete-choice problem. Letting Y^* denote the (random) optimal alternative, we can also write

$$\mathcal{G}(w) = \mathbb{E}\left[w_{Y^*} + \epsilon_{Y^*}\right] = \sum_{y \in \mathcal{Y}} P(Y^* = y)\left(w_y + \mathbb{E}[\epsilon_y | Y^* = y]\right) \tag{3}$$

where the argument of $\mathcal{G}$ is a $|\mathcal{Y}|$-dimensional vector; called *choice specific value functions* in the literature. Hence, if we compare this with the previous equation (2), we obtain

$$V(x) = \mathcal{G}(w_{\cdot}(x)), \text{ where} \tag{4}$$

$$w_y(x) \equiv \bar{u}(y, x) + \beta \mathbb{E}\left[V\left(X'\right) | X = x, Y = y\right]. \tag{5}$$

$\mathcal{G}$ is a convex function. We define $\mathcal{G}^*$, the Legendre-Fenchel conjugate function of $\mathcal{G}$, by

$$\mathcal{G}^*(p) = \sup_{w \in \mathbb{R}^{\mathcal{Y}}} \left\{ \sum_{y \in \mathcal{Y}} p_y w_y - \mathcal{G}(w) \right\} \tag{6}$$

if p is a probability over the set $\mathcal{Y}$, that is $p_y \geq 0$ and $\sum_{y \in \mathcal{Y}} p_y = 1$, and $\mathcal{G}^*(p) = +\infty$ otherwise. From combining Eqs. (3), (4), and (6), we see that the convex conjugate function corresponds to

$$\mathcal{G}^*(p) = -\sum_y P(Y^* = y)\mathbb{E}[\epsilon_y | Y^* = y], \tag{7}$$

the weighted conditional expectations of the utility shocks ϵ_y conditional on choosing the option y.

2.3. **Duality between choice probabilities and choice-specific value functions.** The following result will be the basic of our identification strategy. The subdifferential $\partial\varphi$ of a convex functions is defined in the Appendix.

Proposition 1. *The following pair of equivalent statements identify $w_{\cdot}(x)$:*
(i) p is in the subdifferential of $\mathcal{G}$ at w

$$p(.|x) \in \partial\mathcal{G}(w_{\cdot}(x)), \tag{8}$$

(ii) w is in the subdifferential of $\mathcal{G}^*$ at p

$$w_{\cdot}(x) \in \partial\mathcal{G}^*(p(.|x)). \tag{9}$$

In this sense, the observed choice probabilities p and the unobserved choice-specific value functions w are in a duality relationship. The first part of the proposition above corresponds to the Williams-Daly-Zachary theorem, which is analogous to Roy's Identity in discrete choice models (cf. McFadden (1981), Anderson, DePalma and Thisse (1992)). The second part, demonstrating a "reverse" mapping between conditional choice probabilities and choice-specific value functions, is related to, and perhaps a more general statement of, existing results in the literature (cf. Hotz and Miller (1993), Magnac and Thesmar (2002), Arcidiacono and Miller (2012, Lemma 5)).[1] Furthermore, for the true $w_{\cdot}(x)$, the integrated value function V is equal to

$$V(x) = \mathcal{G}(w_{\cdot}(x)) \tag{10}$$

and $\bar{u}(.,x)$ is given by

$$\bar{u}(y,x) = w_y(x) - \beta \mathbb{E}\left[V\left(X'\right)|X = x, Y = y\right]. \tag{11}$$

Equations (9), (10), and (11) above present the relations between the unobserved functions $\bar{u}(y,x)$, $V(x)$ and $w_y(x)$ and the observed choice probabilities $p(y|x)$, as well as the functions $\mathcal{G}$ and $\mathcal{G}^*$, which in principle are computable given distributional assumptions regarding the random utility shocks ϵ_y. In that sense, these equations summarize the empirical content of the dynamic discrete-choice model. Proposition 1 describes the structure of random utility discrete-choice models, while Eq. (11) presents the recursive restrictions of the dynamic discrete-choice model. These equations echo, in perhaps the most general form, analogous derivations in the existing papers on identification and estimation of dynamic discrete-choice models, including Hotz and Miller (1993), Magnac and Thesmar (2002) and Bajari, Chernozhukov, Hong, and Nekipelov (2007).

[1]Clearly, Proposition (1) also applies to static random utility discrete-choice models, with the $w_{\cdot}(x)$ being interpreted as the utility indices obtained from each of the choices. As such, this proposition (esp. part (ii)) is also related to results regarding the mapping between choice probabilities and utilities in static discrete choice models (e.g. Berry (1994); Haile, Hortacsu, and Kosenok (2008)). Similar results have also arisen in the literature on stochastic learning in games (Hofbauer and Sandholm (2002); Cominetti, Melo and Sorin (2010)).

It is important to note that the relation defined by Equation (9) is a multi-valued correspondence. This arises out of two issues:

- *Indeterminacy of the choice-specific value functions*: if $w_{\cdot}(x)$ satisfies (8), then $w_{\cdot}(x) - K(x)$ also satisfies (8). Indeed, the choice probabilities are only affected by the differences in the levels offered by the various alternatives. Although this issue is standard in discrete choice theory, and is addressed by a proper choice of normalization, it is worthwile noting that in the present setting, the problem is complicated by the fact that one cannot impose normalization directly on elements of $\partial\mathcal{G}^{*}(p(.|x))$, as these depend on the (endogenous) quantity V; instead $\bar{u}$ (a primitive of the model) should be normalized. This is addressed in section 3 below.
- A (proper) *partial identification issue*: as will be clear below, if the distribution of the utility shocks ϵ is not absolutely continuous, the set $\partial\mathcal{G}^{*}(p(.|x))$ may be larger than as set of the form $\{w_{\cdot}(x) - K(x)\}$ for a fixed w. In this case the choice-specific value functions are partially identified in a proper sense, and even imposing a normalizing condition on $\bar{u}$, multiple candidates for $\bar{u}$ will identify the model.

We impose the following assumption on the distribution of ϵ.

Assumption 2. *The distribution of ϵ has full support.*

As shown in Theorem 1 below, this assumption addresses the partial identification issue, and we tackle the indeterminacy issue by isolating a particular $w_{\cdot}^{0}(x)$ among those satisfying Equation (9), which we do by imposing

$$\mathcal{G}\left(w_{\cdot}^{0}(x)\right) = 0. \tag{12}$$

We show that when the distribution of the unobservable heterogeneity ϵ has full support, this defines $w_{\cdot}^{0}$ unambiguously.

The utility vector $w_{\cdot}^{0}(x)$ which satisfies Eq. (12) need *not* satisfy Eqs. (10) or (11). However, as expressed in our next result, all $w_{\cdot}(x)$ satisfying (8) are of the form $w_{\cdot}^{0}(x) - K(x)$, for a vector of (state-dependent) constants $K(x)$; hence, the "true" $w_{\cdot}(x)$ – that which satisfies all the Eqs. (8), (10) and (11) – will differ from $w_{\cdot}^{0}(x)$ by a constant term.

Theorem 1. *Under Assumption 2, the following holds:*

(i) There exists a unique $w_{.}^{0}(x) \in \partial\mathcal{G}^{*}\left(p\left(.|x\right)\right)$ *such that* $\mathcal{G}\left(w_{.}^{0}(x)\right) = 0$.

(ii) $w_{.}(x) \in \partial\mathcal{G}^{*}\left(p\left(.|x\right)\right)$ *if and only if there exists* $K\left(x\right)$ *such that* $w_{.}(x) = w_{.}^{0}(x) - K\left(x\right)$.

But before proceeding to estimation, we discuss the example of the logit model, for which the functions and relations above reduce to familiar expressions.

Example 1 (Logit)**.** *As is classical, when the distribution* Q *of* ϵ *obeys an extreme value type I distribution, it follows from Extreme Value theory that* $\mathcal{G}$ *and* $\mathcal{G}^{*}$ *can be obtained in closed form :*

$$\begin{aligned}\mathcal{G}\left(w\right) &= \log(\sum_{y\in\mathcal{Y}}\exp(w_{y})) + \gamma \\ \mathcal{G}^{*}\left(p\right) &= \sum_{y\in\mathcal{Y}} p_{y}\log p_{y} - \gamma,\end{aligned}$$

where $\gamma \approx 0.57$ *(Euler's constant). Hence in this case,* $\mathcal{G}^{*}$ *is the entropy of distribution* p*. As a result,*

$$w_{.}^{0}(x) = \log p\left(y|x\right)$$

and $w_{.}(x) \in \partial\mathcal{G}^{*}\left(p\left(.|x\right)\right)$ *if and only if* $w_{y}(x) = w_{y}^{0}(x) - K\left(x\right)$.

This is (partially) by McFadden's theory of Generalized Extreme Value (GEV): when $\mathbf{F}$*, the cumulative distribution function of the vector of utility shocks* $(\epsilon_{i})_{1\leq i\leq n}$ *is such that*

$$g\left(x_{1},...,x_{n}\right) = -\log\mathbf{F}\left(-\log x_{1},...,-\log x_{n}\right)$$

is positive homogeneous of degree 1*, then* $\mathcal{G}\left(w\right)$ *exists in closed form:*[2]

$$\begin{aligned}\mathcal{G}\left(w\right) &= \log\left(-\log\mathbf{F}\left(-w_{1},...,-w_{n}\right)\right) + \gamma. \\ &= \log g\left(e^{w_{1}},...,e^{w_{n}}\right) + \gamma\end{aligned}$$

■

[2]However, this does not mean that $\mathcal{G}^{*}\left(p\right)$ can be found in closed form. But $\nabla\mathcal{G}\left(w\right)$ is found in closed form, so computation of $\mathcal{G}^{*}$ from Eq. (6) using gradient descent is very efficient. Relatedly, Arcidiacono and Miller (2011, pp. 1839-1841) discuss computational and analytical solutions for the $\mathcal{G}^{*}$ function in the generalized extreme value setting.

3. Estimation procedure

Based upon the derivations in the previous section, we present an estimation procedure, which follows two-steps. From the key Eq. (11) above, we know that identification and estimation of the model boil down to evaluating the unknown functions $\mathcal{G}$ and $\mathcal{G}^*$. We propose an estimation algorithm which only requires computing the $\mathcal{G}^*$ function.

3.1. **First step.** In the first step, we use the convex analysis to recover the vector of choice-specific value functions $w_{\cdot}^0(x) \in \partial\mathcal{G}^*(p(\cdot|x))$ at each vector of observed choice probabilities $p(\cdot|x)$ for each value of x. Using the following proposition, from Galichon and Salanié (2012), Proposition 2, we can characterize the $\mathcal{G}^*$ function completely in terms of the distribution function Q_ϵ of the utility shocks.

Proposition 2 (Galichon and Salanié (2012))**.** *Under Assumption 2, let $(p_y)_{y\in\mathcal{Y}}$ be a vector of choice probabilities. Then the function $\mathcal{G}^*(p)$ is the value of the mass transportation problem in which the distribution Q of utility shocks $\{\epsilon_y\}_{y\in\mathcal{Y}}$ is matched optimally to the distribution of actions y given by the multinomial distribution p, when the cost associated to a match of (ϵ, y) is given by*

$$c(y,\epsilon) = -\epsilon_y$$

where ϵ_y is the utility shock from taking the y-th action. That is,

$$\mathcal{G}^*(p) = \sup_{w(y)+z(e)\leq c(y,e)} \{\mathbb{E}_p[w(Y)] + \mathbb{E}_Q[z(\epsilon)]\} \tag{13}$$

which, by the Monge-Kantorovich duality, coincides with its dual

$$\mathcal{G}^*(p) = \min_{\substack{Y\sim p\\ \epsilon\sim Q}} \mathbb{E}[c(Y,\epsilon)] \tag{14}$$

and $w \in \partial\mathcal{G}^*(p)$ if and only if there exists g such that (w, g) is solution to (13).

The proof follows from the Monge-Kantorovich duality for mass transportation problems, which we include in the Appendix for the sake of completeness.

It follows from Proposition 2 that the problem of identification of $w_{\cdot}^0(x)$ can be recast as an optimal transportation problem, or an assignment game (Shapley and Shubik (1971)).

Even though $w_{\cdot}^{0}(x)$ cannot be obtained in closed form in general, a number of numerical methods allow for its efficient computation. We consider these later in Section 4.

3.2. **Second step.** In the second step, we use the recursive structure of the dynamic model (encapsulated in Eq. (5)), along with a normalization on the per-period utility functions $\bar{u}(y,x)$, to pin down the values of $w_{\cdot}(x)$, related to $w_{\cdot}^{0}(x)$ by $w_{\cdot}(x) = w_{\cdot}^{0}(x) - K(x)$, where the constant will be determined.[3]

Specifically, from the first step, we have obtained estimates of the multi-valued function $\mathcal{G}^{*}(p)$. We assume that accordingly, by knowledge of that function, we are also able to obtain one element $w_{\cdot}^{0}(x) \in \partial G^{*}(p)$ satisfying (12) which will not, however, in general coincide with $w_{\cdot}(x)$, the vector of choice-specific value functions which are of interest. Instead, given the earlier discussion, the two will differ by a constant:

$$w_{\cdot}(x) = w_{\cdot}^{0}(x) - K(x). \tag{15}$$

In the second step, we will exploit the structure of the dynamic optimization problem, as well as a normalization on the per-period utility functions $\bar{u}(y,x)$ in order to determine the value of K, and hence $\bar{u}(y,x)$.

From (12) and (15), it follows that $\mathcal{G}(w_{\cdot}(x)) = -K(x)$; by (10), we get $K(x) = -V(x)$, so that

$$w_{\cdot}(x) = w_{\cdot}^{0}(x) + V(x).$$

Hence, by (11), we get

$$\bar{u}(y,x) = w_{y}^{0}(x) + V(x) - \beta \mathbb{E}\left[V\left(X'\right) | X = x, Y = y\right].$$

In order to nonparametrically identify $\bar{u}(y,x)$, we need to impose a normalization. Following Bajari, Chernozhukov, Hong, and Nekipelov (2009), we set:

[3]Note that an alternative assumption could have been that $w_0(\cdot) = 0$, which corresponds to the observation in the dynamic discrete choice literature (cf. Hotz and Miller (1993), Magnac and Thesmar (2002)) concerning the mapping between the vector of choice probabilities $\{p(\cdot|x)\}_y$ and the vector of choice-specific value function *differences* $\{w_{\cdot}(x) - w_0(x)\}_y$. We choose the normalization $\mathcal{G}(w_{\cdot}(x)) = 0$ because it simplifies the second stage of our estimation procedure, as we will show below.

Assumption 3. $\forall x, \quad \bar{u}\left(y_0, x\right) = 0.$

With this assumption, we get

$$0 = w_{y_0}^0(x) + V\left(x\right) - \beta \mathbb{E}\left[V\left(X'\right)|X = x, Y = y_0\right]. \tag{16}$$

Let W be the column vector whose general term is $\left(w_{y_0}^0(x)\right)_{x \in \mathcal{X}}$, let V be the column vector whose general term is $\left(V\left(x\right)\right)_{x \in \mathcal{X}}$, and let Π be the $|\mathcal{X}| \times |\mathcal{X}|$ matrix whose general term $\Pi_{xx'}$ is $p\left(x'|x, y_0\right)$. Equation (16), rewritten in matrix notation, is

$$W = \beta \Pi V - V$$

and for $\beta < 1$, matrix $I - \beta\Pi$ is a diagonally dominant matrix. Hence, it is invertible and Equation (16) becomes

$$V = (\beta\Pi - I)^{-1} W. \tag{17}$$

The right hand side of this equation is uniquely estimated from the data. After obtaining $V(x)$, $\bar{u}(y, x)$ can be nonparametrically identified by

$$\bar{u}(y, x) = w_y^0(x) + V\left(x\right) - \beta \mathbb{E}[V(X')|X = x, Y = y], \tag{18}$$

where $w_{\cdot}^0\left(x\right)$ is as in Theorem 1, and V is given by (17).

As a sanity check, one recovers $\bar{u}(y_0, .) = W + V - \beta\Pi V = 0$. Also, when $\beta \to 0$, one recovers $\bar{u}(y, x) = w_y^0(x) - w_{y_0}^0(x)$ which is the case in standard static discrete choice.

Eqs. (17) and (18) above, showing how the per-period utilities can be recovered from the choice-specific value functions via a system of linear equations, echoes similar derivations in the existing literature (e.g. Aguirregabiria and Mira (2007), Pesendorfer and Schmidt-Dengler (2008), Arcidiacono and Miller (2011, 2013)). Hence, the innovative aspect of our estimator lies not in the second step, but rather in the first step, in which we show how the choice-specific value functions can be recovered for any assumed distribution of the utility shocks $(\epsilon_0, \ldots, \epsilon_K)$ conditional on X. In the next section, we focus on computational aspects of this first step.

4. Computation Details for First Step

It follows from Section 3.1 that the problem of identification of w can be formulated as an optimal transportation problem. In this section we shall investigate two methods with various degrees of generality to solve this problem and compute $w^0(x)$.

4.1. **Linear programming approach.** If (in contradiction with Assumption 2, which is otherwise maintained throughout the paper) Q were discrete, and if its support were $\epsilon^1, ...\epsilon^S$, letting $q_s = Q(\epsilon = \epsilon^s)$, Problem (13)-(14) has a Linear Programming formulation as

$$\min_{\pi \geq 0} \sum_{y,s} \pi_{ys} c_{ys} \tag{19}$$

$$\sum_{s=1}^{S} \pi_{ys} = p_y, \ \forall y \in \mathcal{Y} \tag{20}$$

$$\sum_{y \in \mathcal{Y}} \pi_{ys} = q_s, \ \forall s \in \{1, ..., S\}. \tag{21}$$

where $c_{ys} = -\epsilon_y^s$. In this case, the set of $w \in \partial \mathcal{G}^*(p)$ is the set of vector (w_y) of Lagrange multipliers corresponding to constraints (20). In this case, because the support of ϵ is discrete, w_y^0 will generally not be unique. This is due to the non-uniqueness of the solution to the dual of the LP problem in Eq. (19), and follows from Shapley and Shubik's (1971) well-known results on the multiplicity of the core in the finite assignment game. Applied to discrete-choice models, it implies that when the support of the utility shocks is finite, the utilities from the discrete-choice model will only be partially identified, an issue which will be addressed in the companion paper (Chiong, Galichon and Shum (2013)).

In the context of Assumption 2, one may discretize Q and solve the discretized problem. Specifically, we can simulate $\mathcal{G}^*$ at a given vector p by drawing S vectors of $\epsilon \sim Q_\epsilon$, and then solving the corresponding linear programming problem. Letting ϵ^s denote vectors drawn from Q_ϵ, each with weight $q_s \equiv 1/S$, this discretized problem is (19), which has dual

formulation

$$\max_{\lambda,z} \sum_{y\in\mathcal{Y}} p_y\lambda_y + \sum_{s=1}^{S} q_s z_s \tag{22}$$
$$s.t.\ \lambda_y + z_s \leq c_{ys}$$

Consider (λ_y, z_s) a dual solution to (22). It is well known that λ_y can be interpreted as a Lagrange multiplier associated to constraint (20), and z_s, as the Lagrange multiplier associated to constraint (21).[4] Also, one has $\mathcal{G}^*(p) = \sum_{y\in\mathcal{Y}} p_y\lambda_y + \sum_{s=1}^{S} q_s z_s$, which implies that $\mathcal{G}(\lambda) = -\sum_{s=1}^{S} q_s z_s$. Hence, to recover the vector $w_\cdot^0(x)$ satisfying $\mathcal{G}(w_\cdot^0(x)) = 0$, one can set

$$w_y^0 = \lambda_y - \mathcal{G}(\lambda) = \lambda_y + \sum_{s=1}^{S} q_s z_s.$$

This quantity converges to the true value of w_y^0 when S is large enough.[5]

In Appendix C, we present an alternative approach to computing the $\mathcal{G}^*$ function, based on "power diagrams".

4.2. **Remarks.** Our estimation procedure is distinctive in several ways. The estimation procedures proposed in much of the literature on identification and estimation of dynamic models require one of the two following steps. First, some procedures require "inverting" the mapping between choice probabilities and choice-specific value functions (eg. Hotz and Miller (1993), Magnac and Thesmar (2002)). This requires knowledge of the social surplus $\mathcal{G}$ function (because, by Proposition 1, the choice probability functions are just the gradient of the $\mathcal{G}$ function).[6] Second, existing procedures also rely on a small class of distributions for the utility shocks – primarily those in the extreme-value family, as in Example 1 above –

[4]Because the two linear programs (19) and (22) are dual to each other, the Lagrange multipliers of interest λ_y can be obtained by computing either program. In practice, for the simulations and empirical application below, we computed the primal problem (19).

[5]As we remarked before, when Q is finite, then the set of Lagrange multipliers is only partially identified, in a "proper" fashion. In light of the discussion in this section, this corresponds to the well-known property of the Shapley-Shubik assignment game that its core is not unique. However, as Gretsky, Ostroy, and Zame (1999, section 6) show, the core of a large finite markets is "approximately" a singleton.

[6]This remark is also relevant for static discrete choice models. In fact, the random-coefficients multinomial demand model of Berry, Levinsohn, and Pakes (1995) does not have a closed-form expression for the choice

because these distribution yield an analytical (or near-analytical) mapping between choice probabilities and $\{\mathbb{E}[\epsilon_y|y]\}_y$, the vector of conditional expectation of the utility shocks for the optimal choices, which is required in order to recover the per-period utility functions (eg. Hotz and Miller (1993), Hotz, Miller, Sanders, Smith (1994), Aguirregabiria and Mira (2007), Pesendorfer and Schmidt-Dengler (2008), Arcidiacono and Miller (2011)).

Our approach, however, works for any assumed distribution for the utility shocks;[7] it is based on a characterization of $\mathcal{G}^*$ which is amenable to simulation and which easily accommodates different choices for Q, the (joint) distribution of the utility shocks $(\epsilon_0, \ldots, \epsilon_K)$ conditional on X. Therefore, our findings expand the set of dynamic discrete-choice models suitable for applied work far beyond those with extreme-value distributed utility shocks.

5. Monte Carlo Experiment

In this section, we illustrate our estimation framework using Rust (1987) model of bus engine replacement. Harold Zurcher, a bus manager decides in each time period t whether to replace ($y_t = 1$) or maintain ($y_t = 0$) the engines of each bus in the company's fleet. The econometrician observes the cumulative mileage since last replacement of a bus engine at each time t, which we denote by $x_t \in X$, where $X = \{0, 1, \ldots, 89\}$. When $y_t = 0$, the change in mileage $(x_{t+1} - x_t)$ follows a multinomial distribution on $\{0, 1, 2\}$ with parameters $\pi = (\pi_0, \pi_1, \pi_2)$.[8] When the engine is replaced at t, the Markovian transition probability is $\Pr(x_{t+1} = i|x_t, y_t = 1) = \pi_i$, for $i = 0, 1, 2$.

The per-period payoff from choosing $y_t = 0$ is $u(y_t = 0, x_t) = -\theta x_t + \epsilon_{t0}$. This is interpreted as the per-period cost of operating the bus whose accumulated engine mileage since last replacement is x_t. On the other hand, the per-period payoff from choosing $y_t = 1$, and replacing the engine, is given by $u(y_t = 1, x_t) = -RC + \epsilon_{t1}$, where RC is the cost of replacing the engine. $(\epsilon_{t0}, \epsilon_{t1}) \in \mathbb{R}^2$ are the unobserved state variables. For this Monte

probabilities, thus necessitating a simulation-based inversion procedure. In section D of the Appendix, we will consider the random-coefficients demand model as an additional application of our estimation procedure.

[7]However, see Norets and Tang (2013) for another approach to estimation in dynamic binary choice models in which the choice probability function is not required to be known.

[8]Since the support of x_t is finite, when $x_t = 88$, $(x_{t+1} - x_t)$ follows a multinomial distribution on $\{0, 1\}$ with parameters $\pi = (\pi_0, 1 - \pi_0)$. When $x_t = 89$, $x_{t+1} = 89$ with probability 1.

Carlo study, we will assume that $\epsilon_{t0} - \epsilon_{t1} \sim N(0,1)$. Following Section 4, $\mathcal{G}^*$ is computed approximately at a vector of conditional choice probabilities using the linear programming approach in Equation 22 - by drawing S vectors from $N(0,1)$. We set $S = 1000$.

The parameters we fix and hold constant for the Monte Carlo study are $\theta = 0.0394$, $RC = 9.7558$, $(\pi_0, \pi_1, \pi_2) = (0.3489, 0.6394, 0.0117)$ and $\beta = 0.99$. These parameter values correspond to Rust's estimates for group 1,2,3 and 4, except that we increase θ by 15 times in order to decrease the number of initial states with zero probability of replacement.

Using these parameters, we solved the corresponding dynamic programming problem and obtained the true values of $w_0(x_t) = -\theta x_t + \beta\,\mathbb{E}[V(x_{t+1})|x_t, y_t = 0]$ and $w_1(x_t) = -RC + \beta\,\mathbb{E}[V(x_{t+1})|x_t, y_t = 1]$. We then determine the actual conditional choice probability (CCP): the probability of replacement at state x_t is $\Pr(y_t = 1|x_t) = \Pr[\epsilon_1 - \epsilon_0 \geq w_0(x_t) - w_1(x_t)]$.

Figure 1 shows the asymptotic performance of our estimation procedure. That is, abstracting from the sampling error in determining the CCPs and the transition probabilities, we first apply our estimation procedure to the vector of true CCPs, and recover the deterministic per-period utility function $\bar{u}(y_t = 0, x_t)$ (assuming we know (π_0, π_1, π_2), β and the distribution of $\epsilon_1 - \epsilon_0$, but w e know nothing about the per-period utilities). We can see from Figure 1 for the result that identification fails when the conditional choice probability is 0 or 1. Beyond the initial states where the probability of engine replacement is zero, the estimated per-period utilities match the true form of the utility function up to a normalization on $\bar{u}(y = 1, x)$.

To test the performance of our estimation procedure under small sample size, we generate simulated panel data of the following form: Data $= \{y_{it}\ , x_{it} : i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T\}$ where $y_{it} = 0$ at x_{it} if and only if $w_0(x_t) - w_1(x_t) \geq \epsilon_1 - \epsilon_0$, where $\epsilon_1 - \epsilon_0$ is independently drawn from $N(0,1)$. We vary the number of buses N and the number periods T, and for each combination of (N, T), we generate 1000 independent datasets. For each dataset, the deterministic per-period utilities $\bar{u}(y = 0, x)$ for $x \in X$ are computed. We then estimate the slope of $\bar{u}(y = 0, x)$ using weighted[9] ordinary least squares and compare it to the true $\theta = 0.0394$. For each dataset, we restrict to the states where there is at least

[9]We use the total number of observations in each state as the weight

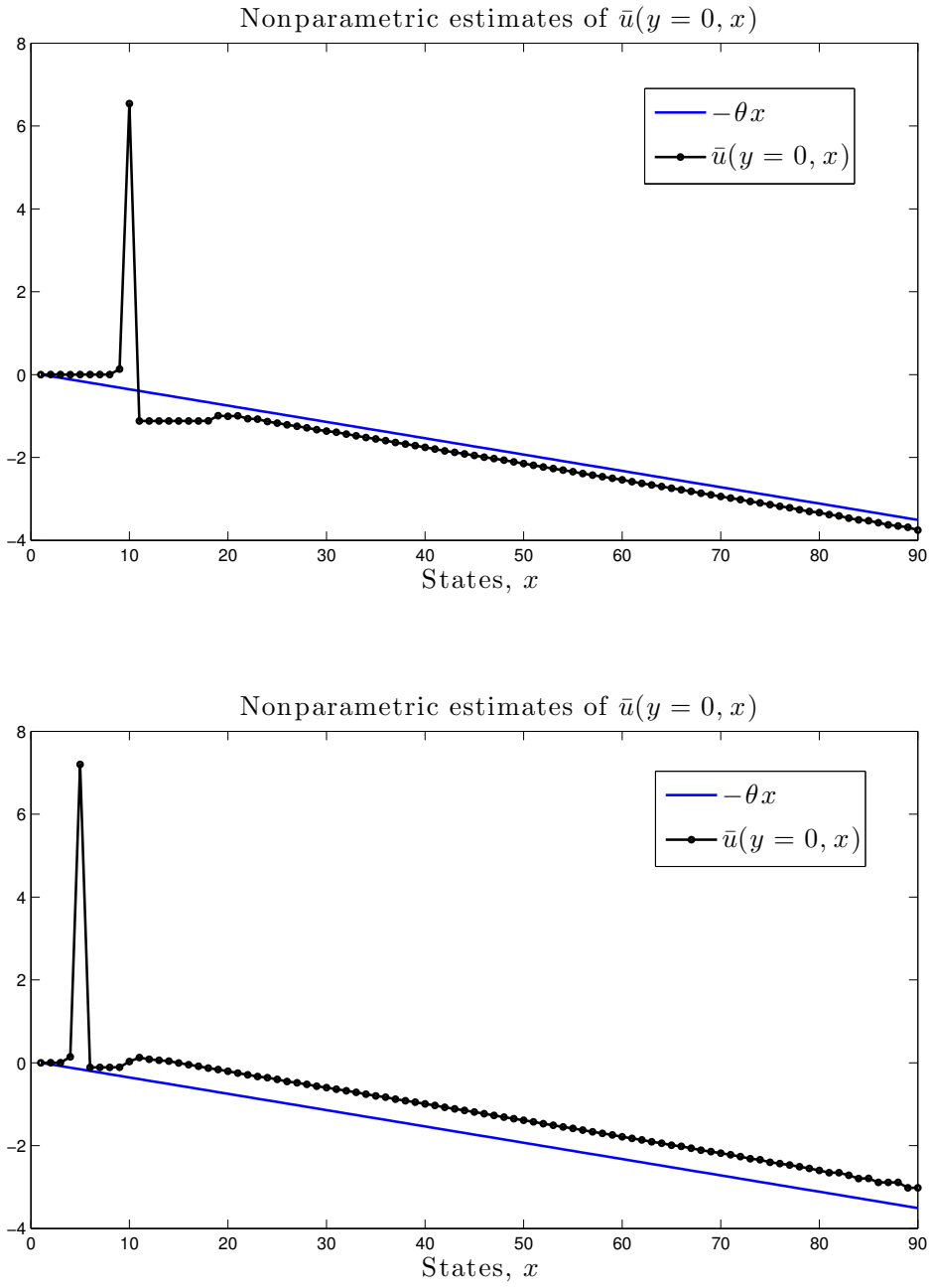


FIGURE 1. Top: $\beta = 0.90$, Bottom: $\beta = 0.99$. Asymptotic behavior of the estimated $\bar{u}(y = 0, x)$. Identification fails when the conditional choice probability is 0 or 1. Beyond the initial states where the probability of engine replacement is zero, the estimated per-period utilities match the true form of the utility function up to a normalization on $\bar{u}(y = 1, x)$.

Design	Mean	Median	Standard deviation	RMSE
$N = 100, T = 30$	0.0327	0.0500	0.0963	0.0965
$N = 100, T = 60$	0.0306	0.0434	0.0485	0.0493
$N = 100, T = 120$	0.0303	0.0390	0.0322	0.0334
$N = 200, T = 30$	0.0286	0.0422	0.0626	0.0635
$N = 200, T = 60$	0.0289	0.0388	0.0360	0.0375
$N = 200, T = 120$	0.0374	0.0382	0.0121	0.0122
$N = 500, T = 30$	0.0255	0.0362	0.0410	0.0432
$N = 500, T = 60$	0.0369	0.0381	0.0108	0.0111
$N = 500, T = 120$	0.0374	0.0377	0.0039	0.0044

TABLE 1. Each row reports the mean, median, SD and RMSE of the estimator $\hat{\theta}$. The true value is $\theta = 0.0394$. For small T, our estimator appears to be biased downward. To illustrate the nature of the bias, a histogram for the design $N = 200, T = 60$ is plotted in Table 4 in Appendix E. Conditional on the cost being positive, the bias disappears.

one observed replacement, since we know that identification fails for those states where the CCP is 1 or 0. The result of the Monte Carlo is reported in Table 1.

6. Empirical Application: Revisiting Harold Zurcher

In this section, we compare parameter estimation from Rust (1987) and our non-parametric procedure using Rust's data on the buses from Group 1 to 4. Using 10 years of monthly data on bus mileage and engine replacement decision for a fleet of 104 buses, Rust (1987) concluded that a linear or a square-root cost function best explains the data. Using the linear specification, the estimated parameters imply that the bus manager, Harold Zurcher, perceives average monthly maintenance costs to increase \$2.17 for every 5,000 accumulated miles on the bus.[10]

[10]Rust (1987) arrived at a figure of \$3.75 for bus groups 1, 2 and 3. We calculated using bus groups 1, 2, 3 and 4.

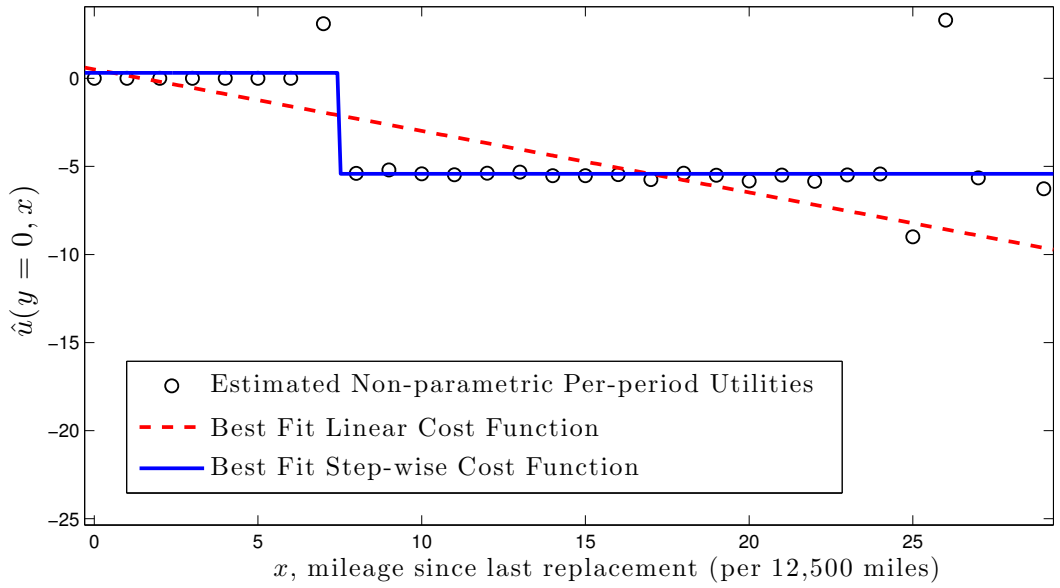
Rust (1987) divided the mileage space into 90 states, each representing a 5,000 increment in mileage since last engine replacement. However, there is only a total of 61 instances of replacements in the entire dataset, which necessarily implies that at least 28 states have zero probability of observing replacement. In the dataset, an engine is replaced approximately once every 10 years, and using monthly observation results in many zero-probability cells. Non-parametric identification is not possible when the vector of conditional choice probability lies on the boundary of the simplex. Typically, one approximates by setting $\hat{p}(y|x) = \epsilon$ when $p(y|x) = 0$, for some small $\epsilon > 0$. We take the view that when replacement is so infrequent, using such a fine grid size introduces substantial noise and errors in the recovered non-parametric utilities. Hence, we used 10 years of quarterly observations of bus mileage and engine replacement decision, and discretized the mileage space into coarser intervals of 12,500 miles.

The states space is now $X = \{0, 1, \dots 29\}$. The first step of the estimation procedure consists of estimating the vector of conditional choice probabilities (CCP) directly from the data set. A vector of CCPs is defined by $p = (p_0, \dots, p_{29})'$, $p_i = (\Pr(y_t = 0|x_t = i), \Pr(y_t = 1|x_t = i))$ for $i = 0, \dots 29$. Also directly obtained from the data in the first step is the Markov transition probabilities for the observed state variable $x_t \in X$, estimated to be of the following:

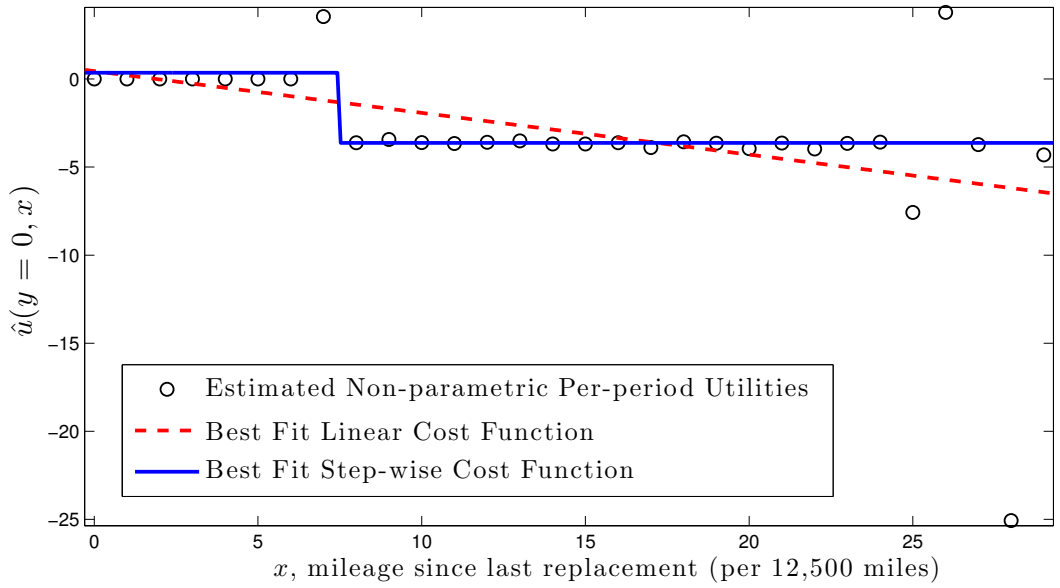
$$\hat{\Pr}(x_{t+1} = j|x_t = i, y_t = 0) = \begin{cases} 0.7405 & \text{if} \quad j = i \\ 0.2595 & \text{if} \quad j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\Pr}(x_{t+1} = j|x_t = i, y_t = 1) = \begin{cases} 0.7405 & \text{if} \quad j = 0 \\ 0.2595 & \text{if} \quad j = 1 \\ 0 & \text{otherwise} \end{cases}$$

The second step of the estimation procedure consists of postulating the distribution of the unobserved state variables, which is sufficient for us to compute ∂G^* at each of $\{\hat{p}_0, \dots, \hat{p}_{29}\}$. For this analysis, we assumed that $\epsilon_{t0} - \epsilon_{t1} \sim N(0, 1)$.

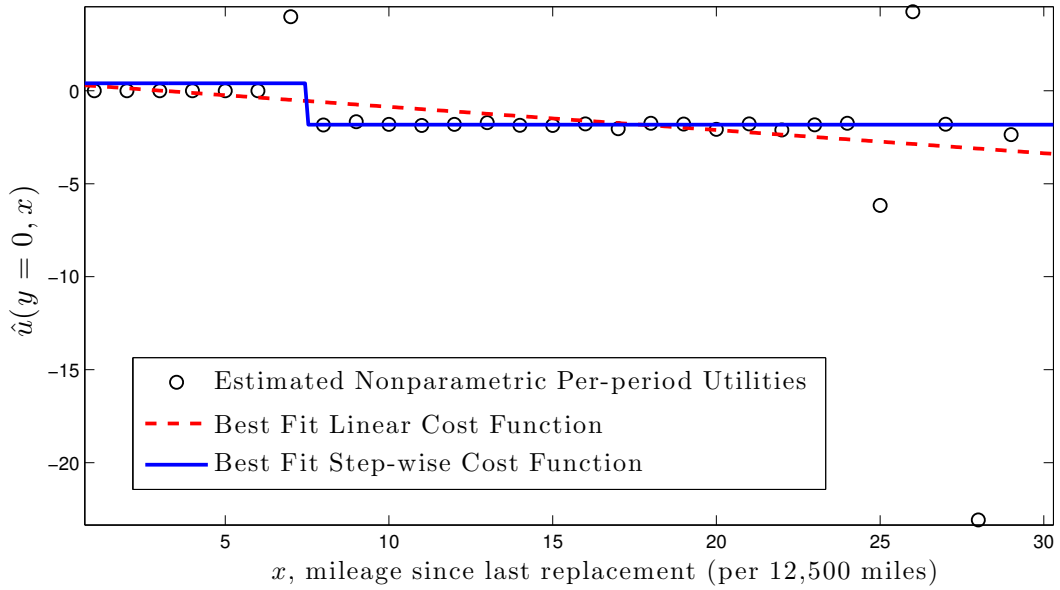


(A) $\beta = 0.7$



(B) $\beta = 0.8$

FIGURE 2



(c) $\beta = 0.9$

FIGURE 2. Point estimates of the per-period utilities when engine is not replaced at each $x \in X$. We fitted the linear and step functions to the points, weighing each point by the number of observations at that state.

To non-parametrically estimate $\bar{u}(y = 0, x)$, we normalized $\bar{u}(y = 1, x)$ to $-RC$ for all $x \in X$. That is, the per-period utility from choosing action $y = 1$ (replacing bus engine) is just minus of the replacement cost. If we further impose that $u(y = 0, x = 0) = 0$, we can estimate RC by $w_1 - w_0$ where $w = (w_0, w_1)$ is any w in $\partial G^*(\hat{p}_0)$.

In Figure 2, we show the result of the estimation for $\beta = 0.7, 0.8, 0.9$. At first glance, the estimated cost function has a step-like shape. We first fitted a linear function to the estimated utilities, and obtained a negative slope that is statistically significant. Qualitatively, this is in agreement with Rust (1987). However, we also fitted a Heaviside step function of the form $aH(x - b) + c$, where $H(x) = 0$ for $x < 0$, $H(x) = 0.5$ for $x = 0$ and $H(x) = 1$ for $x > 0$.

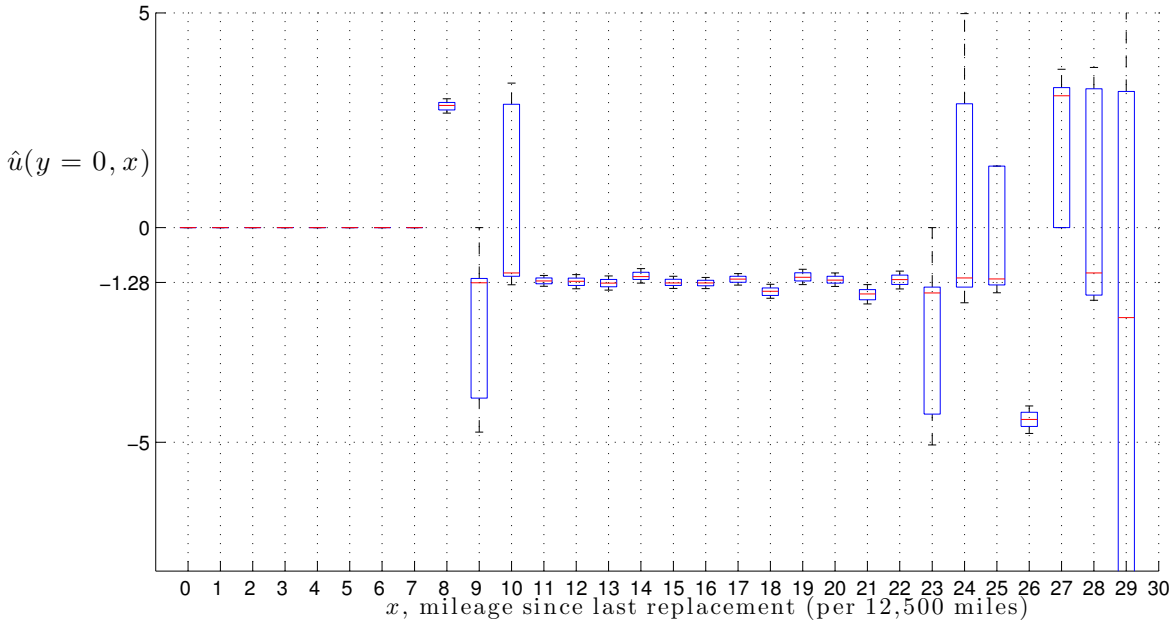


FIGURE 3. Bootstrapped estimates when $\beta = 0.9$

For a large range of β we have considered, the step function fits the estimated utilities much better than the linear function. Table in the appendix shows that at $\beta = 0.9$, the R-squared for fitting the step function is 0.503, while the R-squared for the linear function is 0.272.

Surprisingly, the kink in the cost function occurs between $x = 8$ and $x = 9$, or between 100,000 miles and 112,500 miles. In another words, as soon as the bus has accumulated more than 100,000 miles, Harold Zurcher behaves as if that bus is \$300 per month more expensive to maintain than a bus which has accumulated fewer than 100,000 miles (using $\beta = 0.9$). Harold Zurcher perceives the average quarterly maintenance cost to plateau out when the mileage is above the rule-of-thumb cutoff point of 100,000 miles. It is worth noting that Rust (1987) mentioned that: "According to Zurcher, monthly maintenance costs increase very slowly as a function of accumulated mileage." Our analysis does not identify what happens to the cost function before the cutoff point of 100,000 miles because no engine replacement was observed at those states in the dataset.

7. Conclusions

In this paper, we have shown how results from convex analysis can be fruitfully applied to study identification in dynamic discrete choice models; modulo the use of these tools, a large class of dynamic discrete choice problems with quite general utility shocks becomes no more difficult to compute and estimate than the Logit model encountered in most empirical applications. This has allowed us to provide a natural and holistic framework encompassing the papers of Rust (1987), Hotz and Miller (1993), and Magnac and Thesmar (2002). While the identification results in this paper are comparable to other results in the literature, the convex analysis approach appears new. Far more than providing a mere reformulation, this approach is powerful, and has significant implications in several dimensions:

First, by drawing the (surprising) connection between the computation of the $\mathcal{G}^*$ function and the computation of optimal matchings in the classical assignment game (Shapley and Shubik (1971)), we can apply the powerful tools developed to compute optimal matchings to dynamic discrete-choice models. While the present paper has used standard Linear Programming algorithms such as the Simplex algorithm, other, more powerful matching algorithm such as the Hungarian algorithm may be efficiently put to use when the dimensionality of the problem grows.

Moreover, by reformulating the problem as an optimal matching problem, all existence and uniqueness results are inherited from the theory of optimal transportation. For instance, the uniqueness of a systematic utility rationalizing the consumer's choices follows from the uniqueness of a potential in the Monge-Kantorovich theorem. In addition, the structure of the problem is inherited from the structure of the solution of matching games, which has a lattice structure. Our companion paper Chiong, Galichon and Shum (2013) makes use of this structure in a partial identification setting.

We believe the present paper opens a more flexible way to deal with discrete choice models. While identification is exact for a fixed structure of the unobserved heterogeneity, one may wish to parameterize the distribution of the utility shocks and do inference on that parameter. The results and methods developed in this paper may also extend to dynamic discrete games, with the utility shocks reinterpreted as players' private information (see,

e.g. Aguirregabiria and Mira (2007) or Pesendorfer and Schmidt-Dengler (2008)). These directions, however, we leave for future exploration.

References

[1] V. Aguirregabiria and P. Mira. Sequential estimation of dynamic discrete games. *Econometrica*, 75:1–53, 2007.

[2] S. Anderson, A. DePalma and J. Thisse. *Discrete Choice Theory of Product Differentiation*. MIT Press, 1992.

[3] P. Arcidiacono and R. Miller. Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity. *Econometrica*, 79: 1823-1867, 2011.

[4] P. Arcidiacono and R. Miller. Identifying Dynamic Discrete Choice Models off Short Panels. Working paper, 2013.

[5] F. Aurenhammer. Power diagrams: properties, algorithms, and applications. *SIAM Journal on Computing*, 16(1):78-96, 1987.

[6] C. Aliprantis and K. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, 2006.

[7] P. Bajari, V. Chernozhukov, H. Hong, and D. Nekipelov. Nonparametric and semiparametric analysis of a dynamic game model. Preprint, 2009.

[8] S. Berry. Estimating Discrete-Choice models of Production Differentiation. *RAND Journal of Economics*, 25:242-262, 1994.

[9] S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, 63:841–890, July 1995.

[10] S. Berry and A. Pakes. The pure characteristics demand model. *International Economic Review*, 48:1193–1225, 2007.

[11] K. Chiong, A. Galichon, and M. Shum. The Structure of Partial Identification in Discrete Choice Models. Preprint, 2013.

[12] R. Cominetti, E. Melo, and S. Sorin. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70:71-83, 2010.

[13] A. Galichon and B. Salanié. Cupid's invisible hand: Social surplus and identification in matching models. Preprint, 2012.

[14] N. Gretsky, J. Ostroy, and W. Zame. Perfect Competition in the Continuous Assignment Model. *Journal of Economic Theory*, Vol. 85, pp. 60-118, 1999.

[15] P. Haile, A. Hortacsu, and G. Kosenok. On the Empirical Content of Quantal Response Models. *American Economic Review*, 98:180-200, 2008.

[16] J. Hofbauer and W. Sandholm. On the Global Convergence of Stochastic Fictitious Play. *Econometrica*, 70: 2265-2294, 2002.

[17] J. Hotz and R. Miller. Conditional choice probabilties and the estimation of dynamic models. *Review of Economic Studies*, 60:497–529, 1993.

[18] J. Hotz, R. Miller, S. Sanders, and J. Smith. A Simulation Estimator for Dynamic Models of Discrete Choice. *Review of Economic Studies*, 61:265-289, 1994.

[19] T. Magnac and D. Thesmar. Identifying dynamic discrete decision processes. *Econometrica*, 70:801–816, 2002.

[20] D. McFadden. Modelling the choice of residential location. In A. Karlquist et. al., editor, *Spatial Interaction Theory and Residential Location*. North Holland Pub. Co., 1978.

[21] D. McFadden. Economic Models of Probabilistic Choice. In C. Manski and D. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*, 1981.

[22] A. Norets and S. Takahashi. On the Surjectivity of the Mapping Between Utilities and Choice Probabilities. *Quantitative Economics* 4.1 (2013): 149-155.

[23] A. Norets and X. Tang. Semiparametric Inference in Dynamic Binary Choice Models. Preprint, Princeton University, 2013.

[24] M. Pesendorfer and P. Schmidt-Dengler. Asymptotic least squares estimators for dynamic games. *Review of Economic Studies*, 75:901–928, 2008.

[25] R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[26] J. Rust. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55:999–1033, 1987.

[27] L. Shapley and M. Shubik. The assignment game I: The core. *International Journal of Game Theory*, 1(1):111–130, 1971.

Appendix A. Background results

A.1. **Convex analysis for Discrete-choice Models.** Let $u \in \mathbb{R}^{|\mathcal{Y}|}$ be a closed convex set. Given a set of utility shocks $\{\epsilon_y\}_{y\in\mathcal{Y}}$ distributed according to a joint distribution function Q_ϵ, we define the social surplus function as $\mathcal{G}(u) = \mathbb{E}[\max_y \{u_y + \epsilon_y\}]$, where u_y is the yth component of u. If $\mathbb{E}(\epsilon_y)$ exists and is finite, then the function $\mathcal{G}$ is a proper convex function that is continuous everywhere. Moreover assuming that Q_ϵ is sufficiently well-behaved, $\mathcal{G}$ is differentiable everywhere.

Define the Fenchel conjugate of $\mathcal{G}$ as $\mathcal{G}^*(p) = \sup_{u\in\mathbb{R}^{|\mathcal{Y}|}}\{p \cdot u - \mathcal{G}(u)\}$. The domain of $\mathcal{G}^*$ consists of $p \in \mathbb{R}^{|\mathcal{Y}|}$ for which the supremum is finite. Lemma 1 below shows that $\Delta^{|\mathcal{Y}|}$ is always contained in the domain of $\mathcal{G}^*$. It follows that $\mathcal{G}^*$ is a continuous convex function on its domain since it is the pointwise supremum of the family of affine functions (parameterized by y) $f_y : p \mapsto y \cdot p - \mathcal{G}(y)$. (A closed proper convex function f is the pointwise supremum of the collection of all affine functions h such that $h \leqq f$, see Aliprantis and Border (2006))

Lemma 1. *For any p in the simplex $\Delta^{|\mathcal{Y}|}$, there exists $u^* \in \mathbb{R}^{|\mathcal{Y}|}$ such that $p = \nabla\mathcal{G}(u^*)$. In particular, this implies that $\mathcal{G}^*(p) = \sup_{u\in\mathbb{R}^{|\mathcal{Y}|}}\{p \cdot u - \mathcal{G}(u)\}$ is well-defined and finite for all $p \in \Delta^{|\mathcal{Y}|}$.*

Proof. The proof is a direct implication of Norets and Takahashi (2013). □

We say that u is a subgradient of $\mathcal{G}^*$ at p if $\mathcal{G}^*(p)+u\cdot(p'-p) \leq \mathcal{G}^*(p')$ for all $p' \in \mathbf{dom}\,\mathcal{G}^*$. The notation $u \in \partial\mathcal{G}^*(p)$ is used to convey exactly this. We now show how the subdifferential $\partial\mathcal{G}^*$ is related to $\mathcal{G}$.

Lemma 2. *u is a subgradient of $\mathcal{G}^*$ at p if and only if $p \cdot u - \mathcal{G}(u) \geq p \cdot u' - \mathcal{G}(u')$ for all $u' \in \mathbb{R}^{|\mathcal{Y}|}$. That is, $\partial\mathcal{G}^*(p) = \operatorname{argmax}_{u\in\mathbb{R}^{|\mathcal{Y}|}}\{p \cdot u - \mathcal{G}(u)\}$.*

Proof. First, we show that $p\cdot u-\mathcal{G}(u) \geq p\cdot u'-\mathcal{G}(u')$ for all $u' \implies \mathcal{G}^*(p)+u\cdot(p'-p) \leq \mathcal{G}^*(p')$ for all p' in the domain of $\mathcal{G}^*$.

$$\begin{aligned} & p \cdot u - \mathcal{G}(u) \geq p \cdot u' - \mathcal{G}(u') \quad \forall u' \\ \Longrightarrow \quad & p \cdot u - \mathcal{G}(u) = \sup_{u'} \{p \cdot u' - \mathcal{G}(u')\} \\ \Longrightarrow \quad & p \cdot u - \mathcal{G}(u) = \mathcal{G}^*(p) \end{aligned}$$

Now take any p' in the domain of $\mathcal{G}^*$

$$\begin{aligned} \mathcal{G}^*(p) + u \cdot (p' - p) &= [p \cdot u - \mathcal{G}(u)] + u \cdot (p' - p) \\ \mathcal{G}^*(p) + u \cdot (p' - p) &= p' \cdot u - \mathcal{G}(u) \\ &\leq \sup_{\tilde{u}} \{p' \cdot \tilde{u} - \mathcal{G}(\tilde{u})\} \\ &= \mathcal{G}^*(p') \end{aligned}$$

Conversely, we show that if $u \in \partial \mathcal{G}^*(p)$, then $u \in \text{argmax}_{u'} \{p \cdot u' - \mathcal{G}(u')\}$.

$$\begin{aligned} u \in \partial \mathcal{G}^*(p) \quad \Longrightarrow \quad & \mathcal{G}^*(p') \geq \mathcal{G}^*(p) + u \cdot (p' - p), \ \forall p' \in \mathbf{dom} \mathcal{G}^* \\ \Longrightarrow \quad & u \cdot p - \mathcal{G}^*(p) \geq u \cdot p' - \mathcal{G}^*(p') \\ \Longrightarrow \quad & u \cdot p - \mathcal{G}^*(p) = \sup_{\tilde{p}} \{u \cdot \tilde{p} - \mathcal{G}^*(\tilde{p})\} \\ \Longrightarrow \quad & u \cdot p - \mathcal{G}^*(p) = \mathcal{G}(u) \\ \Longrightarrow \quad & p \cdot u - \mathcal{G}(u) = \sup_{u'} \{p \cdot u' - \mathcal{G}(u')\} \end{aligned}$$

In the second to last line, we have applied the fact if $\mathcal{G}$ is a closed convex function, then the conjugate of the conjugate of $\mathcal{G}$ is itself (cf. Rockafellar (1970, pg. 104). That is, $\mathcal{G}(u) = \sup_{\tilde{p}} \{\tilde{p} \cdot u - \mathcal{G}^*(\tilde{p})\}$. □

Note that Lemma 1 above applies only to discrete-choice models, but Lemma 2 is a general result for any proper closed convex function $\mathcal{G}$. The remaining lemmas in this section are specialized results for discrete-choice models.

Lemma 3. *The function $\mathcal{G}^*$ is not differentiable at any $p \in \Delta^{|\mathcal{Y}|}$.*

Proof. Consider the function $f : \mathbb{R}^{|\mathcal{Y}|} \mapsto \mathbb{R}$ given by $f(u) = p \cdot u - \mathcal{G}(u)$. f is a concave and differentiable function. From Lemma 1, there exists a $u^* \in \mathbb{R}$ such that $p = \nabla\mathcal{G}(u^*)$. This implies that $f'(u) = 0$ has a solution, i.e. $f'(u^*) = 0$. Now, consider adding a constant k to every component of u^*, that is, consider $u^* + k\mathbf{1}$ where $\mathbf{1} \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector of ones.

$$\begin{aligned} f(u^* + k\mathbf{1}) &= p \cdot u^* + k - \mathcal{G}(u^* + k\mathbf{1}) \\ &= p \cdot u^* + k - \mathcal{G}(u^*) - k \\ &= f(u^*) \end{aligned}$$

In the first equality, we have used the fact that $\sum_{y \in \mathcal{Y}} p_y = 1$ since p is in the simplex $\Delta^{|\mathcal{Y}|}$. The last line above shows that $\{u^*, u^* + k\mathbf{1}\}$ belongs to $\text{argmax}\{p \cdot u - \mathcal{G}(u)\}$. By Lemma 2, $u^* + k\mathbf{1}$ and u^* are both subgradient of $\mathcal{G}^*$ at p. By Theorem 25.1 of Rockafellar (1972), a convex function $\mathcal{G}^*$ is differentiable at p if and only if the subdifferential $\partial\mathcal{G}^*(p)$ is a singleton. Since $\{u^*, u^* + k\mathbf{1}\} \in \partial\mathcal{G}^*(p)$, we conclude that $\mathcal{G}^*$ is not differentiable at any p in the simplex $\Delta^{|\mathcal{Y}|}$.

In general, when p is not in the simplex, the $\mathcal{G}^*$ function is usually differentiable. We know this because a convex function is differentiable almost everywhere, and the set of points in $\mathbb{R}^{|\mathcal{Y}|}$ where $\mathcal{G}^*$ is not differentiable has Lebesgue measure zero.

□

We can further characterize the set $\partial\mathcal{G}^*$. For convenience, we will consider the case where $\mathcal{Y} = \{1, 2\}$. The same reasoning applies to $|\mathcal{Y}| > 2$.

Lemma 4. *For any $p = (p_1, p_2)$ such that $0 < p_1, p_2 < 1$, the subdifferential $\partial\mathcal{G}^*(p)$ is a line described by $\{(u_1 + \alpha, u_2 + \alpha) \in \mathcal{U} : \alpha \in \mathbb{R}\}$, where $u = (u_1, u_2)$ is some $u \in \mathbb{R}^2$ satisfying $p = \nabla\mathcal{G}(u)$.*

Proof. Fix an arbitrary $p = (p_1, p_2)$ in the interior of the simplex Δ^2. It is straightforward to show from the definition of $\mathcal{G}$ that if $(p_1, p_2) \in \nabla\mathcal{G}(u_1, u_2)$, then $(p_1, p_2) \in \nabla\mathcal{G}(u_1+\alpha, u_2+\alpha)$ for any constant k. Since (u_1, u_2) and $(u_1+\alpha, u_2+\alpha)$ both maximize $f(u) = p \cdot u - \mathcal{G}(u)$, by Lemma 2 then, (u_1, u_2) and $(u_1+\alpha, u_2+\alpha)$ both belong to the subdifferential $\partial\mathcal{G}^*(p)$. Since the choice of $\alpha \in \mathbb{R}$ is arbitrary, we have then showed that $\{(u_1 + \alpha, u_2 + \alpha) \in \mathbb{R}^2 : \alpha \in \mathbb{R}\}$ is a subset of $\partial\mathcal{G}^*(p)$, where $u = (u_1, u_2)$ is any $u \in \mathbb{R}^2$ satisfying $p = \nabla\mathcal{G}(u)$.

It remains to show that any $u' \notin \{(u_1 + \alpha, u_2 + \alpha) \in \mathbb{R}^2 : \alpha \in \mathbb{R}\}$ does not maximize $f(u) = p \cdot u - \mathcal{G}(u)$, and by Lemma 2, is not a subgradient of $\mathcal{G}$ at p. Take any $u = (u_1, u_2)$ such that $p = \nabla(u)$. Without loss of generality, say that $u' = (u_1 + a, u_2 + b)$ for some $a, b \in \mathbb{R}$, where $a \neq b$. Then,

$$\begin{aligned} f(\tilde{u}) &= p \cdot \tilde{u} - \mathcal{G}(\tilde{u}) \\ f(u') &= p \cdot u + ap_1 + bp_2 - \mathcal{G}(u') \\ &< p \cdot u + ap_1 + bp_2 - [\mathcal{G}(u) + \nabla\mathcal{G}(u) \cdot (a, b)'] \qquad (23) \\ &= f(u) + ap_1 + bp_2 - \nabla\mathcal{G}(u) \cdot (a, b)' \\ &= f(u) \end{aligned}$$

In line 23, we applied the strict inequality that $\mathcal{G}(u') \leq \mathcal{G}(u) + \nabla\mathcal{G}(u) \cdot (u' - u)$, which holds true when $\mathcal{G}$ is a strictly convex function on the line segment joining u and u'. That is, we further need to show $\mathcal{G} : [u, u'] \mapsto \mathbb{R}$ is a strictly convex function, where $[u, u'] = \{(\lambda u + (1 - \lambda)u') : 0 \leq \lambda \leq 1\}$.

$$\begin{aligned}
& \lambda\mathcal{G}(u') + (1-\lambda)\mathcal{G}(u) \\
= \quad & \mathbb{E}[\max\{\lambda u_1 + \lambda a + \lambda\epsilon_1, \lambda u_2 + \lambda b + \lambda\epsilon_2\}] \\
& + \mathbb{E}[\max\{(1-\lambda)u_1 + (1-\lambda)\epsilon_1, (1-\lambda)u_2 + (1-\lambda)\epsilon_2\}] \\
> \quad & \mathbb{E}(\max\{u_1 + \lambda a + \epsilon_1, u_2 + \lambda b + \epsilon_2\}) \\
= \quad & \mathcal{G}\lambda u' + (1-\lambda)u)
\end{aligned}$$

Without loss of generality, assume that $b > a$. The above would hold true if the set of events such that $A = \{u_1 + a + \epsilon_1 \geq u_2 + b + \epsilon_2\} \cap \{u_2 + \epsilon_2 > u_1 + \epsilon_1\}$ occurs with strictly positive probability. Let Φ be the CDF of $\epsilon_2 - \epsilon_1$. Then, we need that $Pr(A) = \Phi(u_1 - u_2 + (a-b)) - \Phi(u_1 - u_2) > 0$. Holds when $\epsilon_2 - \epsilon_1$ has full support. □

Appendix B. Proofs

Proof of Proposition 1. This follows directly from Fenchel's inequality (see Rockafellar (1970), Theorem 23.5). □

Proof of Theorem 1. In this proof we shall drop x from the notation for the sake of clarity.

For a vector w we shall denote $Y(w, \varepsilon)$ be the value of y which maximizes $w_y + \varepsilon_y$.

Let $\tilde{w} \in \partial\mathcal{G}^*(p)$, and let $w_y = \tilde{w}_y - \mathcal{G}(\tilde{w})$. One has $\mathcal{G}(w) = 0$, and an immediate calculation shows that $\partial\mathcal{G}(w) = p$. Let us now show that w is unique. Consider w and w' such that $\mathcal{G}(w) = \mathcal{G}(w') = 0$, and $p \in \partial\mathcal{G}(w)$ and $p \in \partial\mathcal{G}(w')$. Assume $w \neq w'$ to get a contradiction; then there exist two distinct y_0 and y_1 such that $w_{y_0} - w_{y_1} \neq w'_{y_0} - w'_{y_1}$; without loss of generality one may assume

$$w_{y_0} - w_{y_1} > w'_{y_0} - w'_{y_1}.$$

Let S be the set of ε's such that

$$w_{y_0} - w_{y_1} > \varepsilon_{y_1} - \varepsilon_{y_0} > w'_{y_0} - w'_{y_1}$$

$$w_{y_0} + \varepsilon_{y_0} > \max_{y \neq y_0, y_1} w_y + \varepsilon_y$$

$$w'_{y_1} + \varepsilon_{y_1} > \max_{y \neq y_0, y_1} w'_y + \varepsilon_y$$

Because ε has full support, S has positive probability.

Let $\bar{w} = \frac{w+w'}{2}$. Because $p \in \partial \mathcal{G}(w)$ and $p \in \partial \mathcal{G}(w')$, one has $\mathcal{G}(\bar{w}) = 0$, thus

$$\begin{aligned} 0 &= \mathbb{E}\left[\bar{w}_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}\right] = \frac{1}{2}\mathbb{E}\left[w_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}\right] + \frac{1}{2}\mathbb{E}\left[w'_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[w_{Y(w,\varepsilon)} + \varepsilon_{Y(w,\varepsilon)}\right] + \frac{1}{2}\mathbb{E}\left[w'_{Y(w',\varepsilon)} + \varepsilon_{Y(w',\varepsilon)}\right] \\ &= \frac{1}{2}\left(\mathcal{G}(w) + \mathcal{G}\left(w'\right)\right) = 0 \end{aligned}$$

Hence equality holds term by term, and

$$\begin{aligned} w_{Y(w,\varepsilon)} + \varepsilon_{Y(w,\varepsilon)} &= w_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)} \\ w'_{Y(w',\varepsilon)} + \varepsilon_{Y(w',\varepsilon)} &= w'_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)} \end{aligned}$$

For $\varepsilon \in S$, $Y(w,\varepsilon) = Y(\bar{w},\varepsilon) = y_0$ and $Y(w',\varepsilon) = Y(\bar{w},\varepsilon) = y_1$, and we get the desired contradiction.

Hence $w = w'$, and the uniqueness of w follows. □

Proof of Proposition 2. Recall that

$$\begin{aligned} \mathcal{G}^*(p) &= \sup_{w \in \mathbb{R}^{\mathcal{Y}}} \left\{ \sum_y p_y w_y - \mathbb{E}_Q\left[\max_{y \in \mathcal{Y}} (w_y + \epsilon_y)\right] \right\} \\ &= \sup_{w \in \mathbb{R}^{\mathcal{Y}}} \left\{ \sum_y p_y w_y + \mathbb{E}_Q\left[\min_{y \in \mathcal{Y}} (-w_y - \epsilon_y)\right] \right\} \end{aligned}$$

where Q is the distribution of ϵ. Hence, introducing

$$c(y,\epsilon) = -\epsilon_y,$$

one has

$$\mathcal{G}^*(p) = \sup_{w(y)+g(e) \leq c(y,e)} \left\{ \mathbb{E}_p\left[w(Y)\right] + \mathbb{E}_Q\left[g(\epsilon)\right] \right\} \tag{24}$$

which, by the Monge-Kantorovich duality, coincides with its dual

$$\mathcal{G}^*(p) = \min_{\substack{Y\sim p \\ \epsilon\sim Q}} \mathbb{E}\left[c\left(Y,\epsilon\right)\right] \tag{25}$$

where the second equality applies the Monge-Kantorovich duality result. ■ □

Appendix C. Power cells approach for computing $\mathcal{G}^*$ function

Second, we can give geometric insights for the locus of ϵ which lead to the choice of some given y. For this, we need to reinterpret the utility shock ϵ_y as a scalar product in a higher dimensional space – a classical trick. For $y \in \mathcal{Y}$, let $\iota_y \in \{0,1\}^{\mathcal{Y}}$ the vector such that $(\iota_y)_{y'} = 1(y = y')$. Introduce $S_{\mathcal{Y}} = \{\iota_y : y \in \mathcal{Y}\}$, which is nothing else than the canonical basis of $\mathbb{R}^{\mathcal{Y}}$. Denoting $\cdot$ the scalar product in $\mathbb{R}^{\mathcal{Y}}$, one has $\epsilon_y = \epsilon \cdot \iota_y$, and letting P be the distribution over $S_{\mathcal{Y}}$ which gives probability p_y to point ι_y, problem (14) rewrites as

$$\mathcal{G}^*(p) = -\max_{\substack{Z\sim P \\ \epsilon\sim Q_\epsilon}} \mathbb{E}\left[\epsilon \cdot Z\right].$$

Hence, $-\mathcal{G}^*(p)$ is the value of a Monge-Kantorovich problem with a quadratic surplus. This problem is very well studied, and by Brenier's theorem, there exists a convex map $V : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}$ such that the optimal coupling (Z, ϵ) is such that $Z \in \partial V(\epsilon)$. As a result, Y^* is defined in (1) is related to ϵ by

$$\iota_{Y^*} \in \partial V^w(\epsilon) \tag{26}$$

where V is a convex, piecewise linear function given by $V^w(\epsilon) = \max_{y\in\mathcal{Y}}\{\epsilon \cdot \iota_y - w_y\}$. Because V^w is a convex function, it is (Lebesgue-) almost everywhere differentiable, so if the distribution of ϵ is absolutely continuous, then $\nabla V^w(\epsilon)$ exists almost surely, and (26) rewrites as $\iota_{Y^*} = \nabla V^w(\epsilon)$.

Define C_y^w as the set of ϵ which lead to the choice of y, that is

$$C_y^w = \left\{\epsilon \in \mathbb{R}^{\mathcal{Y}} : \iota_y \in \partial V^w(\epsilon)\right\}.$$

C_y^w are closed convex polytopes which are called *Power Diagrams* in combinatorial geometry, see Aurenhammer (1987). The probability of choice of y is hence $Q\left(C_y^w\right)$, the mass assigned

by distribution Q to set C_y^w. Routines in combinatorial geometry provide the computation of the area of C_y^w. Note that

$$Q\left(C_y^w\right) = \int_{C_y^w} dQ$$

which can also be approximated using simulation techniques.

Once $Q\left(C_y^w\right)$ is computed, we can use the following result to obtain w^y:

Theorem 2. *Under Assumption 2, Problem (13)-(14) reformulates as*

$$\mathcal{G}^*(p) = -\min_{w \in \mathbb{R}^{\mathcal{Y}}} \sum_{y \in \mathcal{Y}} p_y w_y + Q\left(C_y^w\right). \tag{27}$$

Problem (27) is a convex optimization problem and can be solved using a gradient descent of the form

$$w_y^{t+1} = w_y^{t+1} + \delta\left(w_y - Q\left(C_y^{w^t}\right)\right).$$

Appendix D. Static discrete choice models

As another application of these results, we can consider the estimation of static discrete choice models, as in Berry, Levinsohn, and Pakes (1995). In particular, we consider a discrete choice model with random coefficients, in which the choice-specific characteristics interact with agent-specific random variables.

We consider a static discrete choice model in which there are a total of J products available in the market. The utility that household i obtains from consuming product j is given by:

$$U_{ij} = \underbrace{X_j'\beta + \xi_j - p_j'\alpha}_{\equiv \delta_j \text{ "mean utility"}} + p_j'\epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \forall j$$

where ϵ_i (which is invariant across all brands j) is household i's "random coefficient" on price. This version is called the "pure characteristics" model of Berry-Pakes (2004). ξ_j is product j's "unobserved product quality", which can be correlated with price p_j and is the source of endogeneity in the BLP demand model.

The identification and estimation of this model is based on the moment conditions

$$\mathbb{E}[\xi_j | Z_j] = 0$$

where Z_j are appropriate instruments. The aggregate choice probabilities, or *market shares*, for this model are

$$p_j = \int \mathbf{1}(U_{ij} > U_{ij'},\ j' \neq j) dF(\epsilon_i)$$

The convex analysis results above can be applied to this case. We can estimate the model using a variant of the "three-step" procedure:

(1) For each set of parameters $\theta = \{\alpha, \beta, \sigma\}$, compute $\mathcal{G}^*_\theta(p)$. This can be done again by using Proposition 1 above, and expressing $\mathcal{G}^*$ as the optimized objective for a mass transportation problem.

(2) The subgradient of the $\mathcal{G}^*$ function contains vectors of "mean utilities":

$$\frac{\partial \mathcal{G}^*_\theta(p)}{\partial p_j} = \delta_j(\theta), \quad j = 1, \ldots, J$$

These can be obtained from the computation of the $\mathcal{G}^*$ function.

(3) Once we have the mean utilities, we can compute sample moment conditions

$$\frac{1}{J} \sum_{j=1}^{J} [(\delta_j(\theta) - X_j'\beta + \alpha p_j) * Z_j].$$

We estimate θ by finding values to minimize these sample moment conditions.

Appendix E. Additional Figures from Monte Carlo

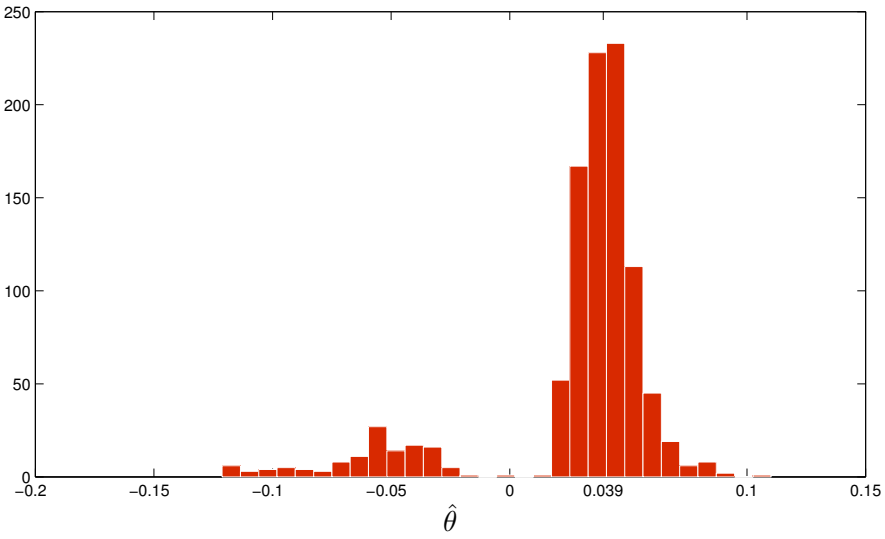


Figure 4. Histogram plot of $\hat{\theta}$, the estimates of the slope of $\bar{u}(y = 0, x)$, for the design $N = 200, T = 60$. The true value of θ is 0.0394 which coincides with the mode and median of the distribution as shown.

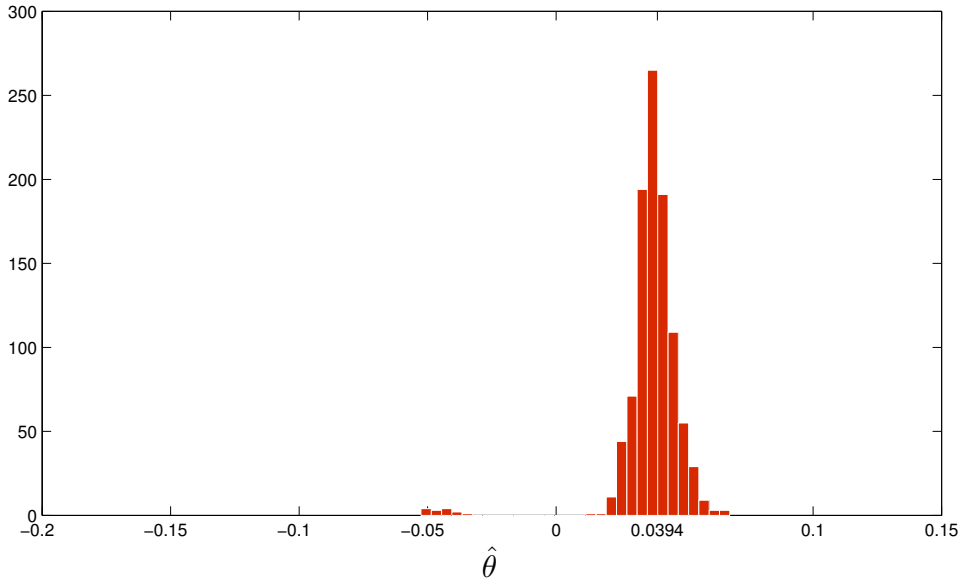


Figure 5. $N = 200, T = 120$.

Design	Mean	Median	Standard deviation	RMSE
$N = 100, T = 30$	0.0756	0.0641	0.0491	0.0610
$N = 100, T = 60$	0.0500	0.0471	0.0174	0.0203
$N = 100, T = 120$	0.0410	0.0404	0.0107	0.0108
$N = 200, T = 30$	0.0536	0.0482	0.0266	0.0301
$N = 200, T = 60$	0.0414	0.0403	0.0119	0.0121
$N = 200, T = 120$	0.0386	0.0383	0.0070	0.0071
$N = 500, T = 30$	0.0395	0.0379	0.0125	0.0125
$N = 500, T = 60$	0.0377	0.0382	0.0070	0.0072
$N = 500, T = 120$	0.0374	0.0377	0.0039	0.0044

TABLE 2. Mean, Median, SD and RMSE conditional on θ being non-negative.