

Duality in Dynamic Discrete Choice Models

Khai Chiong, Alfred Galichon, Matt Shum

▶ **To cite this version:**

Khai Chiong, Alfred Galichon, Matt Shum. Duality in Dynamic Discrete Choice Models. Quantitative Economics, 2015, Forthcoming. hal-03568184

HAL Id: hal-03568184
https://sciencespo.hal.science/hal-03568184

Submitted on 12 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DUALITY IN DYNAMIC DISCRETE CHOICE MODELS

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

Abstract. Using results from convex analysis, we investigate a novel approach to identification and estimation of discrete choice models which we call the "Mass Transport Approach" (MTA). We show that the conditional choice probabilities and the choice-specific payoffs in these models are related in the sense of *conjugate duality*, and that the identification problem is a mass transport problem. Based on this, we propose a new two-step estimator for these models; interestingly, the first step of our estimator involves solving a linear program which is identical to the classic assignment (two-sided matching) game of Shapley and Shubik (1971). The application of convex-analytic tools to dynamic discrete choice models, and the connection with two-sided matching models, is new in the literature.

Date: First draft: April 2013. This version: May 2015.

The authors thank the Editor, three anonymous referees, as well as Benjamin Connault, Thierry Magnac, Emerson Melo, Bob Miller, Sergio Montero, John Rust, Sorawoot (Tang) Srisuma, and Haiqing Xu for useful comments. We are especially grateful to Guillaume Carlier for providing decisive help with the proof of Theorem 5. We also thank audiences at Michigan, Northwestern, NYU, Pittsburgh, UCSD, the CEMMAP conference on inference in game-theoretic models (June 2013), UCLA econometrics mini-conference (June 2013), the Boston College Econometrics of Demand Conference (December 2013) and the Toulouse conference on "Recent Advances in Set Identification" (December 2013) for helpful comments. 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).

[§]Division of the Humanities and Social Sciences, California Institute of Technology; kchiong@caltech.edu

[†]Department of Economics, Sciences Po; alfred.galichon@sciences-po.fr

[♣]Division of the Humanities and Social Sciences, California Institute of Technology; mshum@caltech.edu.

1. Introduction

Empirical research utilizing dynamic discrete choice models of economic decision-making has flourished in recent decades, with applications in all areas of applied microeconomics including labor economics, industrial organization, public finance, and health economics. The existing literature on the identification and estimation of these models has recognized a close link between the conditional choice probabilities (hereafter, CCP, which can be observed and estimated from the data) and the payoffs (or *choice-specific value functions*, which are unobservable to the researcher); indeed, most estimation procedures contain an "inversion" step in which the choice-specific value functions are recovered given the estimated choice probabilities.

This paper has two contributions. First, we explicitly characterize this duality relationship between the choice probabilities and choice-specific payoffs. Specifically, in discrete choice models, the social surplus function (McFadden (1978)) provides us with the mapping from payoffs to the probabilities with which a choice is chosen at each state (conditional choice probabilities). Recognizing that the social surplus function is convex, we develop the idea that the convex conjugate of the social surplus function gives us the inverse mapping - from choice probabilities to utility indices. More precisely, the subdifferential of the convex conjugate is a correspondence that maps from the observed choice probabilities to an identified set of payoffs. In short, the choice probabilities and utility indices are related in the sense of *conjugate duality*. The discovery of this relationship allows us to succinctly characterize the empirical content of discrete choice models, both static and dynamic.

Not only is the convex conjugate of the social surplus function a useful theoretical object; it also provides a new and practical way to "invert" from a given vector of choice probabilities back to the underlying utility indices which generated these probabilities. This is the second contribution of this paper. We show how the conjugate along with its set of subgradients can be efficiently computed by means of linear programming. This linear programming formulation has the structure of an optimal assignment problem (as in Shapley-Shubik's (1971) classic work). This surprising connection enables us to apply insights developed in the optimal transport literature, e.g. Villani (2003, 2009), to discrete choice models. We call this new methodology the "Mass Transport Approach" to CCP inversion.

This paper focuses on the estimation of dynamic discrete-choice models via two-step estimation procedures in which conditional choice probabilities are estimated in the initial stage; this estimation approach was pioneered in Hotz and Miller (HM, 1993) and Hotz, Miller, Sanders, Smith (1994).[1] Our use of tools and concepts from convex analysis to study identification and estimation in this dynamic discrete choice setting is novel in the literature. Based on our findings, we propose a new two-step estimator for DDC models. A nice feature of our estimator is that it works for practically any assumed distribution of the utility shocks.[2] Thus, our estimator would make possible the task of evaluating the robustness of estimation to different distributional assumptions.[3]

Section 2 contains our main results regarding duality between choice probabilities and payoffs in discrete choice models. Based on these results, we propose, in Section 3, a two-step estimation approach for these models. We also emphasize here the surprising connection between dynamic discrete-choice and optimal matching models. In Section 4 we discuss computational details for our estimator, focusing on the use of linear programming to compute (approximately) the convex conjugate function from the dynamic discrete-choice model. Monte Carlo experiments (in Section 5) show that our estimator performs well in practice, and we apply the estimator to Rust's (1987) bus engine replacement data (Section 6). Section 7 concludes. The Appendix contains proofs and also a brief primer on relevant results from convex analysis. Sections 2.2 and 2.3, as well as Section 4, are not specific to dynamic discrete choice problems but are also true for any (static) discrete choice model.

[1] Subsequent contributions include Aguirregabiria and Mira (2002, 2007), Magnac and Thesmar (2002), Pesendorfer and Schmidt-Dengler (2008), Bajari, et. al. (2009), Arcidiacono and Miller (2011), and Norets and Tang (2013).

[2] While existing identification results for dynamic discrete choice models allow for quite general specifications of the additive choice-specific utility shocks, many applications of these two-step estimators maintain the restrictive assumption that the utility shocks are distributed i.i.d. type I extreme value, independently of the state variables, leading to choice probabilities which take the multinomial logit form.

[3] While they are not the focus in this paper, many applications of dynamic choice models do not utilize HM-type two step estimation procedures, and they allow for quite flexible distributions of the utility shocks, and also for serial correlation in these shocks (examples include Pakes (1986) and Keane and Wolpin (1997)). This literature typically employs simulated method of moments, or simulated maximum likelihood for estimation (see Rust (1994, section 3.3)).

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 $x \in \mathcal{X}$ which we assume to take only a finite number of values. Agents choose actions $y \in \mathcal{Y}$ from a finite space $\mathcal{Y} = \{0, 1, \ldots, D\}$. The single-period utility flow which an agent derives from choosing y in a given period is

$$\bar{u}_y(x) + \varepsilon_y$$

where ε_y denotes the utility shock pertaining to action y, which differs across agents. Across agents and time periods, the set of utility shocks $\varepsilon \equiv (\varepsilon_y)_{y \in \mathcal{Y}}$ is distributed according to a joint distribution function $Q(\cdots; x)$ which can depend on the current values of the state variable x. We assume that this distribution Q is known to the researcher.

Throughout, we consider a stationary setting in which the agent's decision environment remains unchanged across time periods; thus, for any given period, we use primes ($'$) to denote next-period values. 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[4]):

Assumption 1 (Conditional Independence)**.** (x, ε) *evolves across time periods as a controlled first-order Markov process, with transition*

$$\begin{aligned} Pr(x', \varepsilon' | y, x, \varepsilon) =& Pr(\varepsilon' | x', y, x, \varepsilon) \cdot Pr(x' | y, x, \varepsilon) \\ =& Pr(\varepsilon' | x') \cdot Pr(x' | y, x). \end{aligned}$$

The discount rate is β. Agents are dynamic optimizers whose choices each period satisfy[5]

$$y \in \arg\max_{\tilde{y} \in \mathcal{Y}} \left\{ \bar{u}_{\tilde{y}}(x) + \varepsilon_{\tilde{y}} + \beta \mathbb{E}\left[\bar{V}\left(x', \varepsilon'\right) | x, \tilde{y} \right] \right\}, \tag{1}$$

where the value function $\bar{V}$ is recursively defined via Bellman's equation as[6]

$$\bar{V}(x, \varepsilon) = \max_{\tilde{y} \in \mathcal{Y}} \left\{ \bar{u}_{\tilde{y}}(x) + \varepsilon_{\tilde{y}} + \beta \mathbb{E}\left[\bar{V}\left(x', \varepsilon'\right) | x, \tilde{y} \right] \right\}.$$

[4] See Norets (2009), Kasahara and Shimotsu (2009), Arcidiacono and Miller (2011), and Hu and Shum (2012).

[5] We have used Assumption 1 to eliminate ε as a conditioning variable in the expectation in Eq. (1).

[6] See, eg., Bertsekas (1987, chap. 5) for an introduction and derivation of this equation.

$V(x)$, the ex-ante value function, is defined as:[7]

$$V(x) = \mathbb{E}\left[\bar{V}(x, \varepsilon) | x\right].$$

The expectation above is conditional on the current state x. In the literature, $V(x)$ is called the ex-ante (or integrated) 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.

Next we define the *choice-specific value functions* as consisting of two terms: the per-period utility flow and the discounted continuation payoff:

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

In this paper, the utility flows $\{u_y(x); \forall y \in \mathcal{Y}, \forall x \in \mathcal{X}\}$, and subsequently also the choice-specific value functions $\{w_y(x), \forall y, x\}$, will be treated as unknown parameters; and we will study the identification and estimation of these parameters. For this reason, in the initial part of the paper, we will suppress the explicit dependence of w_y on x for convenience.

Given these preliminaries, we derive the duality which is central to this paper.

2.2. **The social surplus function and its convex conjugate.** We start by introducing the expected indirect utility of a decision maker facing the $|\mathcal{Y}|$-dimensional vector of choice-specific values $w \equiv \{w_y, y \in \mathcal{Y}\}'$:

$$\mathcal{G}(w; x) = \mathbb{E}\left[\max_{y \in \mathcal{Y}} (w_y + \varepsilon_y) | x\right] \tag{2}$$

where the expectation is assumed to be finite and is taken over the distribution of the utility shocks, $Q(\cdot; x)$. This function $\mathcal{G}(\cdot; x) : \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$, 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.

[7] There is a difference between the definition of $V(x)$ and the last terms in Equation (1) above. Here, we are considering the expectation of the value function $\bar{V}(x, \varepsilon)$ taken over the distribution of $\varepsilon|x$ (ie. holding the first argument fixed). In the last term of Eq. (1), however, we are considering the expectation over the *joint* distribution of $(x', \varepsilon')|x$ (ie. holding neither argument fixed).

For convenience in what follows, we introduce the notation $Y(w, \varepsilon)$ to denote an agent's optimal choice given the vector of choice-specific value functions w and the vector of utility shocks ε; that is, $Y(w, \varepsilon) = \text{argmax}_{y \in \mathcal{Y}}(w_y + \varepsilon_y)$.[8] This notation makes explicit the randomness in the optimal alternative (arising from the utility shocks ε). We get

$$\mathcal{G}(w; x) = \mathbb{E}\left[w_{Y(w,\varepsilon)} + \varepsilon_{Y(w,\varepsilon)} | x\right] = \sum_{y \in \mathcal{Y}} \underbrace{Pr(Y(w, \varepsilon) = y|x)}_{\equiv p_y(x)} (w_y + \mathbb{E}[\varepsilon_y | Y(w, \varepsilon) = y, x]) \quad (3)$$

which shows an alternative expression for the social surplus function as a weighted average, where the weights are the components of the vector of *conditional choice probabilities* $p(x)$. For the remainder of this section, we suppress the dependence of all quantities on x for convenience. In later sections, we will reintroduce this dependence when it is necessary.

In the case when the social surplus function $\mathcal{G}(w)$ is differentiable (which holds for most discrete-choice model specifications considered in the literature[9]), we obtain a well-known fact that the vector of choice probabilities p compatible with rational choice coincides with the gradient of $\mathcal{G}$ at w:

Proposition 1 (The Williams-Daly-Zachary (WDZ) Theorem)**.**

$$p = \nabla \mathcal{G}(w).$$

This result, which is analogous to Roy's Identity in discrete choice models, is expounded in McFadden (1978) and Rust (1994; Thm. 3.1)). It characterizes the vector of choice probabilities corresponding to optimal behavior in a discrete choice model as the gradient of the social surplus function. For completeness, we include a proof in the Appendix. The WDZ theorem provides a mapping from the choice-specific value functions (which are unobserved by researchers) to the observed choice probabilities p.

However, the identification problem is the reverse problem, namely to determine the set of w which would lead to a given vector of choice probabilities. This problem is exactly solved by convex duality and the introduction of the convex conjugate of $\mathcal{G}$, which we denote as $\mathcal{G}^*$:[10]

[8] We use w and ε (and also p below) to denote vectors, while w_y and ε_y (and p_y) denote the y-th component of these vectors.

[9] This includes logit, nested logit, multinomial probit, etc. in which the distribution of the utility shocks is absolutely continuous and w is bounded, cf. Lemma 1 in Shi, Shum and Wong (2014).

[10] Details of convex conjugates are expounded in the Appendix. Convex conjugates are also encountered in classic producer and consumer theory. For instance, when f is the convex cost function of the firm

Definition 1 (Convex Conjugate)**.** *We define $\mathcal{G}^*$, the Legendre-Fenchel conjugate function of $\mathcal{G}$ (a convex function), by*

$$\mathcal{G}^*(p) = \sup_{w \in \mathbb{R}^{\mathcal{Y}}} \left\{ \sum_{y \in \mathcal{Y}} p_y w_y - \mathcal{G}(w) \right\}. \tag{4}$$

Equation (4) above has the property that if p is not a probability, that is if either conditions $p_y \geq 0$ or $\sum_{y \in \mathcal{Y}} p_y = 1$ do not hold, then $\mathcal{G}^*(p) = +\infty$. Because the choice-specific value functions w and the choice probabilities p are, respectively, the arguments of the functions $\mathcal{G}$ and its convex conjugate function $\mathcal{G}^*$, we say that w and p are related in the sense of *conjugate duality*. The theorem below states an implication of this duality, and provides an "inverse" correspondence from the observed choice probabilities back to the unobserved w, which is a necessary step for identification and estimation.

Theorem 1. *The following pair of equivalent statements capture the empirical content of the DDC model:*
(i) p is in the subdifferential of $\mathcal{G}$ at w

$$p \in \partial \mathcal{G}(w), \tag{5}$$

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

$$w \in \partial \mathcal{G}^*(p). \tag{6}$$

The definition and properties of the subdifferential of a convex function are provided in Appendix A.[11] Part (i) is, of course, connected to the WDZ theorem above; indeed, it is the WDZ theorem when $\mathcal{G}(w)$ is differentiable at w. Hence, it encapsulates an optimality requirement that the vector of observed choice probabilities p be derived from optimal discrete-choice decision making for some unknown vector w of choice-specific value functions.

(decreasing returns to scale in production), then the convex conjugate of the cost function, f^*, is in fact the firm's optimal profit function.

[11] $\mathcal{G}$ is differentiable at w if and only if $\partial \mathcal{G}(w)$ is single-valued. In that case, part (i) of Th. 1 reduces to $p = \nabla \mathcal{G}(w)$, which is the WDZ theorem. If, in addition, $\nabla \mathcal{G}$ is one-to-one, then we immediately get $w = (\nabla \mathcal{G})^{-1}(p)$, or $\nabla \mathcal{G}^*(p) = (\nabla \mathcal{G})^{-1}(p)$, which is the case of the classical Legendre transform. However, as we show below, $\nabla \mathcal{G}(w)$ is not typically one-to-one in discrete choice models, so that the statement in part (ii) of Th. 1 is more suitable.

Part (ii) of this proposition, which describes the "inverse" mapping from conditional choice probabilities to choice-specific value functions, does not appear to have been exploited in the literature on dynamic discrete choice. It relates to Galichon and Salanié (2012) who use convex analysis to estimate matching games with transferable utilities. It specifically states that the vector of choice-specific value functions can be identified from the corresponding vector of observed choice probabilities p as the subgradient of the convex conjugate function $\mathcal{G}^*(p)$. Eq. (6) is also constructive, and suggests a procedure for computing the choice-specific value functions corresponding to observed choice probabilities. We will fully elaborate this procedure in subsequent sections[12].

Appendix A contains additional derivations related to the subgradient of a convex function. Specifically, it is known (Eq. (25)) that $\mathcal{G}(w) + \mathcal{G}^*(p) = \sum_{y\in\mathcal{Y}} p_y w_y$ if and only if $p \in \partial\mathcal{G}(w)$. Combining this with Eq. (3), we obtain an alternative expression for the convex conjugate function $\mathcal{G}^*$:

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

corresponding to the weighted expectations of the utility shocks ε_y conditional on choosing the option y. It is also known that the subdifferential $\partial\mathcal{G}^*(p)$ corresponds to the set of maximizers in the program (4) which define the conjugate function $\mathcal{G}^*(p)$; that is,

$$w \in \partial\mathcal{G}^*(p) \quad \Leftrightarrow \quad w \in \text{argmax}_{w\in\mathbb{R}^{\mathcal{Y}}} \left\{ \sum_{y\in\mathcal{Y}} p_y w_y - \mathcal{G}(w) \right\}. \tag{8}$$

Later, we will exploit this variational representation of the subdifferential $\mathcal{G}^*(p)$ for computational purposes; cf. Section 4 below.

Example 1 (Logit)**.** *Before proceeding, we discuss the logit model, for which the functions and relations above reduce to familiar expressions. 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*[13]*: $\mathcal{G}(w) = \log(\sum_{y\in\mathcal{Y}} \exp(w_y)) + \gamma$, while $\mathcal{G}^*(p) = \sum_{y\in\mathcal{Y}} p_y \log p_y - \gamma$*

[12] Clearly, Theorem 1 also applies to static random utility discrete-choice models, with the $w(x)$ being interpreted as the utility indices for each of the choices. As such, Eq. (6) relates to results regarding the invertibility of the mapping from utilities to choice probabilities in static discrete choice models (e.g. Berry (1994); Haile, Hortacsu, and Kosenok (2008); Berry, Gandhi, and Haile (2013)). Similar results have also arisen in the literature on stochastic learning in games (Hofbauer and Sandholm (2002); Cominetti, Melo and Sorin (2010)).

[13] Relatedly, Arcidiacono and Miller (2011, pp. 1839-1841) discuss computational and analytical solutions for the $\mathcal{G}^*$ function in the generalized extreme value setting.

if p belongs in the interior of the simplex, $\mathcal{G}^*(p) = +\infty$ otherwise ($\gamma \approx 0.57$ is Euler's constant). Hence in this case, $\mathcal{G}^*$ is the entropy of distribution p(see Anderson, de Palma, Thisse (1988) and references therein).

The subdifferential of $\mathcal{G}^*$ is characterized as follows: $w \in \partial\mathcal{G}^*(p)$ if and only if $w_y = \log p_y - K$, for some $K \in \mathbb{R}$. In this logit case the convex conjugate function $\mathcal{G}^*$ is the entropy of distribution p, which explains why it can be called a generalized entropy *function even in non-logit contexts.* ■

2.3. **Identification.** It follows from Theorem 1 that the identification of systematic utilities boils down to the problem of computing the subgradient of a generalized entropy function. However, from examining the social surplus function $\mathcal{G}$, we see that if $w \in \partial\mathcal{G}^*(p)$, then it is also true that $w - K \in \partial\mathcal{G}^*(p)$, where $K \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector taking values of K across all $\mathcal{Y}$ components. Indeed, the choice probabilities are only affected by the differences in the levels offered by the various alternatives. In what follows, we shall tackle this indeterminacy problem by isolating a particular w^0 among those satisfying $w \in \partial\mathcal{G}^*(p)$, where we choose

$$\mathcal{G}\left(w^0\right) = 0. \tag{9}$$

We will impose the following assumption on the heterogeneity.

Assumption 2 (Full Support)**.** *Assume the distribution Q of the vector of utility shocks ε is such that the distribution of the vector $(\varepsilon_y - \varepsilon_1)_{y \neq 1}$ has full support.*

Under this assumption, Theorem 2 below shows that Eq. (9) defines w^0 uniquely. Theorem 3 will then show that the knowledge of w^0 allows for easy recovery of all vectors w satisfying $p \in \partial\mathcal{G}(w)$.

Theorem 2. *Under Assumption 2, let p be in the interior of the simplex $\Delta^{|\mathcal{Y}|}$, (i.e. $p_y > 0$ for each y and $\sum_y p_y = 1$). Then there exists a unique $w^0 \in \partial\mathcal{G}^*(p)$ such that $\mathcal{G}\left(w^0\right) = 0$.*

The proof of this theorem is in the Appendix. Moreover, even when Assumption 2 is not satisfied, w^0 will still be set-identified; Theorem 4 below describes the identified set of w^0 corresponding to a given vector of choice probabilities p.

Our next result is our main tool for identification; it shows that our choice of $w^0(x)$, as defined in Eq. (9) is without loss of generality; it is not an additional model restriction, but

merely a convenient way of *representing* all $w(x)$ in $\partial\mathcal{G}^*(p(x))$ with respect to a natural and convenient reference point.[14]

Theorem 3. *Maintain Assumption 2, and let K denote any scalar $K \in \mathbb{R}$. The set of conditions*

$$w \in \partial\mathcal{G}^*(p) \text{ and } \mathcal{G}(w) = K$$

is equivalent to

$$w_y = w_y^0 + K, \ \forall y \in \mathcal{Y}.$$

This theorem shows that any vector within the set $\partial\mathcal{G}^*(p)$ can be characterized as the sum of the (uniquely-determined, by Theorem 3) vector w^0 and a constant $K \in \mathbb{R}$. As we will see below, this is our invertibility result for dynamic discrete choice problems, as it will imply unique identification of the vector of choice-specific value functions corresponding to any observed vector of conditional choice probabilities.[15]

2.4. **Empirical Content of Dynamic Discrete Choice Model.** To summarize the empirical content of the model, we recall the fact that the ex-ante value function V solves the following equation

$$V(x) = \sum_{y\in\mathcal{Y}} p_y(x) \left(\bar{u}_y(x) + \mathbb{E}[\varepsilon_y | Y(w,\varepsilon) = y, x] + \beta \sum_{x'} p\left(x'|x,y\right) V\left(x'\right) \right)$$

(derived in Pesendorfer and Schmidt-Dengler (2008), among others), where we write $p(x'|x,y) = Pr(x_{t+1} = x'|x_t = x, y_t = y)$. Noting that the choice-specific value function is just

$$w_y(x) = \bar{u}_y(x) + \beta \sum_{x'} p\left(x'|x,y\right) V\left(x'\right), \tag{10}$$

and, comparing with Eq. (3),

$$V(x) = \mathcal{G}(w(x);x) \text{ and } p(x) \in \partial\mathcal{G}(w(x);x).$$

[14] This indeterminacy issue has been resolved in the existing literature on dynamic discrete choice models (eg. Hotz and Miller (1993), Rust (1994), Magnac and Thesmar (2002) by focusing on the *differences* between choice-specific value functions, which is equivalent to setting $w_{y_0}(x)$, the choice-specific value function for a benchmark choice y_0, equal to zero. Compared to this, our choice of $w^0(x)$ satisfying $\mathcal{G}(w^0(x)) = 0$ is more convenient in our context, as it leads to a simple expression for the constant K (see Section 2.4).

[15] See Berry (1994), Chiappori and Komunjer (2010), Berry, Gandhi, and Haile (2012), among others, for conditions ensuring the invertibility or "univalence" of demand systems stemming from multinomial choice models, under settings more general than the random utility framework considered here.

Hence, by Theorem 3, the true $w(x)$ will differ from $w^0(x)$ by a constant term $V(x)$:

$$w(x) = w^0(x) + V(x)$$

where $w^0(x)$ is defined in Theorem 2. This result is also convenient for identification purposes, as it separates identification of w into two subproblems, the determination of w^0 and the determination of V. Once w^0 and V are known, the utility flows are determined from Eq. (10). This motivates our two-step estimation procedure, which we describe next.

3. Estimation using the Mass Transport Approach (MTA)

Based upon the derivations in the previous section, we present a two-step estimation procedure. In the first step, we use the results from Theorem 3 to recover the vector of choice-specific value functions $w^0(x)$ corresponding to each observed vector of choice probabilities $p(x)$. In the second step, we recover the utility flow functions $\bar{u}_y(x)$ given the $w^0(x)$ obtained from the first step.

3.1. **First step.** In the first step, the goal is to recover the vector of choice-specific value functions $w^0(x) \in \partial\mathcal{G}^*(p(x))$ corresponding to the vector of observed choice probabilities $p(x)$ for each value of x. In doing this, we use Theorem 1 above and Proposition 2 below, which show how $w^0(x)$ belongs to the subdifferential of the conjugate function $\mathcal{G}^*(p(x))$. We delay discussing these details until Section 4. There, we will show how this problem of obtaining $w^0(x)$ can be reformulated in terms of a class of mathematical programming problems, the Monge-Kantorovich *mass transport* problems, which leads to convenient computational procedures. Since this is the central component of our estimation procedure, we have named it the *mass transport approach* (MTA).

3.2. **Second step.** From the first step, we obtained $w^0(x)$ such that $w(x) = w^0(x) + V(x)$. Now in the second step, we use the recursive structure of the dynamic model, along with fixing one of the utility flows, to *jointly* pin down the values of $w(x)$ and $V(x)$. Finally, once $w(x)$ and $V(x)$ are known, the utility flows can be obtained from $\bar{u}_y(x) = w_y(x) - \beta\mathbb{E}[V(x')|x, y]$.

In order to nonparametrically identify $\bar{u}_y(x)$, we need to fix some values of the utility flows. Following Bajari, Chernozhukov, Hong, and Nekipelov (2009), we fix the utility flow corresponding to a benchmark choice y_0 to be constant at zero:[16]

Assumption 3 (Fix utility flow for benchmark choice)**.** $\forall x, \quad \bar{u}_{y_0}(x) = 0.$

With this assumption, we get

$$0 = w^0_{y_0}(x) + V(x) - \beta \mathbb{E}\left[V\left(x'\right)|x, y = y_0\right]. \tag{11}$$

Let W be the column vector whose general term is $\left(w^0_{y_0}(x)\right)_{x\in\mathcal{X}}$, let V be the column vector whose general term is $\left(V(x)\right)_{x\in\mathcal{X}}$, and let Π^0 be the $|\mathcal{X}| \times |\mathcal{X}|$ matrix whose general term Π^0_{ij} is $Pr\left(x_{t+1} = j | x_t = i, y = y_0\right)$. Equation (11), rewritten in matrix notation, is

$$W = \beta \Pi^0 V - V$$

and for $\beta < 1$, matrix $I - \beta\Pi^0$ is a diagonally dominant matrix. Hence, it is invertible and Equation (11) becomes

$$V = (\beta\Pi^0 - I)^{-1} W. \tag{12}$$

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

$$\bar{u}_y(x) = w^0_y(x) + V(x) - \beta\mathbb{E}[V(x')|x, y], \tag{13}$$

where $w^0(x)$ is as in Theorem 3, and V is given by (12).

As a sanity check, one recovers $\bar{u}_{y_0}(.) = W + V - \beta\Pi^0 V = 0$. Also, when $\beta \to 0$, one recovers $\bar{u}_y(x) = w^0_y(x) - w^0_{y_0}(x)$ which is the case in standard static discrete choice. Moreover, since our approach to identifying the utility flows is nonparametric, our MTA

[16] In a static discrete-choice setting (i.e. $\beta = 0$), this assumption would be a normalization, and without loss of generality. In a dynamic discrete-choice setting, however, this entails some loss of generality because different values for the utility flows imply different values for the choice-specific value functions, which leads to differences in the optimal choice behavior. Norets and Tang (2013) discuss this issue in greater detail.

approach does not leverage any known restrictions on the flow utility (including parametric or shape restrictions) in identifying or estimating the flow utilities.[17]

Eqs. (12) and (13) above, showing how the per-period utility flows 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 MTA estimator lies not in the second step, but rather in the first step. In the next section, we delve into computational aspects of this first step.

Existing procedures for estimating DDC models typically rely on a small class of distributions for the utility shocks – primarily those in the extreme-value family, as in Example 1 above – because these distributions yield analytical (or near-analytical) formulas for the choice probabilities and $\{\mathbb{E}[\varepsilon_y|Y(w,\varepsilon)=y,x]\}_y$, the vector of conditional expectation of the utility shocks for the optimal choices, which is required in order to recover the utility flows[18]. Our approach, however, which is based on computing the $\mathcal{G}^*$ function, easily accommodates different choices for Q_ε, the (joint) distribution of the utility shocks 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.[19]

4. Computational details for the MTA estimator

In Section 4.1, we show that the problem of identification in DDC models can be formulated as a mass transport problem, and also how this may be implemented in practice. In showing how to compute $\mathcal{G}^*$, we exploit the connection, alluded to above, between this

[17] To ensure that the inverted w satisfies certain shape restrictions, the linkage between w and the CCP will no longer be stipulated by the subdifferential of the convex conjugate function. It is possible that there exists a modification of the convex conjugate function that is *equivalent* to imposing certain shape restrictions on utilities. This is an interesting avenue for future research.

[18] Related papers include Hotz and Miller (1993), Hotz, Miller, Sanders, Smith (1994), Aguirregabiria and Mira (2007), Pesendorfer and Schmidt-Dengler (2008), Arcidiacono and Miller (2011). Norets and Tang (2013) propose another estimation approach for binary dynamic choice models in which the choice probability function is not required to be known.

[19] 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 probabilities, thus necessitating a simulation-based inversion procedure. In ongoing work (Chiong, Galichon, Shum (2013)), we are exploring the estimation of random-coefficients discrete-choice demand models using our approach.

function and the assignment game, a model of two-sided matching with transferable utility which has been used to model marriage and housing markets (such as Shapley and Shubik (1971) and Becker (1973)).

4.1. **Mass Transport formulation.** Much of our computational strategy will be based on the following proposition, which was derived in Galichon and Salanié (2012, Proposition 2). It characterizes the $\mathcal{G}^*$ function as an optimum of a well-studied mathematical program: the "mass transport,"problem, see Villani (2003).

Proposition 2 (Galichon and Salanié)**.** *Given Assumption (2), the function $\mathcal{G}^*(p)$ is the value of the mass transport problem in which the distribution Q of vectors of utility shocks ε 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*

$$c(y, \varepsilon) = -\varepsilon_y$$

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

$$\mathcal{G}^*(p) = \sup_{\substack{w,z \\ s.t.\ w_y + z(\varepsilon) \leq c(y,\varepsilon)}} \{\mathbb{E}_p[w_Y] + \mathbb{E}_Q[z(\varepsilon)]\}, \tag{14}$$

where the supremum is taken over the pair (w, z), where w_y is a vector of dimension $|\mathcal{Y}|$ and $z(\cdot)$ is a Q-measurable random variable. By Monge-Kantorovich duality, (14) coincides with its dual

$$\mathcal{G}^*(p) = \min_{\substack{Y \sim p \\ \varepsilon \sim Q}} \mathbb{E}[c(Y, \varepsilon)], \tag{15}$$

where the minimum is taken over the joint distribution of (Y, ε) such that the the first margin Y has distribution p and the second margin ε has distribution Q. Moreover, $w \in \partial\mathcal{G}^*(p)$ if and only if there exists z such that (w, z) solves (14). Finally, $w^0 \in \partial\mathcal{G}^*(p)$ and $\mathcal{G}(w^0) = 0$ if and only if there exists z such that (w^0, z) solves (14) and z is such that $\mathbb{E}_Q[z(\varepsilon)] = 0$.

In Eq. (15) above, the minimum is taken across all joint distributions of (Y, ε) with marginal distribution equal to, respectively, p and Q. It follows from the proposition that the main problem of identification of the choice-specific value functions w can be recast as a mass transport problem (Villani (2003)), in which the set of optimizers to Eq. (14) yield vectors of choice-specific value functions $w \in \partial\mathcal{G}^*(p)$.

Moreover, the mass transport problem can be interpreted as an optimal matching problem. Using a marriage market analogy, consider a setting in which a matched couple consisting of a "man" (with characteristics $y \sim p$) and a "woman" (with characteristics $\varepsilon \sim Q$)

obtain a joint marital surplus $-c(y,\varepsilon)=\varepsilon_y$. Accordingly, Eq. (15) is an optimal matching problem in which the joint distribution of characteristics (y,ε) of matched couples is chosen to maximize the aggregate marital surplus.

In the case when Q is a discrete distribution, the mass transport problem in the above proposition reduces to a linear-programming problem which coincides with the assignment game of Shapley and Shubik (1971). This connection suggests a convenient way for efficiently computing the $\mathcal{G}^*$ function (along with its subgradient). Specifically, we will show how the dual problem (Eq. (15)) takes the form of a linear programming problem or assignment game, for which some of the associated Lagrange multipliers correspond to the the subgradient $\partial\mathcal{G}^*$, and hence the choice-specific value functions. These computational details are the focus of Section 4 below. We include the proof of Proposition 2 in the Appendix for completeness.

4.2. **Linear programming computation.** Let $\hat{Q}$ be a discrete approximation to the distribution Q. Specifically, consider a S-point approximation to Q, where the support is $\text{Supp}(\hat{Q})=\{\varepsilon^1,\ldots,\varepsilon^S\}$. Let $Pr(\hat{Q}=\varepsilon^s)=q_s$. The best S-point approximation is such that the support points are equally weighted, $q_s=\frac{1}{S}$, i.e. the best $\hat{Q}$ is a uniform distribution, see Kennan (2006). Therefore, let $\hat{Q}$ be a uniform distribution whose support can be constructed by drawing S points from the distribution Q. Moreover, $\hat{Q}$ converges to Q uniformly as $S\to\infty$,[20] so that the approximation error from this discretization will vanish when S is large. Under these assumptions, Problem (14)-(15) has a Linear Programming formulation as

$$\max_{\pi\geq 0}\sum_{y,s}\pi_{ys}\varepsilon_y^s \tag{16}$$

$$\sum_{s=1}^{S}\pi_{ys}=p_y,\ \forall y\in\mathcal{Y} \tag{17}$$

$$\sum_{y\in\mathcal{Y}}\pi_{ys}=q_s,\ \forall s\in\{1,...,S\}\,. \tag{18}$$

For this discretized problem, the set of $w\in\partial\mathcal{G}^*(p)$ is the set of vectors w of Lagrange multipliers corresponding to constraints (17). To see how we recover w^0, the specific element

[20] Because $\hat{Q}$ is constructed from i.i.d. draws from Q, this uniform convergence follows from the Glivenko-Cantelli Theorem.

in $\partial\mathcal{G}^*(p)$ as defined in Theorem 1, we begin with the dual problem

$$\begin{aligned} &\min_{\lambda,z} \sum_{y\in\mathcal{Y}} p_y\lambda_y + \sum_{s=1}^{S} q_s z_s \\ &s.t.\ \lambda_y + z_s \geq \varepsilon_y^s \end{aligned} \tag{19}$$

Consider (λ, z) a solution to (19). By duality, λ and z are, respectively, vectors of Lagrange multipliers associated to constraints (17) and (18).[21] We have $\mathcal{G}^*(p) = \sum_{y\in\mathcal{Y}} p_y\lambda_y + \sum_{s=1}^{S} q_s z_s$, which implies[22] that $\mathcal{G}(\lambda) = -\sum_{s=1}^{S} q_s z_s$. Also, for any two elements $\lambda, w^0 \in \partial\mathcal{G}^*(p)$, we have $\sum_{y\in\mathcal{Y}} p_y\lambda_y - \mathcal{G}(\lambda) = \sum_{y\in\mathcal{Y}} p_y w_y^0 - \mathcal{G}(w^0)$.

Hence, because $\mathcal{G}(w^0) = 0$, we get

$$w_y^0 = \lambda_y - \mathcal{G}(\lambda) = \lambda_y + \sum_{s=1}^{S} q_s z_s. \tag{20}$$

In Theorem 5 below, we establish the consistency of this estimate of w^0.

4.3. **Discretization of Q and a second type of indeterminacy issue.** Thus far, we have proposed a procedure for computing $\mathcal{G}^*$ (and the choice-specific value functions w^0) by discretizing the otherwise continuous distribution Q. However, because the support of ε is discrete, w_y^0 will generally not be unique.[23] This is due to the non-uniqueness of the solution to the dual of the LP problem in Eq. (16), and corresponds to 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. In this section, we discuss this partial identification, or indeterminacy, problem further.

Recall that

$$\mathcal{G}^*(p) = \sup_{w_y + z(\varepsilon) \leq c(y,\varepsilon)} \{\mathbb{E}_p[w_Y] + \mathbb{E}_Q[z(\varepsilon)]\} \tag{21}$$

[21] Because the two linear programs (16) and (19) 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 (16).

[22] This uses Eq. (25) in Appendix A, which (in our setup) states that $\mathcal{G}^*(p) + \mathcal{G}(\lambda) = p \cdot \lambda$, for all Lagrange multiplier vectors $\lambda \in \partial\mathcal{G}^*(p)$.

[23] Note that Theorem 1 requires ε to have full support.

where $c(y,\varepsilon) = -\varepsilon_y$. In Proposition 2, this problem was shown to be the dual formulation of an optimal assignment problem.

We call *identified set of payoff vectors*, denoted by $\mathcal{I}(p)$, the set of vectors w such that

$$\Pr\left(w_y + \varepsilon_y \geq \max_{y'}\{w_{y'} + \varepsilon_{y'}\}\right) = p_y \tag{22}$$

and we denote by $\mathcal{I}_0(p)$ the *normalized identified set of payoff vectors*, that is the set of $w \in \mathcal{I}(p)$ such that $\mathcal{G}(w) = 0$. If Q were to have full support, $\mathcal{I}_0(p)$ would contain only the singleton $\{w^0\}$ as in Theorem 3. Instead, when the distribution Q is discrete, the set $\mathcal{I}_0(p)$ contains a multiplicity of vectors w which satisfy (5). One has:

Theorem 4. *The following holds:*

(i) The set $\mathcal{I}(p)$ coincides with the set of w such that there exists z such that (w,z) is a solution to (21). Thus

$$\mathcal{I}(p) = \left\{ w : \exists z, \quad \begin{array}{c} w_y + z_\varepsilon \leq c(y,\varepsilon) \\ \mathbb{E}_p[w_Y] + \mathbb{E}_Q[z_\varepsilon] = \mathcal{G}^*(p) \end{array} \right\}.$$

(ii) The set $\mathcal{I}_0(p)$ is determined by the following set of linear inequalities

$$\mathcal{I}_0(p) = \left\{ w : \exists z, \quad \begin{array}{c} w_y + z_\varepsilon \leq c(y,\varepsilon) \\ \mathbb{E}_p[w_Y] = \mathcal{G}^*(p) \\ \mathbb{E}_Q[z_\varepsilon] = 0 \end{array} \right\}.$$

This result allows us to easily derive bounds on the individual components of w^0 using the characterization of the identified set using linear inequalities. Indeed, for each $y \in \mathcal{Y}$, we can obtain upper (resp. lower) bounds on w_y by maximizing (resp. minimizing) w_y subject to the linear inequalities characterizing $\mathcal{I}_0(p)$,[24] which is a linear programming problem.[25]

Furthermore, when the dimensionality of discretization, S, is high, the core shrinks to a singleton, and the core collapses to $\{w^0\}$. This is a consequence of our next theorem, which

[24] However, letting $\bar{w}_y$ (resp. $\underline{w}_y$) denote the upper (resp. lower) bound on w_y, we note that typically the vector $(w_y, y \in \mathcal{Y})' \notin \mathcal{I}_0(p)$.

[25] Moreover, partial identification in w^0 (due to discretization of the shock distribution $Q(\varepsilon)$ will naturally also imply partial identification in the utility flows u^0. For a given identified vector w^0 (and also given the choice probabilities p and transition matrix Π^0 from the data), we can recover the corresponding u^0 using Eqs. (12)-(13).

is a consistency result.[26] In our Monte Carlo experiments below, we provide evidence for the magnitude of this indeterminacy problem under different levels of discretization.

4.4. **Consistency of MTA estimator.** Here we show (strong) consistency for our MTA estimator of w^0, the normalized choice-specific value functions. In our proof, we accommodate two types of error: (i) approximation error from discretizing the distribution Q of ε, and (ii) sampling error from our finite-sample observations of the choice probabilities. We use Q^n to denote the discretized distributions of ε, and p^n to denote the sample estimates of the choice probabilities. The limiting vector of choice probabilities is denoted p^0. For a given (Q^n, p^n), let w_y^n denote the choice-specific value functions estimated using our MTA approach.

Theorem 5. *Assume:*

(i) The sequence of vectors $\left\{p_y^n\right\}_{y\in\mathcal{Y}}$*, viewed as the multinomial distribution of* y*, converges weakly to* p^0*;*

(ii) The discretized distributions of ε *converge weakly to* Q*:* $Q^n \xrightarrow{d} Q$*;*

(iii) The second moments of Q^n *are uniformly bounded.*

Then the convergence $w_y^n \to w_y^0$ *for each* $y \in \mathcal{Y}$ *holds almost surely.*

The proof, which is in the appendix, may be of independent interest as the main argument relies on approximation results from mass transport theory, which we believe to be the first use of such results for proving consistency in an econometrics context.

5. Monte Carlo Evidence

In this section, we illustrate our estimation framework using a dynamic model of resource extraction. To illustrate how our method can tractably handle any general distribution of the unobservables, we use a distribution in which shocks to different choices are correlated. We will begin by describing the setup.

At each time t, let $x_t \in \{1, 2, \ldots, 30\}$ be the state variable denoting the size of the resource pool. There are three choices,

[26] Gretsky, Ostroy, and Zame (1999) also discusses this phenomenon in their paper.

$y_t = 0$**:** The pool of resources is extracted fully. $x_{t+1}|x_t, y_t = 0$ follows a multinomial distribution on $\{1, 2, 3, 4\}$ with parameter $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$. The utility flow is $\bar{u}(y_t = 0, x_t) = 0.5\sqrt{x_t} - 2 + \varepsilon_0$.

$y_t = 1$**:** The pool of resources is extracted partially. $x_{t+1}|x_t, y_t = 1$ follows a multinomial distribution on $\{\max\{1, x_t - 10\}, \max\{2, x_t - 9\}, \max\{3, x_t - 8\}, \max\{4, x_t - 7\}\}$ with parameter π. The utility flow is $\bar{u}(y_t = 1, x_t) = 0.4\sqrt{x_t} - 2 + \varepsilon_1$.

$y_t = 2$**:** Agent waits for the pool to grow and does not extract. $x_{t+1}|x_t, y_t = 3$ follows a multinomial distribution on $\{x_t, x_t + 1, x_t + 2, x_t + 3\}$ with parameter π. We normalize the utility flow to be $\bar{u}(y_t = 2, x_t) = \varepsilon_2$.

The joint distribution of the unobserved state variables is given by $(\varepsilon_0 - \varepsilon_2, \varepsilon_1 - \varepsilon_2) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$. Other parameters we fix and hold constant for the Monte Carlo study are the discount rate, $\beta = 0.9$ and $\pi = (0.3, 0.35, 0.25, 0.10)$.

5.1. **Asymptotic performance.** As a preliminary check of our estimation procedure, we show that we are able to recover the utility flows using the actual conditional choice probabilities implied by the underlying model. We discretized the distribution of ε using $S = 5000$ support points. As is clear from Figure 1, the estimated utility flows (plotted as dots) as a function of states matched the actual utility functions very well.

5.2. **Finite sample performance.** To test the performance of our estimation procedure when there is sampling error in the CCPs, we generate simulated panel data of the following form: $\{y_{it}\ , x_{it} : i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T\}$ where $y_{it} \in \{0, 1, 2\}$ is the dynamically optimal choice at x_{it} after the realization of simulated shocks. We vary the number of cross-section observations N and the number periods T, and for each combination of (N, T), we generate 100 independent datasets.[27]

For each replication or simulated dataset, the root-mean-square error (RMSE) and R^2 are calculated, showing how well the estimated $\bar{u}_y(x)$ fits the true utility function for each y. The averages are reported in Table 1.

[27] In each dataset, we initialized x_{i1} with a random state in $\mathcal{X}$. When calculating RMSE and R^2, we restrict to states where the probability is in the interior of the simplex Δ^3, otherwise utilities are not identified and the estimates are meaningless.

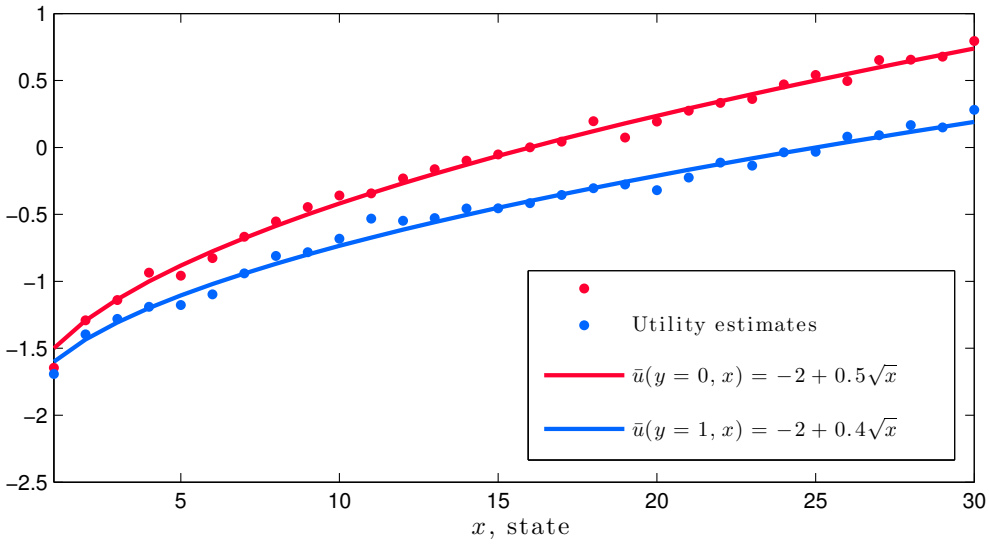


FIGURE 1. Comparison between the estimated and true utility flows.

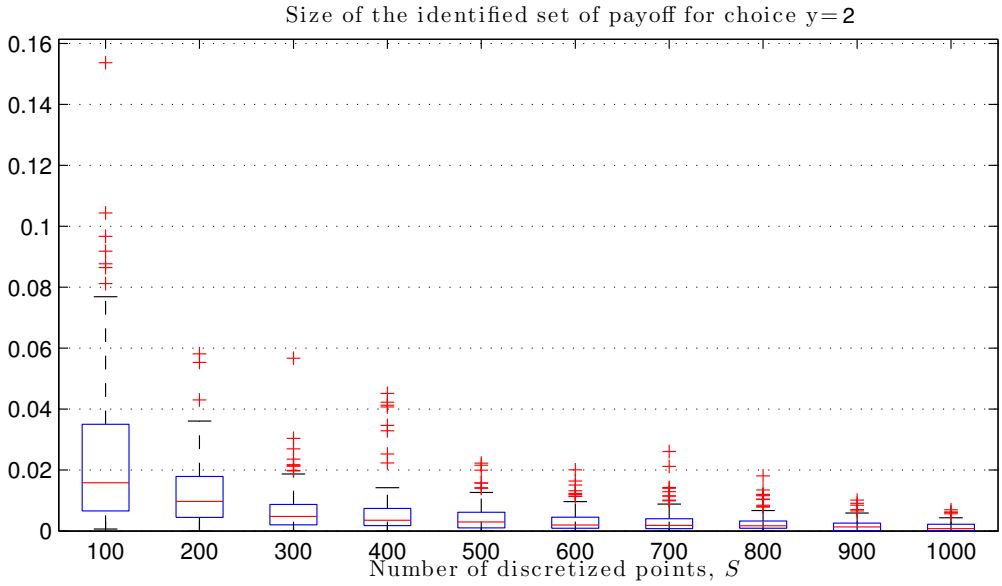
Design	RMSE($y=0$)	RMSE($y=1$)	$R^2(y=0)$	$R^2(y=1)$
$N=100, T=100$	0.5586	0.2435	0.3438	0.7708
$N=100, T=500$	0.1070	0.1389	0.7212	0.9119
$N=100, T=1000$	0.0810	0.1090	0.8553	0.9501
$N=200, T=100$	0.1244	0.1642	0.5773	0.8736
$N=200, T=200$	0.1177	0.1500	0.7044	0.9040
$N=500, T=100$	0.0871	0.1162	0.8109	0.9348
$N=500, T=500$	0.0665	0.0829	0.8899	0.9678
$N=1000, T=100$	0.0718	0.0928	0.8777	0.9647
$N=1000, T=1000$	0.0543	0.0643	0.9322	0.9820

TABLE 1. Average fit across all replications. Standard deviations are reported in the Appendix.

5.3. **Size of the identified set of payoffs.** As mentioned in Section 4.3, using a discrete approximation to the distribution of the unobserved state variable introduces a partial identification problem: the identified choice-specific value functions might not be unique. Using simulations, we next show that the identified set of choice-specific value functions

(which we will simply refer to as "payoffs") shrinks to a singleton as S increases, where S is the number of support points in the discrete approximation of the continuous error distribution. For S ranging from 100 to 1000, we plot in Figure 2, the differences between the largest and smallest choice-specific value function for $y = 2$ across all values of $p \in \Delta^3$ (using the linear programming procedures described in Section 4.3).

FIGURE 2. The identified set of payoffs shrinks to a singleton across Δ^3.



For each value of S, we plot the values of the differences $\max_{w \in \partial \mathcal{G}^*(p)} w - \min_{w \in \partial \mathcal{G}^*(p)} w$ across all values of $p \in \Delta^3$. In the boxplot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

As is evident, even at small S, the identified payoffs are very close to each other in magnitude. At $S = 1000$, where computation is near-instantaneous, for most of the values in the discretised grid of Δ^3, the core is a singleton; when it is not, the difference in the estimated payoff is less than 0.01. Similar results hold for the choice-specific value functions for choices $y = 0$ and $y = 1$, which are plotted in Figures 6 and 7 in the Appendix. To sum up, it appears that this indeterminacy issue in the payoffs is not a worrisome problem for even very modest values of S.

5.4. **Comparison: MTA vs. Simulated Maximum Likelihood.** One common technique used in the literature to estimate dynamic discrete choice models with non-standard distribution of unobservables is the Simulated Maximum Likelihood (SML). Our MTA method has a distinct advantage over SML – while MTA allows the utility flows $\bar{u}_y(x)$ for different choices y and states x to be nonparametric, the SML approach typically requires parameterizing these utility flows as a function of a low-dimensional parameter vector. This makes comparison of these two approaches awkward. Nevertheless, here we undertake a comparison of the nonparametric MTA vs. the parametric SML approach. First we compare the performance of the two alternative approaches in terms of computational time. The computations were performed on a Quad Core Intel Xeon 2.93GHz UNIX workstation, and the results are presented in Table 2.

From a computational point of view, the disadvantage of SML is that the dynamic programming problem must be solved (via Bellman function iteration) for each trial parameter vector, whereas the MTA requires solving a large-scale linear programming problem – but only *once*. Table 2 shows that our MTA procedure significantly outperforms SML in terms of computational speed. This finding, along with the results in Table 1, show that MTA has the desirable properties of speed and accuracy, and also allows for nonparametric specification of the utility flows $\bar{u}_y(x)$.

Furthermore, the SML approach can be very inaccurate, in addition to being slower. This is illustrated in Figure 3, which contains a histogram, across 250 replications, of SML estimates of the utility parameters (same setup as before, with $S = 5000$). We can see that estimates from SML depart from the true value of 2. The reason appears to be that the SML optimization algorithm[28] tends to terminate at local rather than global optima; even worse, the termination points are highly dependent upon the user-provided initial values. In contrast, the MTA estimator, by virtue of its being a linear programming problem, does not require specifying initial estimates, always finds the global optimum due to the linearity, and accordingly, has only negligible bias in finite samples (as we previously documented in Figure 1 and Table 1).

[28] We used MATLAB's nonlinear constrained optimization algorithm (*fmincon*) here, a typical choice for researchers. Unfortunately an exhaustive search for the global maximum is too computationally prohibitive in this non-smooth optimization problem. Even then it will be dependent on what specified intervals of parameters to search for.

TABLE 2. Comparison: MTA vs. Simulated Maximum Likelihood (SML)

S discretized points	SML:$^{+}$ Avg. seconds	MTA:$^{++}$ Avg. seconds
2000	19.8	2.6
3000	24.5	4.4
4000	26.5	6.6
5000	40.9	9.6
6000	70.5	13.4
7000	105.0	17.5
8000	129.4	21.5

$^{+}$:In this column we report time it takes to estimate the parameters $\theta = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$ as a local maximum of a simulated maximum likelihood, where θ corresponds to $\bar{u}_{y=0}(x) = \theta_{00} + \theta_{01}\sqrt{x}$, and $\bar{u}_{y=1}(x) = \theta_{10} + \theta_{11}\sqrt{x}$.

$^{++}$:In this column we report the time it takes to nonparametrically estimate the per-period utility flow.

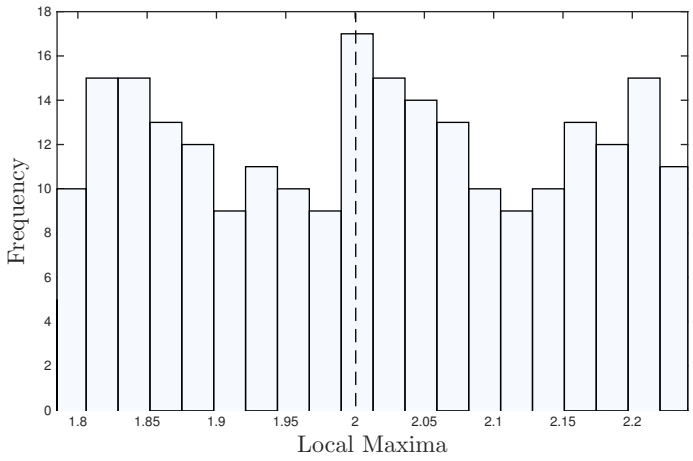


FIGURE 3. Histogram of SML estimates, across 250 replications. The true value is $\theta_{10} = 2$ in the per-period utility $\bar{u}(y = 1, x) = -\theta_{10} + \theta_{11}\sqrt{x}$. Initial values are uniformly drawn in the interval $[1.75, 2.25]$, and comparable result is obtained with much wider interval.

Indeed, one advantage of the Hotz-Miller approach is that, under the logistic assumption on unobservables and linear-in-parameters utility, the system of equations that defines the

estimator has a unique (global) solution, which avoids the problem with the SML terminating at a local optimum; in their discussion of this, Aguirregabiria and Mira (2010, pg. 48) remark that "extending the range of applicability of ... CCP methods to models which do not impose the CLOGIT [logistic] assumption is a topic for further research." This paper fills the gap: our MTA estimator is an invertibility result using the CCP, but works for non-logistic models. In this sense, the MTA estimator is a generalized CCP estimator.

6. Empirical Application: Revisiting Harold Zurcher

In this section, we apply our estimation procedure to the bus engine replacement dataset first analyzed in Rust (1987). In each week t, Harold Zurcher (bus depot manager), chooses $y_t \in \{0, 1\}$ after observing the mileage $x_t \in \mathcal{X}$ and the realized shocks ε_t. If $y_t = 0$, then he chooses not to replace the bus engine, and $y_t = 1$ means that he chooses to replace the bus engine. The states space is $\mathcal{X} = \{0, 1, \dots 29\}$, that is, we divided the mileage space into 30 states, each representing a 12,500 increment in mileage since the last engine replacement.[29] Harold Zurcher manages a fleet of 104 identical buses, and we observe the decisions that he made, as well as the corresponding bus mileage at each time period t. The duration between $t + 1$ and t is a quarter of a year, and the dataset spans 10 years. Figures 8 and 9 in the Appendix summarize the frequencies and mileage at which replacements take place in the dataset.

Firstly, we can directly estimate the probability of choosing to replace and not to replace the engine for each state in $\mathcal{X}$. Also directly obtained from the data is the Markov transition probabilities for the observed state variable $x_t \in X$, estimated as:

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

[29] This grid is coarser compared to Rust's (1987) original analysis of this data, in which he divided the mileage space into increments of 5,000 miles. However, because replacement of engines occurred so infrequently (there were only 61 replacement in the entire ten-year sample period), using such a fine grid size leads to many states that have zero probability of choosing replacement. Our procedure – like all other CCP-based approaches – fails when the vector of conditional choice probability lies on the boundary of the simplex.

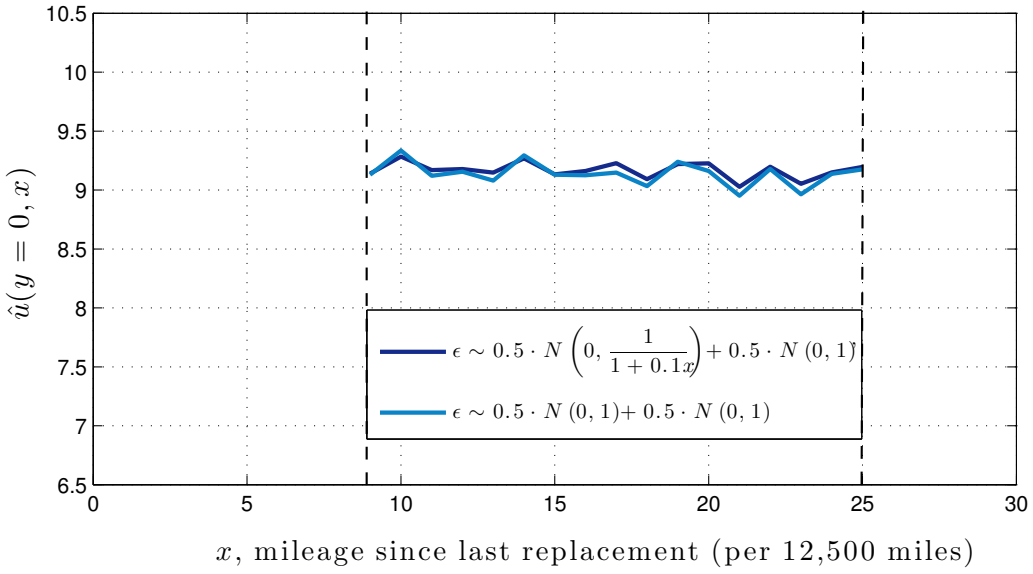


FIGURE 4. Estimates of utility flows $\bar{u}_{y=0}(x)$, across values of mileage x

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

For this analysis, we assumed a normal mixture distribution of the error term, specifically, $\varepsilon_{t0}-\varepsilon_{t1} \sim \frac{1}{2}N(0,1)+\frac{1}{2}N(0,\frac{1}{1+0.1x})$.[30] We chose this mixture distribution in order to allow the utility shocks to depend on mileage – which accommodates, for instance, operating costs which may be more volatile and unpredictable at different levels of mileage. At the same time, these specifications for the utility shock distribution showcase the flexibility of our procedure in estimating dynamic discrete choice models for any general error distribution.

[30] In this paper, we restrict attention to the case where the researcher fully knows the distribution of the unobservables $Q_{\vec{\varepsilon}}$, so that there are no unknown parameters in these distributions. In principle, the two-step procedure proposed here can be nested inside an additional "outer loop" in which unknown parameters of $Q_{\vec{\varepsilon}}$ are considered, but identification and estimation in this case must rely on additional model restrictions in addition to those considered in this paper. We are currently exploring such a model in the context of the simpler static discrete choice setting (Chiong, Galichon and Shum (2014, work in progress)).

For comparison, we repeat this exercise using an error distribution that is homoskedastic, i.e., its variance does not depend on the state variable x_t. The result appears to be robust to using different distributions of $\varepsilon_{t0} - \varepsilon_{t1}$. We set the discount rate $\beta = 0.9$.

To non-parametrically estimate $\bar{u}_{y=0}(x)$, we fixed $\bar{u}_{y=1}(x)$ to 0 for all $x \in X$. Hence, our estimates of $\bar{u}_{y=0}(x)$ should be interpreted as the magnitude of operating costs[31] relative to replacement costs[32], with positive values implying that replacement costs exceed operating costs. The estimated utility flows from choosing $y = 0$ (don't replace) relative to $y = 1$ (replace engine) are plotted in Figure 4. We only present estimates for mileage within the range $x \in [9, 25]$, because within this range, the CCPs are in the interior of the probability simplex (cf. footnote 29 and Figure 9 in appendix).

Within this range, the estimated utility function does not vary much with increasing mileages, i.e. it has slope that is not significantly different from zero. The recovered utilities fall within the narrow band of 9 and 9.5, which implies that on average the replacement cost is much higher than the maintenance cost, by a magnitude of 18 to 19 times the variance of the utility shocks. It is somewhat surprising that our results suggest that when the mileage goes beyond the cutoff point of 100,000 miles, Harold Zurcher perceived the operating costs to be inelastic with respect to accumulated mileage. It is worth noting that Rust (1987) mentioned: "According to Zurcher, monthly maintenance costs increase very slowly as a function of accumulated mileage."

To get an idea for the effect of sampling error on our estimates, we bootstrapped our estimation procedure. For each of 100 resamples, we randomly drew 80 buses with replacement from the dataset, and re-estimated the utility flows $\bar{u}_{y=0}(x)$ using our procedure. The results are plotted in Figure 5. The evidence suggests that we are able to obtain fairly tight cost estimates for states where there is at least one replacement, i.e. for $x \geq 9$ ($x \geq 112,500$ miles), and for states that are reached often enough; i.e. for $x \leq 22$ ($x \leq 275,000$ miles).

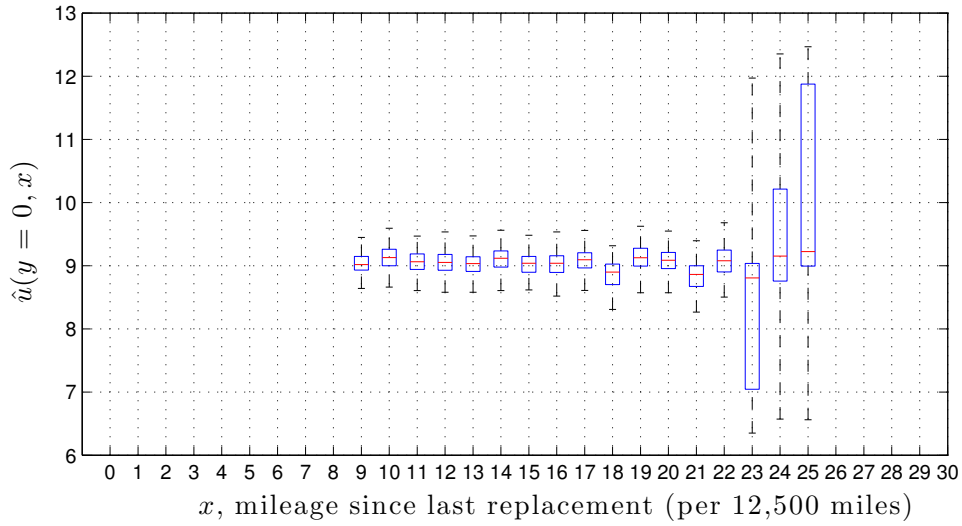
7. Conclusion

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

[31] Operating costs include maintenance, fuel, insurance costs, plus Zurcher's estimate of the costs of lost ridership and goodwill due to unexpected breakdowns.

[32] To be pedantic, this also includes the operating cost at $x = 0$.

FIGURE 5. Bootstrapped estimates of utility flows $\bar{u}_{y=0}(x)$



We plot the values of the bootstrapped resampled estimates of $\bar{u}_{y=0}(x)$. In each boxplot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the 5th and 95th percentiles.

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 approach we take, based on the convexity of the social surplus function $\mathcal{G}$ and the resulting duality between choice probabilities and choice-specific value functions, 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, we can apply the powerful tools developed to compute optimal matchings to dynamic discrete-choice

models.[33] Moreover, by reformulating the problem as an optimal matching problem, all existence and uniqueness results are inherited from the theory of optimal transport. 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.

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.[34] However, we leave these directions for future exploration.

References

[1] V. Aguirregabiria and P. Mira. Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica*, 70:1519-1543, 2002.

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

[3] V. Aguirregabiria and P. Mira. Dynamic discrete choice structural models: a survey. *Journal of Econometrics*, 156:38–67, 2010.

[4] Anderson, S., de Palma, A., and Thisse, J.-F. A Representative Consumer Theory of the Logit Model. *International Economic Review*, 29(3), 461-466, 1988.

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

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

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

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

[9] S. Berry, A. Gandhi, and P. Haile. Connected Substitutes and Invertibility of Demand. *Econometrica* 81: 2087-2111, 2013.

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

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

[12] D. Bertsekas. *Dynamic Programming Deterministic and Stochastic Models*. Prentice-Hall, 1987.

[33] While the present paper has used standard Linear Programming algorithms such as the Simplex algorithm, other, more powerful matching algorithms such as the Hungarian algorithm may be efficiently put to use when the dimensionality of the problem grows.

[34] See, e.g. Aguirregabiria and Mira (2007) or Pesendorfer and Schmidt-Dengler (2008)).

[13] P. Chiappori and I. Komunjer. On the Nonparametric Identification of Multiple Choice Models. Working paper, 2010.

[14] K. Chiong, A. Galichon, and M. Shum. Simulation and Partial Identification in Random Coefficient Discrete Choice Demand Models. Work in progress, 2014.

[15] 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.

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

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

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

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

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

[21] 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.

[22] Y. Hu and M. Shum. Nonparametric Identification of Dynamic Models with Unobserved Heterogeneity. *Journal of Econometrics*, 171: 32-44, 2012.

[23] H. Kasahara and K. Shimotsu. Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choice. *Econometrica*, 77: 135–175, 2009.

[24] M. Keane and K. Wolpin. The career decisions of young men. *Journal of Political Economy*, 105: 473–522, 1997.

[25] J. Kennan. A Note on Discrete Approximations of Continuous Distributions. Mimeo, University of Wisconsin at Madison, 2006.

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

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

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

[29] A. Norets. Inference in dynamic discrete choice models with serially correlated unobserved state variables. *Econometrica*, 77: 1665-1682, 2009.

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

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

[32] A. Pakes. Patents as options: some estimates of the value of holding European patent stocks. *Econometrica*, 54:1027-1057, 1986.

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

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

[35] J. Rust. Structural Estimation of Markov Decision Processes. *Handbook of Econometrics*, Volume 4 (ed. R. Engle and D. McFadden). North-Holland, 1994.

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

[37] X. Shi, M. Shum, and W. Song. Estimating Multinomial Models using Cyclic Monotonicity. Caltech Social Science Working Paper 1397, 2014.

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

[39] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics, Vol. 58. American Mathematical Society, 2003.

[40] C. Villani. *Optimal Transport, Old and New*. Springer, 2009.

8. Background results

8.1. **Convex Analysis for Discrete-choice Models .** Here, we give a brief review of the main notions and results used in the paper. We keep an informal style and do not give proofs, but we refer to Rockafellar (1970) for an extensive treatment of the subject.

Let $u \in \mathbb{R}^{|\mathcal{Y}|}$ be a vector of utility indices. For utility shocks $\{\varepsilon_y\}_{y\in\mathcal{Y}}$ distributed according to a joint distribution function Q, we define the social surplus function as

$$\mathcal{G}(u) = \mathbb{E}[\max_y \{u_y + \varepsilon_y\}], \tag{23}$$

where u_y is the y-th component of u. If $\mathbb{E}(\varepsilon_y)$ exists and is finite, then the function $\mathcal{G}$ is a proper convex function that is continuous everywhere. Moreover assuming that Q is sufficiently well-behaved (for instance, if it has a density with respect to the Lebesgue measure), $\mathcal{G}$ is differentiable everywhere.

Define the *Legendre-Fenchel conjugate*, or *convex conjugate* of $\mathcal{G}$ as $\mathcal{G}^*(p) = \sup_{u\in\mathbb{R}^{|\mathcal{Y}|}}\{p \cdot u - \mathcal{G}(u)\}$. Clearly, $\mathcal{G}^*$ is a convex function as it is the supremum of affine functions. Note that the inequality

$$\mathcal{G}(u) + \mathcal{G}^*(p) \geq p \cdot u \tag{24}$$

holds in general. The domain of $\mathcal{G}^*$ consists of $p \in \mathbb{R}^{|\mathcal{Y}|}$ for which the supremum is finite. In the case when $\mathcal{G}$ is defined by (23), it follows from Norets and Takahashi (2013) that the domain of $\mathcal{G}^*$ contains the simplex $\Delta^{|\mathcal{Y}|}$, which is the set of $p \in \mathbb{R}^{|\mathcal{Y}|}$ such that $p_y \geq 0$ and $\sum_{y\in\mathcal{Y}} p_y = 1$. This means that our convex conjugate function is always well-defined.

The *subgradient* $\partial\mathcal{G}(u)$ of $\mathcal{G}$ at u is the set of $p \in \mathbb{R}^{|\mathcal{Y}|}$ such that

$$p \cdot u - \mathcal{G}(u) \geq p \cdot u' - \mathcal{G}(u')$$

holds for all $u' \in \mathbb{R}^{|\mathcal{Y}|}$. Hence $\partial\mathcal{G}$ is a set-valued function or correspondence. $\partial\mathcal{G}(u)$ is a singleton if and only if $\mathcal{G}(u)$ is differentiable at u; in this case, $\partial\mathcal{G}(u) = \nabla\mathcal{G}(u)$.

One sees that $p \in \partial\mathcal{G}(u)$ if and only if $p \cdot u - \mathcal{G}(u) = \mathcal{G}^*(p)$, that is if equality is reached in inequality (24):

$$\mathcal{G}(u) + \mathcal{G}^*(p) = p \cdot u. \tag{25}$$

This equation is itself of interest, and is known in the literature as "Fenchel's equality". By symmetry in (25), one sees that $p \in \partial\mathcal{G}(u)$ if and only if $u \in \partial\mathcal{G}^*(p)$. In particular, when both $\mathcal{G}$ and $\mathcal{G}^*$ are differentiable, then $\nabla\mathcal{G}^* = \nabla\mathcal{G}^{-1}$.

9. Proofs

Proof of Proposition 1. Consider the y-th component, corresponding to $\frac{\partial\mathcal{G}(w)}{\partial w_y}$:

$$\frac{\partial\mathcal{G}(w)}{\partial w_y} = \frac{\partial}{\partial w_y}\int \max_y [w_y + \varepsilon_y] dQ \tag{26}$$

$$= \int \frac{\partial}{\partial w_y} \max_y [w_y + \varepsilon_y] dQ \tag{27}$$

$$= \int \mathbb{1}(w_y + \varepsilon_y \geq w_{y'} + \varepsilon_{y'}), \forall y' \neq y) dQ = p(y). \tag{28}$$

(We have suppressed the dependence on x for convenience.) □

Proof of Theorem 1. This follows directly from Fenchel's equality (see Rockafellar (1970), Theorem 23.5, see also Appendix 8.1), which states that

$$p \in \partial\mathcal{G}(w)$$

is equivalent to $\mathcal{G}(w) + \mathcal{G}^*(p) = \sum_y p_y w_y$, which is equivalent in turn to

$$w \in \partial\mathcal{G}^*(p).$$

□

Proof of Theorem 2. Because ε has full support, the choice probabilities p will lie strictly in the interior of the simplex $\Delta^{|\mathcal{Y}|}$. 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

$$\begin{aligned} &w_{y_0} - w_{y_1} > \varepsilon_{y_1} - \varepsilon_{y_0} > w'_{y_0} - w'_{y_1} \\ &w_{y_0} + \varepsilon_{y_0} > \max_{y \neq y_0, y_1} \{w_y + \varepsilon_y\} \\ &w'_{y_1} + \varepsilon_{y_1} > \max_{y \neq y_0, y_1} \{w'_y + \varepsilon_y\} \end{aligned}$$

Because ε has full support, S has positive probability.

Let $\bar{w} = \frac{w+w'}{2}$. Because $p \in \partial \mathcal{G}(w)$ and $p \in \partial \mathcal{G}(w')$, $\mathcal{G}$ is linear on the segment $[w, w']$, thus $\mathcal{G}(\bar{w}) = 0$, 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 Theorem 3. From $\mathcal{G}\left(w^0\right) = 0$ and $\partial \mathcal{G}(w - \mathcal{G}(w)) = \partial \mathcal{G}(w)$, and by the uniqueness result in Theorem 2, it follows that

$$w^0 = w - \mathcal{G}(w).$$

□

Proof of Proposition 2. The proof is in Galichon and Salanié (2012), but we include it here for self-containedness. This connection between the $\mathcal{G}^*$ function and a matching model

follows from manipulation of the variational problem in the definition of $\mathcal{G}^*$:

$$\begin{aligned}\mathcal{G}^*(p) &= \sup_{w\in\mathbb{R}^{\mathcal{Y}}}\left\{\sum_y p_y w_y - \mathbb{E}_Q\left[\max_{y\in\mathcal{Y}}(w_y+\varepsilon_y)\right]\right\} \\ &= \sup_{w\in\mathbb{R}^{\mathcal{Y}}}\left\{\sum_y p_y w_y + \mathbb{E}_Q\underbrace{\left[\min_{y\in\mathcal{Y}}(-w_y-\varepsilon_y)\right]}_{\equiv z(\varepsilon)}\right\}.\end{aligned} \tag{29}$$

Defining $c(y,\varepsilon)\equiv-\varepsilon_y$, one can rewrite the above as

$$\mathcal{G}^*(p) = \sup_{w_y+z(\varepsilon)\le c(y,\varepsilon)}\left\{\mathbb{E}_p[w_Y]+\mathbb{E}_Q[z(\varepsilon)]\right\}. \tag{30}$$

As is well-known from the results of Monge-Kantorovich (Villani (2003), Thm. 1.3), this is the dual-problem for a mass transport problem. The corresponding primal problem is

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

which is equivalent to (16)-(18). Comparing Eqs. (29) and (30), we see that the subdifferential $\partial\mathcal{G}^*(p)$ is identified with those elements w such that (w,z), for some z, solves the dual problem (30). □

Proof of Theorem 4. (i) follows from Proposition 2 and the fact that if $w_y+z(\varepsilon)\le c(y,\varepsilon)$, then $\mathbb{E}_p[w_Y]+\mathbb{E}_Q[z(\varepsilon)]=\mathcal{G}^*(p)$ if and only if (w,z) is a solution to the dual problem.

(ii) follows from the fact that $-z(\varepsilon)=\sup_y\{w_y-c(y,\varepsilon)\}=\sup_y\{w_y+\varepsilon_y\}$, thus $\mathbb{E}_Q[z(\varepsilon)]=0$ is equivalent to $\mathbb{E}_Q\left[\sup_y\{w_y+\varepsilon_y\}\right]=0$, that is $\mathcal{G}(w)=0$. □

Proof of Theorem 5. We shall show that the vector of choice-specific value functions derived from the MTA estimation procedure, denoted w^n, converges to the true vector w^0. In our procedure, there are two sources of estimation error. The first is the sampling error in the vector of choice probabilities, denoted p^n. The second is the simulation error involved in the discretization of the distribution of ε; we let Q^n denote this discretized distribution.

A distinctive aspect of our proof is that it utilizes the theory of mass transport; namely convergence results for sequences of mass transport problems. For $y\in\mathcal{Y}$, let ι^y denote the $|\mathcal{Y}|$-dimensional row vector with all zeros except a 1 in the y-th column. This discretized mass transport problem from which we obtain w^n is:

$$\sup_{\gamma\in\mathcal{M}(Q^n,p^n)}\int_{\mathbb{R}^d\times\mathbb{R}^d}(\iota\cdot\varepsilon)\,\gamma(d\varepsilon,d\iota) \tag{31}$$

where $\mathcal{M}(Q^n, p^n)$ denotes the set of joint (discrete) probability measures with marginal distributions Q^n and p^n. In the above, ι denotes a random vector which is equal to ι^y with probability p_y^n, for $y \in \mathcal{Y}$. The dual problem used in the MTA procedure is

$$\inf_{z,w} \quad \int z(\varepsilon)\, dQ^n(\varepsilon) + \sum_y w_y p_y^n \ : \tag{32}$$

$$s.t. \quad z(\varepsilon) \geq \iota^y \cdot \varepsilon - w_y, \ \forall y, \ \forall \varepsilon \tag{33}$$

$$\mathcal{G}_n(w_y^n) = 0, \tag{34}$$

where $\mathcal{G}_n(w) \equiv \mathbb{E}_{Q^n}(w_y + \epsilon_y)$. We let (z^n, w^n) denote solutions to this discretized dual problem (32). Recall (from the discussion in Section 2.3) that the extra constraint (34) in the dual problem just selects among the many dual optimizing arguments (w^n, z^n) corresponding to the optimal primal solution γ^n, and so does not affect the primal problem.[35]

Next we derive a more manageable representation of this constraint (34). From Fenchel's Equality (Eq. (25)), we have $\sum_y p_y^n w_y^n = \mathcal{G}_n(w^n) + \mathcal{G}_n^*(p^n) = \mathcal{G}_n^*(p^n)$ (with $\mathcal{G}_n^*$ defined as the convex conjugate function of $\mathcal{G}_n$). Moreover, from Proposition 2, we know that $\mathcal{G}_n^*(p^n)$ can be characterized as the optimized dual objective function in (32). Hence, we see that the constraint $\mathcal{G}_n(w^n) = 0$ is equivalent to $\int z^n(\varepsilon) dQ^n(\varepsilon) = 0$. We introduce this latter constraint directly and rewrite the dual program

$$\inf_{z,w} \quad \sum_y w_y p_y^n + \int z(\varepsilon)\, dQ^n(\varepsilon) \tag{35}$$

$$s.t. \quad z(\varepsilon) \geq \iota^y \cdot \varepsilon - w_y, \ \forall y, \ \forall \varepsilon \tag{36}$$

$$\int z(\varepsilon)\, dQ^n(\varepsilon) = 0. \tag{37}$$

We will demonstrate consistency by showing that (z^n, w^n) converge a.s. to the dual optimizers in the "limit" dual problem, given by

$$\inf_{z,w} \quad \sum_y w_y p_y^0 \tag{38}$$

$$z(\varepsilon) \geq \iota^y \cdot \varepsilon - w_y, \ \forall y, \ \forall \varepsilon \tag{39}$$

$$\int z(\varepsilon)\, dQ = 0 \tag{40}$$

We denote the optimizers in this limit problem by (w^0, z^0), where, by construction, w^0 are the "true" values of the choice-specific value functions. The difference between the

[35] We note that, as discussed before, the discreteness of Q^n implies that (z^n, w^n) will not be uniquely determined, as the core of the assignment game for a finite market is not a singleton. But this does not affect the proof, as our arguments below hold for any sequence of selections $\{z^n, w^n\}_n$.

discretized and limit dual problems is that Q^n in the former has been replaced by Q, the continuous distribution of ε, and the estimated choice probabilities p^n have been replaced by the limit p^0.

We proceed in two steps. First, we argue that the sequence of optimized dual programs (35) converges to the optimized limit dual program (38), a.s. Based upon this, we then argue that the sequence of dual optimizers, (w^n, z^n), necessarily converge to their unique limit optimizers, (w^0, z^0), a.s.

First step. By the Kantorovich duality theorem, we know that the optimized values for the limit primal and dual programs coincide

$$\sup_{\gamma \in \Pi(Q^0, p^0)} \int_{\mathbb{R}^d \times \mathbb{R}^d} (\iota \cdot \varepsilon) \, \gamma(d\varepsilon, d\iota) = \inf \sum_y w_y p_y^0 + \int z(\varepsilon) \, dQ. \tag{41}$$

Moreover, both the primal and dual problems in the discretized case are finite-dimensional linear programming problem, and by the usual LP duality, the optimal primal and dual problems for the discretized case also coincide:

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\iota \cdot \varepsilon) \, \gamma_n(d\varepsilon, d\iota) = \sum_y w_y^n p_y^n + \int z^n(\varepsilon) \, dQ^n.$$

Given Assumption 1, and by Theorem 5.20 in Villani (2009), p. 77, we have that, up to a subsequence extraction, γ^n (the optimizing argument of (31)) converges weakly. In addition, by Theorem 5.30 in Villani (2009), the left-hand side of (41) has a unique solution γ; hence, the sequence γ^n must converge generally to γ. This implies a.s. convergence of the value of the primal problems:

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\iota \cdot \varepsilon) \, \gamma_n(d\varepsilon, d\iota) \to \int_{\mathbb{R}^d \times \mathbb{R}^d} (\iota \cdot \varepsilon) \, \gamma(d\varepsilon, d\iota), \quad a.s.,$$

and, by duality, we must also have a.s. convergence of the discretized dual problem to the limit problem:

$$\sum_y w_y^n p_y^n + \int z^n(\varepsilon) \, dQ^n \to \sum_y w_y p_y^0 + \int z(\varepsilon) \, dQ, \quad a.s. \tag{42}$$

Second step. Next, we show that the discretized dual minimizers (z^n, w^n) converge a.s. For convenience, in what follows we will suppress the qualifier "a.s." from all the statements below. Let

$$\underline{w}^n = \min_y w_y^n. \tag{43}$$

From examination of the dual problem (35), we see that z^n is the piecewise affine function

$$z^n(\varepsilon) = \max_y \{\iota^y \cdot \varepsilon - w_y^n\}, \tag{44}$$

thus z^n is M-Lipschitz with $M := \max_y |\iota^y| = 1$. Now observe that

$$z^n(\varepsilon) + \underline{\mathrm{w}}^n = \max_y \{\iota^y \cdot \varepsilon - w_y^n + \underline{\mathrm{w}}^n\} \leq \max_y \{\iota^y \cdot \varepsilon\} =: \overline{z}(\varepsilon) \tag{45}$$

and, letting y' be the argument of the minimum in (43),

$$z^n(\varepsilon) + \underline{\mathrm{w}}^n \geq \iota^{y'} \cdot \varepsilon - w_{y'}^n + \underline{\mathrm{w}}^n = \iota^{y'} \cdot \varepsilon \geq \min_y \{\iota^y \cdot \varepsilon\} =: \underline{z}(\varepsilon) \tag{46}$$

thus, by a combination of (45) and (46),

$$\underline{z}(\varepsilon) \leq z^n(\varepsilon) + \underline{\mathrm{w}}^n \leq \overline{z}(\varepsilon). \tag{47}$$

By $\int z^n(\varepsilon)\, dQ^n(\varepsilon) = 0$, we have that that $\underline{\mathrm{w}}^n$ is uniformly bounded (sublinear): for some constant K, $|z^n(\varepsilon)| \leq C(1+|\varepsilon|)$ for every n and every ε. Hence the sequence z^n is uniformly equicontinuous, and converges locally uniformly up to a subsequence extraction by Ascoli's theorem. Let this limit function be denoted z^0. By (42), and Theorem 2, we deduce that z, the optimizer in the limit dual problem is unique[36], so that it must coincide with the limit function z^0.

By the definition of (w^n, z^n) as optimizing arguments for (35), we have $\sum_y w_y^n p_y^n \leq \sum_y \underline{\mathrm{w}}^n p_y + \int [\overline{z}(\varepsilon)] dQ^n(\varepsilon)$ or

$$\sum_y \left(w_y^n - \underline{\mathrm{w}}^n\right) p_y^n \leq \int [\overline{z}(\varepsilon)] dQ^n(\varepsilon) = \mathbb{E}_{Q^n} \overline{z}$$

The second moment restrictions on Q^n (condition (ii) in the theorem) imply that $\mathbb{E}_{Q^n} \overline{z}(\varepsilon)$ exists and converges to $\mathbb{E}_Q \overline{z}$. Hence, the nonnegative vectors $\left(w_y^n - \underline{\mathrm{w}}^n\right)$ are bounded; accordingly, the vectors $\left(w_y^n\right)$ are themselves bounded. This implies that w^n converges up to a subsequence to some limit point w^*, using the Bolzano-Weierstrass theorem. This implies that $\sum_y w_y^n p_y^n \to \sum_y w_y^* p_y$ by bounded convergence. By Theorem 2, we know that the limit point w^* must coincide with w^0, which is the unique optimizer in the dual limit problem (38). Thus, we have shown that w^n converges to w^0, a.s. □

[36] Although the support of ε is not bounded, the locally uniform convergence of z^n and the fact that the second moments of Q^n are uniformly bounded are enough to conclude.

10. Additional Figures

Design	RMSE($y=0$)	RMSE($y=1$)	$R^2(y=0)$	$R^2(y=1)$
$N=100, T=100$	0.5586 (3.7134)	0.2435 (0.1155)	0.3438 (0.7298)	0.7708 (0.2073)
$N=100, T=500$	0.1070 (0.0541)	0.1389 (0.0638)	0.7212 (0.2788)	0.9119 (0.0820)
$N=100, T=1000$	0.0810 (0.0376)	0.1090 (0.0425)	0.8553 (0.1285)	0.9501 (0.0352)
$N=200, T=100$	0.1244 (0.0594)	0.1642 (0.0628)	0.5773 (0.6875)	0.8736 (0.1112)
$N=200, T=200$	0.1177 (0.0736)	0.1500 (0.0816)	0.7044 (0.2813)	0.9040 (0.0842)
$N=500, T=100$	0.0871 (0.0375)	0.1162 (0.0430)	0.8109 (0.2468)	0.9348 (0.0650)
$N=500, T=500$	0.0665 (0.0261)	0.0829 (0.0290)	0.8899 (0.1601)	0.9678 (0.0374)
$N=1000, T=100$	0.0718 (0.0340)	0.0928 (0.0344)	0.8777 (0.1320)	0.9647 (0.0314)
$N=1000, T=1000$	0.0543 (0.0176)	0.0643 (0.0162)	0.9322 (0.0577)	0.9820 (0.0101)

Table 3

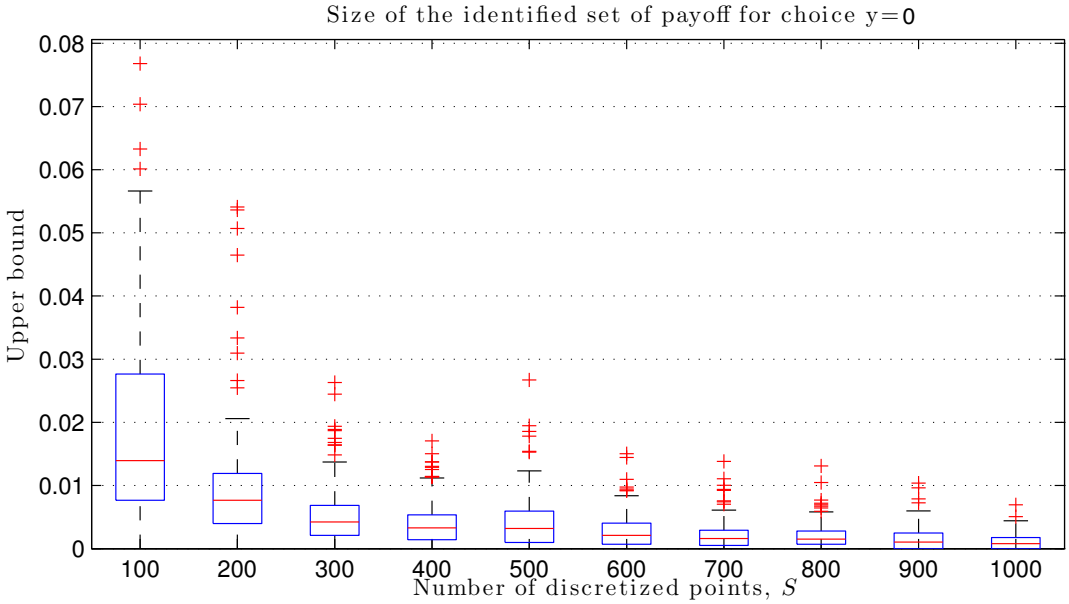


FIGURE 6. For each value of S, we plot the values of the differences $\max_{w\in\partial\mathcal{G}^*(p)} w_1 - \min_{w\in\partial\mathcal{G}^*(p)} w_1$ across all values of $p \in \Delta^3$. In the boxplot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

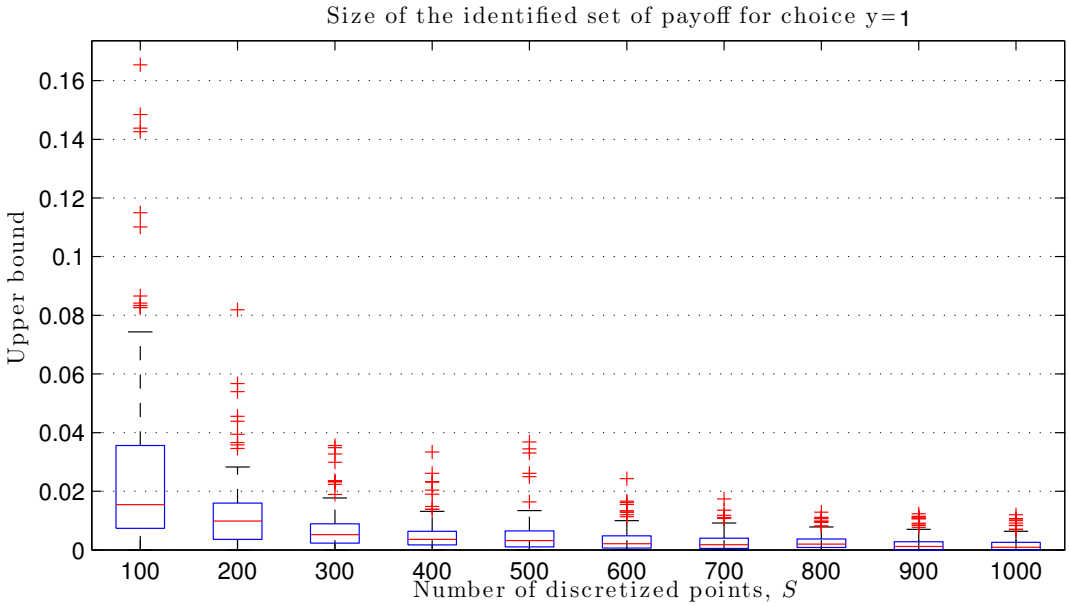


FIGURE 7. For each value of S, we plot the values of the differences $\max_{w \in \partial \mathcal{G}^*(p)} w_2 - \min_{w \in \partial \mathcal{G}^*(p)} w_2$ across all values of $p \in \Delta^3$. In the box-plot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

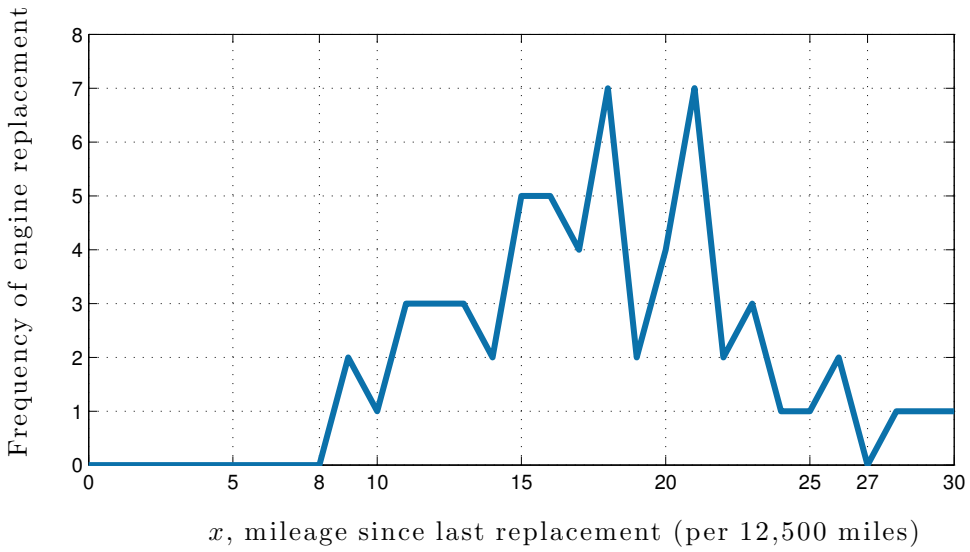


FIGURE 8

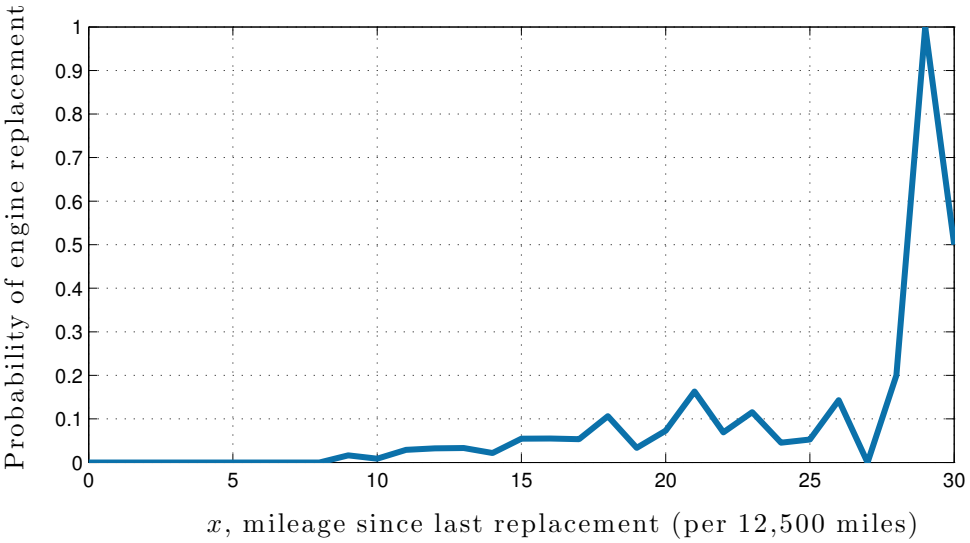


FIGURE 9