ROBUST DECISIONS FOR INCOMPLETE MODELS OF STRATEGIC INTERACTION

KONRAD MENZEL[†] AND TOBIAS SALZ[♯]

Abstract. We propose Monte Carlo Markov Chain (MCMC) methods for estimation and inference in game-theoretic models with a particular focus on settings in which only a small number of observations for a given type of game is available. In particular we do not assume that it is possible to concentrate out or estimate consistently an equilibrium selection mechanism linking a parametric distribution of unobserved payoffs to observable choices. The algorithm developed in this paper can in particular be used to analyze structural models of social interactions with multiple equilibria using data augmentation techniques. This study adapts the multiple prior framework from Gilboa and Schmeidler (1989) to compute Gamma-posterior expected loss (GPEL) optimal decisions that are robust with respect to assumptions on equilibrium selection, and gives conditions under which it is possible to solve the GPEL problem using one single Markov chain. The practical usefulness of the generic MCMC algorithm is illustrated with an application to revealed preference analysis of two-sided marriage markets with non-transferable utilities.

JEL Classification: C12, C13, C14, C15

Keywords: MCMC Methods, Incomplete Models, Partial Identification, Multiple Priors, Choquet Integral, Matching Markets

1. Introduction

The defining feature of economic models of social interactions is that individuals' payoffs are affected by other agents' actions. Relevant examples include models of firm competition, network formation, matching markets, or individual choice in the presence of social spillovers.

As has been noticed in the literature, this interdependence raises two important practical problems for estimation of structural parameters in this context: for one, standard economic solution concepts - as e.g. Nash equilibrium or match stability - may allow for multiple solutions of the model for a given set state of nature (including both observable and unobservable characteristics). In addition, even when a likelihood of the structural model is defined, interdependencies between individual actions and preferences often make it very difficult to evaluate a likelihood of a structural model directly, especially in models with a large number of participants.

Date: first draft: June 10, 2011. This version: August 2013.

[†] NYU, Department of Economics, Email: konrad.menzel@nyu.edu

[♯] NYU, Department of Economics, Email: ts1552@nyu.edu

PRELIMINARY - any comments and suggestions are welcome.

Even though in general observed outcomes of a game are informative about the equilibrium selection rule that generated the data, there are many settings in which it cannot be estimated consistently under realistic sampling assumptions. For one, we may in some cases only observe a very small number of instances of a game (e.g. realization of a network or marriage market), furthermore, if the game has a large number of players or rich action spaces, the number of potential equilibria may be very large, so that a very rich parameter space would be needed to characterize the equilibrium selection mechanism. Finally the researcher may in general be reluctant to impose too many restrictions across observed instances of the game if equilibrium selection may be related to a very rich set of other explanatory variables.

We can view the difficulty in dealing with the unknown equilibrium selection rule with a small number of observations of a game as an instance of the incidental parameters problem (Neyman and Scott (1948)). Most of the recent frequentist approaches to the problem concentrate out the nuisance parameter explicitly or in other cases implicitly as we will argue in section 2, see e.g. Chen, Tamer, and Torgovitsky (2011), Galichon and Henry (2011), and Beresteanu, Molchanov, and Molinari (2009). Imposing a specific prior or parametric distribution over equilibrium selection rules will in general require that we solve for all equilibria of the game, and any particular choice will in general be difficult to justify on economic or statistical considerations.

Instead we opt for a robust Bayes approach which is conservative with respect to the nuisance parameter. More specifically we consider decisions that are optimal with respect to Gamma posterior expected loss with respect to a class of priors over the equilibrium selection rule. A computational advantage arises from that evaluation of robust criteria often only depends on particular "extremal" points in the parameter space for the nuisance component. In particular, the methods considered in this paper require only verification of stability and uniqueness conditions instead of finding the full set of equilibria for a given realization of payoffs. However, the simulation methods described in this paper can also be adapted for a concentrated likelihood approach.

The main aim of this paper is to propose an approach that exploits some of the practical advantages of Bayesian computation - especially the use of MCMC simulation techniques - while maintaining a full set of likelihood functions arising from different selection rules for selecting from a multiplicity of predictions of the economic model. An important motivation for a Bayesian approach to structural estimation is that it often leads to computationally attractive procedures, especially in the context of latent variable models, see e.g. Tanner and Wong (1987), McCulloch and Rossi (1994), and Norets (2009). In particular, multi-stage sampling techniques and data augmentation procedures can simplify the evaluation of complex likelihood functions, especially if a closed form of the model is only available conditional on a latent variable.

As a motivating empirical application, we consider revealed-preference analysis in a model of a two-sided marriage market, where the observed marriages are assumed to constitute a stable matching as in the theoretical analysis of Gale and Shapley (1962). This model has previously been analyzed by Logan, Hoff, and Newton (2008), and the procedure proposed in this paper leads to a modification of their approach which accounts for the multiplicity of stable matchings. In general the matching market model without transferable utilities will not be point-identified unless we impose a rule for selecting among multiple stable matchings.

Structural models of social interactions have been considered in many contexts in economics, prominent examples are firm entry decisions in concentrated markets (e.g. Bresnahan and Reiss (1990) and Ciliberto and Tamer (2009)), firm mergers (Fox (2010)), neighborhood effects and spillovers (see Brock and Durlauf (2001) and references therein), or network formation (Christakis, Fowler, Imbens, and Kalyanaraman (2010)). Many of these models have the structure of a simultaneous discrete choice model which was first analyzed by Heckman (1978). A central difficulty in many of these models is the existence of multiple equilibria, that is a failure of the solution concept used for the econometric model to generate a unique prediction given the underlying state of nature (see e.g. Jovanovic (1989),Bresnahan and Reiss (1991a),Tamer (2003)). Bajari, Hong, and Ryan (2010) analyze identification and estimation of an equilibrium selection mechanism for a discrete, complete information game where all equilibria are computed explicitly.

Some of the most influential papers in the recent literature on games estimate structural parameters by only imposing best response or other stability conditions that are only necessary but not sufficient for determining the observed profile of actions (most importantly Pakes, Porter, Ho, and Ishii (2006) and Fox (2010)). In the framework of this paper, this approach can be loosely interpreted as estimation based on the upper probability, or alternatively, the most favorable equilibrium selection mechanism for generating the observed market outcome. A potential advantage for following a (computationally more cumbersome) procedure based on a full likelihood rather than a moment-based method is that it becomes more natural to incorporate a parametric model for unobserved heterogeneity and link the structural parameters in payoff functions to empirical counterparts like choice probabilities or substitution elasticities.

Logan, Hoff, and Newton (2008) estimate a model for a matching market, and Christakis, Fowler, Imbens, and Kalyanaraman (2010) use MCMC techniques to estimate a model of strategic network formation, but do not explicitly account for the possibility of multiple equilibria. The matching market model analyzed in this paper is different from that in Fox (2010) and Galichon and Salanié (2010) in that I do not assume transferable utilities, so that stable matchings do not necessarily maximize joint surplus across matched pairs. Echenique, Lee, and Shum (2010) and **?** consider inference based on implications of matching stability assuming that agents' types are discrete and fully observed by the econometrician. Pakes, Porter, Ho, and Ishii (2006), Baccara, Imrohoroglu, Wilson, and Yariv (2012) and Uetake and Watanabe (2012) estimate matching games via inequality restrictions on the conditional mean or median of payoff functions derived from necessary conditions for optimal choice, whereas the approach in this project aims to model the conditional distribution of payoffs. In many cases, this will require a parametric model for the distribution of unobserved heterogeneity, but in general knowledge of the full distribution of heterogeneity is necessary to compute policy-relevant counterfactuals (e.g. conditional choice probabilities) from estimated payoff parameters.

There is a vast literature on robustness for Bayesian decisions with respect to prior information, see e.g. Kudō (1967), Berger (1984), and Berger, Insua, and Ruggeri (2000) and references therein for an overview. Robust approaches to decisions under Uncertainty have also been used in prescriptive analysis of economic decisions and modeling of individual choice behavior, see e.g. Gilboa and Schmeidler (1989), Hansen and Sargent (2008), Strzalecky (2011), and references therein. We are going to cast the statistical problem of estimation and inference in the maxmin expected utility framework with multiple priors proposed by Gilboa and Schmeidler (1989) which can be interpreted as reflecting a group decision among agents with different prior beliefs, or representing true ambiguity - as opposed to risk - regarding the choice of a correct model for the observed data on the true state of nature. Kitagawa (2010) analyzes Gamma-posterior expected loss and Gamma-minimax decisions for partially identified models based on the posterior distribution for a sufficient parameter and an inverse mapping from this sufficient parameter to the "structural" parameters of interest. In contrast the approach taken in this study does not presuppose such a formulation for the statistical decision problem but sets up the decision problem by parameterizing explicitly the "completion" of the model without placing any restrictions on the prior over this auxiliary parameter.

While this paper is mainly aiming at providing computational tools, the question of optimal decisions in incompletely specified models is an important area of research of its own right. For example Manski (2000), and subsequently Stoye (2009) and Song (2009) consider settings that are not point-identified from a frequentist's perspective and analyze optimal point decisions and inference. Liao and Jiang (2010) imposed a specific prior on the unspecified component of the model. Moon and Schorfheide (2010) consider Bayesian inference when the decision maker has a prior for the parameter of interest, but maintains a set of models for the data-generating process. They point out that for large samples, Bayesian credible sets will typically be strict subsets of the identification region. Our procedure is going to share that feature since we assume a prior over the structural parameter.

The next section characterizes a class of statistical decision problems arising in estimation and inference with incomplete models, and discusses optimal statistical decisions using Gamma-posterior expected loss (GPEL). Section 3 proposes a generic algorithm for simulating the integral corresponding to GPEL and establishes consistency for GPEL-optimal decisions, and section 4 shows how to incorporate independence restrictions in the class of priors. In section 5 we illustrate the practical usefulness of our method with an empirical analysis of mate preference parameters based on marriage outcomes in a two-sided matching market. Section 6 concludes.

2. Equilibrium Selection and Likelihood

This paper considers structural latent variable models that may be incomplete in the sense that for some states there is no well-defined reduced form for the observable outcomes in that the model does not map each state to a unique observable outcome. We observe M instances of the game ("markets"), i.e. an element $y = (y_1, \dots, y_M)$ of the sample space $\mathcal{Y} = \mathcal{Y}_1 \times \dots, \mathcal{Y}_M$, where $y_m \in \mathcal{Y}_m$ contains information about players' characteristics and other payoff-relevant information as well as the actions chosen by the players in the mth instance of that game. More specifically we let the data be of the form $y_m = (s_m', x_m')$, where given the action space for the mth market $\mathcal{S}_m$, $s_m \in \mathcal{S}_m$ is the observed action profile in game m, and x_m is a vector of observed agent- and game-level covariates. For the purposes of this paper, we restrict our attention to the case in which $\mathcal{S}_m := \left\{ s_m^{(1)}, \dots, s_m^{(p_m)} \right\}$ is finite and denote the number of action profiles for the mth game with $p_m := \sharp \mathcal{S}_m$.

For the mth game, suppose there is a set $\mathcal{N}_m$ of n_m players, and we denote player i's payoff from action profile $s_m \in \mathcal{S}_m$ with $U_{mi}(s_m)$. We stack the vector of payoffs as $\mathbf{u}_m := (U_{m1}(s_1), \dots, U_{mn_m}(s_1), \dots, U_{m1}(s_{p_i}), \dots, U_{mn_m}(s_{p_i}))$ and denote the payoff space for the game $\mathcal{U}_m \subset \mathbb{R}^{n_m p_m}$. We assume a parametric model for the distribution of U_m

$$\mathbf{u}_m \sim g_m(u|x_m, \theta)$$

where $\theta \in \Theta$ is a k-dimensional parameter. Let $\Delta\mathcal{S}_m \subset [0,1]^{p_m}$ be the set of probability distributions over $\mathcal{S}_m$, so that an equilibrium selection rule for game m corresponds to a measurable map

$$\lambda_m : \begin{cases} \mathcal{U}_m & \to & \Delta\mathcal{S}_m \\ u & \mapsto & \lambda_m(\mathbf{u}_m) = \left(\lambda_m(\mathbf{u}_m, s_m^{(1)}), \dots, \lambda_m(\mathbf{u}_m, s_m^{(p_i)})\right)' \end{cases}$$

Denoting the set of Nash equilibria for a given payoff profile $\mathbf{u}_m$ with $\Sigma_m^*(\mathbf{u}_m) \subset \Delta\mathcal{S}_m$, the parameter space for $\lambda := (\lambda_1, \dots, \lambda_M)'$ is given by

$$\Lambda := \left\{ \lambda : \lambda_m(\mathbf{u}_m, s_m^{(p)}) = 0 \text{ if } s_m^{(p)} \notin \text{support } \Sigma_m^*(\mathbf{u}_m), \text{ for all } u \in \mathcal{U}_m, \ p = 1, \dots, p_m, \text{ and } m = 1, \dots, M \right\}$$

Then the likelihood of $y = y_1, \ldots, y_m$ can be written as

$$f(y|\theta, \lambda) = \prod_{m=1}^{M} \left(\int_{\mathcal{U}_m} \lambda_m(\mathbf{u}, s_m) g_m(u|x_m, \theta) d\mathbf{u} \right) h(x)$$

where $h(x)$ is the joint density of $x_1, \ldots, x_M$. The resulting family of distributions of y is indexed by two parameters, the parameter of interest $\theta \in \Theta$ which will be taken to be a k-dimensional vector, and $\lambda \in \Lambda$, a general parameter space that depends on the particular problem at hand. Hence for any fixed value of λ we have a fully parametric likelihood, i.e. there is a set of distributions that is indexed with (θ, λ),

$$y \sim f(y|\theta, \lambda) \quad \theta \in \Theta, \lambda \in \Lambda \tag{2.1}$$

Note that in this formulation, there is no formal difference between the roles of θ and λ. However in the following, θ will be the parameter of interest for which we specify a prior that does not depend on λ, whereas λ will be treated as nuisance parameter for which we do not want to impose a specific prior, but only a number of restrictions on its parameter space Λ.

In order to develop the main ideas, we will first consider the bivariate game with stochastic payoffs developed in Tamer (2003).

Example 2.1. (Bivariate Game) *Suppose there are two players, each of whom can choose from two actions, $s_1 \in \{0, 1\}$ and $s_2 \in \{0, 1\}$, and players' payoffs are given by $u_i(s_1, s_2) = (x_i'\beta + \Delta_i s_{-i} + \varepsilon_i)s_i$ for $i = 1, 2$, where $(\varepsilon_1, \varepsilon_2) \sim N(0, I)$. It is possible to verify, along the lines of the argument in Tamer (2003), that for $\Delta_1\Delta_2 \leq 0$, the equilibrium is unique for any realization of $(\varepsilon_1, \varepsilon_2)$, whereas if $\Delta_1\Delta_2 > 0$, there is a rectangular region in the support of the random shocks for which there are three equilibria.*[1] *If equilibrium selection is assumed not to depend on payoffs and covariates, Λ is the probability simplex $D^3 := \left\{(\lambda_1, \lambda_2, \lambda_3) \in [0, 1]^3 : \sum_{k=1}^{3} \lambda_k = 1\right\}$. More generally, an equilibrium selection rule will be given by $\lambda : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathbb{R}^2 \to D^3$, specifying the probability of either of the pure equilibria for each payoff profile in the region of multiplicity for the cases $\Delta_1, \Delta_2 > 0$ and $\Delta_1, \Delta_2 < 0$, respectively.*

2.1. **Upper and Lower Likelihood.** Of particular interest are the *upper* and the *lower likelihoods* of the sample y, which we define as

$$f^*(y|\theta) := \sup_{\lambda \in \Lambda} f(y|\theta, \lambda), \quad \text{and } f_*(y|\theta) := \inf_{\lambda \in \Lambda} f(y|\theta, \lambda)$$

respectively. Furthermore let $\lambda_*(\theta; y) \in \arg\min_{\lambda \in \Lambda} f(y|\theta, \lambda)$ and $\lambda^*(\theta; y) \in \arg\max_{\lambda \in \Lambda} f(y|\theta, \lambda)$. Since the dimension of Λ may be infinite, these extrema need not always exist, but we will

[1] If $\Delta_1, \Delta_2 > 0$, the region of multiplicity takes the form of "battle of the sexes" with the pure equilibria $(0, 0), (1, 1)$ and a mixed equilibrium. For $\Delta_1, \Delta_2 < 0$, the region of multiplicity consists of games that are strategic equivalents of "chicken" with pure equilibria $(0, 1), (1, 0)$ and one mixed equilibrium.

show how to construct the upper and lower likelihoods for some classes of problems, and assume existence for our formal results.

Example 2.2. *We will now use the bivariate game from Example 2.1 to illustrate our general approach to simulating upper and lower probabilities in latent variable models with set-valued predictions. As figure 2.2 illustrates, the distribution of the latent states conditional on the observed outcome changes according to how nature selects among the multiple predictions of the model if the latent utilities fall into the intersection of the shaded areas corresponding to the outcomes* $(0,1)$ *and* $(1,0)$*.*

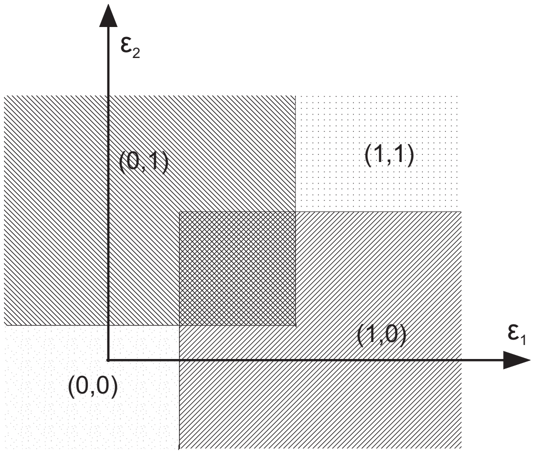

Figure 1. Pure Nash equilibria in the bivariate discrete game

If we leave the equilibrium selection rule across the n *instances of the game unrestricted, then we can obtain the bounds on the likelihood by considering only the two extremal points of the set* Λ*: (i) the "most favorable" selection mechanism in which nature produces the observed outcome if it constitutes one out of possible many equilibria given the realization of* $(\varepsilon_1, \varepsilon_2)$*, and (ii) the "least favorable" selection rule in which nature chooses any other possible equilibrium over the observed outcome. E.g. if our data set records the outcome* $(1,0)$*, then a simulation rule based on the first selection rule allows for all points in the state space for which* $(1,0)$ *is a Nash equilibrium (i.e. the shaded rectangle in the left graph in Figure 3), whereas in the second case we only consider states for which* $(1,0)$ *is the* unique *Nash equilibrium, excluding the region of multiplicity.*

This example shows that instead of having to consider the class of all possible selection mechanisms which may be difficult to parameterize, it is in many cases sufficient to simulate from distributions that correspond to the lower and the upper probabilities for the observed outcomes, which correspond to the probability bounds derived in Ciliberto and Tamer (2009). Clearly, this particular problem is fairly basic, and there is no compelling reason to favor the use of simulation techniques over explicit evaluation of the relevant choice probabilities

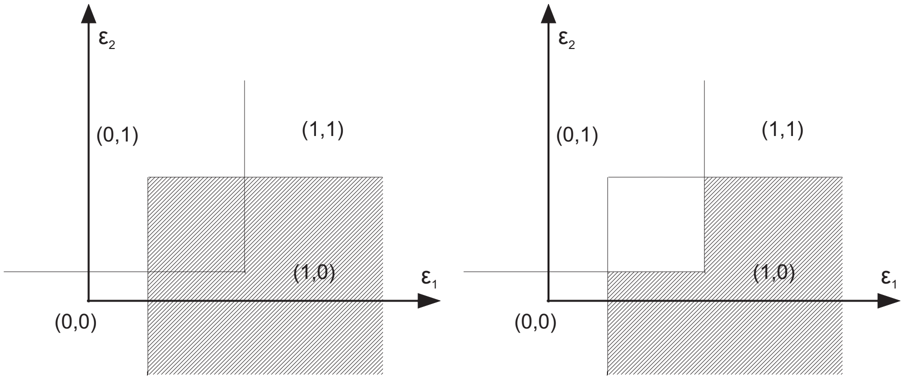

FIGURE 2. Most and least favorable equilibrium selection with respect to the outcome $(1, 0)$ given the unobserved state

in this case - e.g. Ciliberto and Tamer (2009) estimated a six-player version of this game for the US airline market using a minimum distance approach.

Another way to interpret the role of the equilibrium selection mechanism in estimation is that the most favorable selection rule generates the observed outcome *whenever* the (necessary) Nash conditions hold. Hence the upper probability results from imposing the same set of stochastic restrictions that have been used in moment-based estimation of games, most importantly Pakes, Porter, Ho, and Ishii (2006) and Fox (2010), but also in work following a Bayesian approach Christakis, Fowler, Imbens, and Kalyanaraman (2010) and Logan, Hoff, and Newton (2008). Especially the last approach can be directly interpreted as conducting inference based on the concentrated likelihood with respect to the equilibrium selection rule.

2.2. **Exchangeability.** We will now discuss a particular type of restrictions on the likelihood that reflect a particular kind of symmetry among different components of the observable data from the econometrician's point of view. For one, the identity of individual agents in a game may be unknown or irrelevant from the econometrician's point of view, so that the likelihood $f(y|\theta, \lambda)$ should be invariant under permutations of, or within subsets of, the set of agents for each instance of the game. On the other hand, we may want to treat different instances $y_1, \dots, y_M$ of the game as exchangeable.[2]

More specifically, let $\mathcal{N} := \{(m, i) : m = 1, \dots, M; i = 1, \dots, n_m\}$ be the index set for the M games and n_m players in the mth market. Now let σ denote a permutation of $\mathcal{N}$, and we let Σ^* denote the set of permutations σ such that the model is invariant with respect to σ in the following sense:

Assumption 2.1. ***(Exchangeability)*** *There is a set of permutations Σ^* of the set $\mathcal{N}$ such that for all $\lambda \in \Lambda$ and $\sigma \in \Sigma^*$, there exists λ_σ such that $f(y|\theta, \lambda) = f(\sigma(y)|\theta, \lambda_\sigma)$.*

[2] Recall that for a random sample $Z = (Z_1, \dots, Z_n)$ with joint distribution $f(z_1, \dots, z_n)$, we say that z_i and z_j are exchangeable if $f(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_{j-1}, z_j, z_{j+1}, \dots, z_n) = f(z_1, \dots, z_{i-1}, z_j, z_{i+1}, \dots, z_{j-1}, z_i, z_{j+1}, \dots, z_n)$.

Note that this is an invariance property for the *family* of models $\{f(y|\theta,\lambda)\}_{\lambda\in\Lambda}$ as a whole, and is also substantially weaker than exchangeability for every $\lambda \in \Lambda$.

Many previous papers in the literature on estimation of games have assumed that the observed sample consists of i.i.d. draws of markets (see e.g. Ciliberto and Tamer (2009), Galichon and Henry (2011) and Beresteanu, Molchanov, and Molinari (2009)), which implies exchangeability of observed instances of the game. If the econometric analysis doesn't distinguish between individuals players' identities, then Σ^* contains all permutations of agents within a given market. For example, Bresnahan and Reiss (1991b) treat potential entrants in a game of market entry with Cournot competition as symmetric so that for the purposes of the econometric analysis, the number of market entrants is a sufficient statistic for the actual equilibrium played in the game.

In order to impose the symmetry implied by the permutations in the set Σ^* for any equilibrium selection rule $\lambda \in \Lambda$, our analysis will be based on the *invariant likelihood*, which we define as

$$f_{inv}(y|\theta,\lambda) := \frac{1}{|\Sigma^*|}\sum_{\sigma\in\Sigma^*} f(\sigma(y)|\theta,\lambda) = \frac{1}{|\Sigma^*|}\sum_{\sigma\in\Sigma^*}\int_{\mathcal{U}} \lambda(u;\sigma(s))g(u|\sigma(x),\theta)duh(x) \tag{2.2}$$

In practice the number of permutations $|\Sigma^*|$ can be very large, so that we may choose to approximate the average in (2.2) by an average over several random draws from a uniform distribution over Σ^*. We can now define the upper and lower invariant likelihood, respectively, as

$$f^*_{inv}(y|\theta) = \sup_{\lambda\in\Lambda} f_{inv}(y|\theta,\lambda), \text{ and } f_{inv,*}(y|\theta) = \inf_{\lambda\in\Lambda} f_{inv}(y|\theta,\lambda) \tag{2.3}$$

In particular, the upper invariant likelihood is given by

$$f^*_{inv}(y|\theta) = \frac{1}{|\Sigma^*|}\int_{\mathcal{U}} v(u,\sigma,y)du$$

where for $\tilde{S}(s,\Sigma^*) := \{s \in \mathcal{S} : \exists\sigma : \sigma(\tilde{s}) \in \mathcal{S}^*(u)\}$,

$$v(u,\sigma,y) := \max_{\tilde{s}\in\mathcal{S}(s,\Sigma^*)} \sum_{\sigma:\sigma(s)=\tilde{s}} g(u|\sigma(x),\theta)\mathbb{1}\{\sigma(s) \in S^*(u)\}$$

, and we take the minimum and the maximum over an empty set to be equal to zero since if u does not support the equilibrium s, it does not contribute to the likelihood. Hence if for all $u \in \mathcal{U}$ we have $\mathcal{S}(u) \subset \{\sigma(s) : \sigma \in \Sigma^*\}$, and in the absence of agent-specific covariates, the invariant likelihood ratio between the most and least favorable model is equal to 1 so that the invariant likelihood $f_{inv}(y|\theta,\lambda)$ is the same for all $\lambda \in \Lambda$ despite the multiplicity of equilibria.

For example in the symmetric entry game considered by Bresnahan and Reiss (1991b), for every payoff profile u the number of entrants is the same for all equilibria in $\mathcal{S}^*(u)$ implying

$\mathcal{S}(u) \subset \{\sigma(s) : \sigma \in \Sigma^*\}$. Furthermore there are market-level but no firm-specific characteristics, so that the invariant likelihood is uniquely defined and inference is not conservative with respect to equilibrium selection.

As in the baseline case, the maximization and minimization, respectively, in (2.3) can be done pointwise given the draws of u and the corresponding likelihood ratios $\frac{g(u|\sigma(x),\theta)}{g(u|x,\theta)}$, and for a Gaussian conditional distribution, this results in a constrained quadratic assignment problem which is in general known to be NP-hard but a solution to which can be approximated in polynomial time, and some special cases have solutions that are easier to compute.

Example 2.3. ***(Binary Action Game with Exchangeable Agents)*** *Let y_m be an observation of a game with n_m players, each of whom can choose an action $s_i \in \{0,1\}$. Suppose the difference in player i's payoff from choosing action $s_i = 0$ is normalized to zero, and her payoff from choosing $s_i = 1$ is given by $u_i = \mu(x_i,\theta) + \Delta \sum_{j\neq i} s_j + \varepsilon_i$, where $\varepsilon_i, i = 1,\dots,n_m$ are independent draws from a standard normal distribution. Then if Σ^* allows for all permutations of $\{1,\dots,n_m\}$, the permutations $\sigma^* := \arg\max_{\sigma(s)\in\mathcal{S}^*(u)} g(u|\sigma(x),\theta)$ and $\sigma_* := \arg\min_{\sigma(s)\in\mathcal{S}^*(u)} g(u|\sigma(x),\theta)$ are given by a (ascending and descending, respectively) assortative matching of $u_i - \Delta\sum_{j=1}^{n_m} s_j$ with the conditional means $\mu(x_i,\theta)$ subject to the Nash conditions. The computational complexity of sorting the observations is of the order $O(n_m \log n_m)$, and the cost of finding the optimal assignment based on an ordered list is linear in n_m.*

In section 5, we will describe a procedure for sampling from a posterior with respect to the upper and lower invariant likelihoods in more detail for the problem of structural estimation of matching markets.

The invariant likelihood ratio, or its logarithm

$$LR_{inv}(\theta) := \log f^*_{inv}(y|\theta) - \log f_{inv,*}(y|\theta)$$

is not only useful for the methods considered in this paper, but the likelihood ratio statistic $LR_{inv}(\theta)$ may be of interest for frequentist inference in settings for which it is not possible to concentrate out the equilibrium selection rule consistently - e.g. when only one realization of a game with a large number of players is observed. Note that by de Finetti's theorem, infinite exchangeable sequences can be represented as i.i.d. conditional on a common sigma-algebra $\mathcal{F}_\infty$, so that for a fixed selection rule λ conditional LLNs and CLTs can be obtained using martingale convergence theory, see Loève (1963). This sigma-algebra may in general contain information about the equilibrium selection rule, so it is not obvious whether exchangeability of individual players can justify inference based on a concentrated likelihood alone. However establishing procedures for construction of critical values and a frequentist asymptotic theory for inference based on $LR_{inv}(\theta)$ under appropriate exchangeability conditions is beyond the scope of this paper and will be left for future research.

There are two potential limitations to the approach in this paper: for one it may be very difficult to verify or impose uniqueness of an equilibrium for a given realization of payoffs. For games with strategic complementarities, uniqueness of a given equilibrium can be verified systematically by checking whether tâtonnement from both the supremum and the infimum of the strategy set converges to the same profile (see e.g. Theorem 8 in Milgrom and Roberts (1990)). For the two-sided matching market model considered in our empirical application, we can exploit the structure of the problem in a very similar fashion and check whether the stable matching is unique using the deferred acceptance algorithm (see section 4 of this paper). However in general an efficient implementation of our method will require a good understanding of the structure of the economic model to find such shortcuts.

The other potential pitfall of the procedure arises from the maximization step in the definition of (3.1): if the likelihood of the observed outcome, $f(y|\theta,\lambda)$, varies too much between the most and least favorable selection rule λ^* and λ_*, respectively, then the simulated approximation to GPEL will put nonzero weight only on a small fraction of the simulated sample of parameter values, so that the effective number of simulation draws can be very small. This is reflected in the requirement of Assumption 4.3 below, which restricts the range of likelihood ratios for different values of $\lambda \in \Lambda$ to be uniformly bounded in θ. If the probability of the observed equilibrium being unique is very small relative to the probability of the outcome being an equilibrium, the bounding constant in that assumption would have to be chosen very small and likely lead to a poor performance of the simulation algorithm.

Both issues are particularly relevant for "large" models of social interactions, but it should be expected that alternative approaches would be affected by these problems to a comparable degree.

3. Robust Decision Rules

We consider the problem of a decision-maker who, after observing a sample y, chooses an action from a set $\mathcal{A}$, and who is concerned whether a statistical decision is robust with respect to her beliefs about the model incompleteness $\lambda \in \Lambda$. The literature on Bayes decisions has proposed a number of optimality and evaluative criteria that capture various notions of local and global robustness (see e.g. Berger (1985) for a discussion). The main focus of this paper is on a class of statistical decision rules that minimize *Gamma-posterior expected loss*, a modification of Bayes risk that is globally robust with respect to the model incompleteness. This approach will allow us to exploit some computational advantages of Bayes procedures without imposing priors on how the model incompleteness is resolved in the observed sample.

We will now develop the decision-theoretic framework motivating the main simulation procedure, and we will in general follow the exposition from Ferguson (1967); for a general treatment of the standard Bayesian framework, see also Berger (1985): The *action space* $\mathcal{A}$ is the set of actions available to the decision maker - for example, for the construction of confidence sets for the parameter θ the action space is a suitable collection of subsets of Θ, or for the estimation of lower bounds for a component of θ, $\mathcal{A}$ is a subset of the real line. A (deterministic) *decision rule* $d : \mathcal{Y} \to \mathcal{A}$ prescribes an action a for every point y in the sample space.

Finally we have to choose a *loss function* $L : \Theta \times \mathcal{A} \to \mathbb{R}$ to evaluate an action $a \in \mathcal{A}$ if the true parameter is θ. We will assume that the decision maker's objective does not depend on the "true" value of the nuisance parameter λ. I will now give a few examples of loss functions corresponding to some important statistical problems in estimation of bounds and set inference.

Bounds on the Posterior Mean. Let $b \in \mathbb{R}^l$. For an l-tuple $r = (r_1, \dots, r_l)$ of indices in $\{1, \dots, k\}$, let $\theta_{(r)}$ be the subvector $(\theta_{r_1}, \dots, \theta_{r_l})'$ of components of θ corresponding to the indices in r. Then the loss function corresponding to a joint lower bound for the posterior mean of the subvector $\theta_{(r)}$ is

$$L(\theta, b) = \|\theta_{(r)} - b\|_-^2$$

where $\|x\|_-$ denotes the norm of the (component-wise) negative parts of x, $\|x\|_- := \sum_{i=1}^{\dim(x)} |\min\{x_i, 0\}|$

Quadratic Loss. Given a prior on θ we might want to find an optimal point decision on the true parameter θ_0 with respect to quadratic loss,

$$L(\theta, a) = (\theta - a)'W(\theta - a)$$

for a positive semi-definite matrix W.

Credible Sets. For a value $\alpha \in [0, 1]$, an $1 - \alpha$ Bayesian confidence set is a set $\mathcal{C}(1 - \alpha) \subset \Theta$ such that $1 - \alpha \leq P_n(\theta \in \mathcal{C}(1 - \alpha) | Y = y) = \int_{\mathcal{C}(1-\alpha)} p_n(\theta, \mathcal{C}(1 - \alpha)) d\theta$. The problem of constructing credible sets is associated with the loss function

$$L(\theta, \mathcal{C}(1 - \alpha)) = 1 - \mathbb{1}\{\theta \in \mathcal{C}(1 - \alpha)\} = \begin{cases} 0 & \text{if } \theta \in \mathcal{C}(1 - \alpha) \\ 1 & \text{otherwise} \end{cases}$$

From a more conceptual perspective, we can also apply our approach to loss functions that are direct measures of the (negative of the) welfare effect of a policy recommendation a if the true state of nature is θ. However, for the purposes of this paper, we will primarily consider statistical decision problems related to estimation and inference.

For a given loss function, we can define the *risk function* as the expected loss from following a decision rule $d(y)$ for each value of θ,

$$R(\theta, \lambda; d) := \mathbb{E}_{\theta,\lambda}\left[L(\theta, d(Y))\right] = \int_{\mathcal{Y}} L(\theta, d(y)) f(y|\theta; \lambda) dy$$

where the expectation is taken over a random variable Y with realizations in $\mathcal{Y}$ which are distributed according to (2.1).

It will in general not be possible to rank decision rules unambiguously according to $R(\theta, \lambda; d)$ unless we assume a particular value for the parameter θ, λ. We will therefore define a new criterion resulting from taking a suitable average over the parameter space.

More specifically, suppose that the marginal prior with respect to θ is given by $\pi_0(\theta)$, and there is a set Γ of joint priors $\pi(\theta, \lambda)$ for θ and λ whose marginals with respect to λ equal $\pi_0(\theta)$, i.e.

$$\Gamma := \left\{ \pi(\theta, \lambda) \in \mathcal{M}_{\theta,\lambda} : \int \pi(\theta, d\lambda) = \pi_0(\theta) \ \pi_0\text{-a.s. in } \theta \right\}$$

where $\mathcal{M}_{\theta,\lambda}$ denotes the set of probability measures over θ, λ. Our decision theoretic framework will not presuppose one single prior $\pi(\theta, \lambda)$ but allow for a *set of priors* that the decision maker all regards as equally "reasonable" but does not want to assign respective probabilities.

3.1. **Gamma-Posterior Expected Loss.** Gamma-posterior expected loss (GEPL) is motivated as a robust version of Bayes average risk over the family of priors, Γ. Let $\pi(\theta, \lambda) \in \Gamma$ be a prior distribution for (θ, λ). Then we can define the *average risk* with respect to π as

$$r^*(d, \pi) := \int_{\Theta \times \Lambda} R(\theta, \lambda; d) d\pi(\theta, \lambda)$$

By a change of variables

$$\begin{aligned} r^*(d, \pi) &= \int_{\Theta \times \Lambda} \int_{\mathcal{Y}} L(\theta, d(y)) f(y|\theta, \lambda) dy d\pi(\theta, \lambda) \\ &= \int_{\mathcal{Y}} \int_{\Theta \times \Lambda} L(\theta, d(y)) f(y|\theta, \lambda) d\pi(\theta, \lambda) dy \end{aligned}$$

so that for a given value of y, the average-risk optimal decision $d(y)$ with respect to a particular prior $\pi(\theta, \lambda)$ solves

$$d^*(y, \pi) := \arg\min_{a \in \mathcal{A}} \int_{\Theta \times \Lambda} L(\theta, a) f(y|\theta, \lambda) \pi(d\theta, d\lambda) \equiv \arg\min_{a \in \mathcal{A}} \int_{\Theta \times \Lambda} L(\theta, a) p(d\theta, d\lambda; y, \pi)$$

where

$$p(\theta, \lambda; y, \pi) := \frac{f(y|\theta, \lambda) \pi(\theta, \lambda)}{\int f(y|\theta, \lambda) d\pi(\theta, \lambda)}$$

is the joint *posterior* for (θ, λ) given y and prior π. In the following we denote

$$\mathcal{Q} := \left\{ p(\theta) : \frac{\int f(y|\theta, \lambda) \pi(\theta, d\lambda)}{\int \int f(y|\theta, \lambda) d\pi(\theta, \lambda)} : \pi \in \Gamma \right\}$$

the set of posteriors obtained after updating each prior in the set Γ.

We now define the main criterion for evaluating statistical procedures in the context of this study:

Definition 3.1. *The* Gamma-Posterior Expected Loss *(GPEL) of an action $a \in \mathcal{A}$ at a point $y \in \mathcal{Y}$ in the sample space is given by*

$$\varrho(y, a, \Gamma) := \sup_{\pi(\theta,\lambda)\in\Gamma} \frac{\int_{\Theta\times\Lambda} L(\theta, a) f(y|\theta; \lambda) d\pi(\lambda, \theta)}{\int_{\Theta\times\Lambda} f(y|\theta; \lambda) d\pi(\lambda, \theta)} \tag{3.1}$$

Here we adopt the terminology from the literature on robustness in Bayes decisions, in which "Gamma-" refers to the class of priors Γ - usually denoted by Λ - rather than the parameter space Λ in the notation of this paper. In accordance with the literature, we will also refer to Gamma-posterior expected loss as conditionally minimax (i.e. minimax given the sample $y \in \mathcal{Y}$) with respect to the set of priors Γ.

We will show in section 3 that under the assumptions of this paper, the supremum in (3.1) is attained at one particular element in Γ which we will call the *least favorable prior* which we will denote by $\pi^*(\theta, \lambda; a, y)$. It is important to note that the least favorable prior will in general depend on the loss function and the particular action $a \in \mathcal{A}$ and the observed point in the sample space, $y \in \mathcal{Y}$. In particular, minimization over the set Γ gives rise to a dynamic inconsistency: the prior implicitly chosen in the calculation of GPEL, i.e. conditional on the sample, will in general be different from the (unconditional) minimax decision with respect to Γ which minimizes average risk $r^*(d, \pi)$ over $\pi \in \Gamma$.

This dynamic inconsistency does not arise in the framework considered by Kitagawa (2010) who considers decisions minimizing GPEL and Gamma-minimax risk under a smaller class of priors $\tilde{\Gamma}$ for which the sampling distribution $f(y, \pi) := \int f(y|\theta, \lambda) d\pi(\theta, \lambda)$ is held fixed across $\pi \in \tilde{\Gamma}$. We prefer not to impose such a restriction on the set of priors, so that the procedures arising from minimizing GPEL will in general not have minimax properties, but be more conservative with respect to the model incompleteness $\lambda \in \Lambda$.[3] However, if the parameter θ is sufficient for the distribution of y, GPEL-optimal actions will also be Gamma-minimax by the same arguments as in Kitagawa (2010).

There are two main reasons why we choose to focus on Gamma-posterior loss rather than (unconditional) Gamma-minimax risk as a decision criterion: for one, in the Bayesian approach it is very natural to consider a rule that is robust given the observed sample, and therefore the researcher's information set, rather than considering a notion of "ex-ante" robustness with respect to a sampling experiment. There is also a direct connection to the Maximin Expected Utility framework for decisions under ambiguity analyzed by Gilboa and Schmeidler (1989). The other advantage of a conditional minimax rule is computational

[3] Note that by Jensen's Inequality,

$$\begin{aligned}
\sup_{\pi\in\Gamma} r^*(d, \pi) &:= \sup_{\pi\in\Gamma} \left\{ \int_{\mathcal{Y}} \int_{\Theta\times\Lambda} L(d(y), \theta) p(d\theta, d\lambda; y, \pi) dy \right\} \\
&\leq \int_{\mathcal{Y}} \sup_{\pi\in\Gamma} \left\{ \int_{\Theta\times\Lambda} L(d(y), \theta) p(d\theta, d\lambda; y, \pi) \right\} dy = \int_{\mathcal{Y}} \varrho(y, d(y), \Gamma) dy
\end{aligned}$$

in that the conditional Gamma-minimax rule is determined by extremal points of Λ that are relatively easy to find in many practically relevant cases as we will argue below. In particular, conditional Gamma-minimax does not require averaging over the sample space $\mathcal{Y}$ which makes it easier to adapt simulation algorithms from (non-robust) Bayesian statistics.

The main motivation for the Monte Carlo algorithms in this paper is to provide computationally tractable uniform approximations to the GPEL objective that can be used to determine the optimal action given the data at hand. More broadly, the procedures analyzed in section 3 can be used to simulate Choquet integrals over a capacity whose core is given by the posteriors obtained from updating the class of priors Γ element by element.

In addition to decision problems defined by minimization of GPEL for some loss function, we can extend the same ideas to problems in which GPEL plays the role of a size constraint, as e.g. in the following definition of a Gamma-posterior credible set for the parameter θ:

Definition 3.2. *We define a Gamma-posterior* $1-\alpha$ *credible set with respect to the family of measures* $\mathcal{Q}$ *as* $\mathcal{C}^*(1-\alpha) \subset \Theta$ *such that* $\inf_{Q\in\mathcal{Q}} Q\left(\theta \in \mathcal{C}^*(1-\alpha)\right) \geq 1-\alpha$.

For a Gamma-posterior $(1-\alpha)$-credible set we have that the Gamma-posterior expected loss

$$\varrho(y, \mathcal{C}(1-\alpha), \Gamma) = \int L(\theta, \mathcal{C}(1-\alpha)) dT(\theta, \mathcal{Q}) = \inf_{Q\in\mathcal{Q}} \int_{\mathcal{C}(1-\alpha)} dP = \inf_{Q\in\mathcal{Q}} Q\left(\theta \in \mathcal{C}(1-\alpha)\right)$$

To compare this to the frequentist literature on bounds, it should be pointed out that since we do impose a prior over the parameter θ, for very "informative" samples credible sets are going to be strict subsets of the identification region with large probability, whereas minimax confidence sets are constructed to cover the entire identification region with a pre-specified probability. On the other hand, as pointed out above, GPEL-optimal decisions are conditionally minimax, and therefore more conservative with respect to the model incompleteness $\lambda \in \Lambda$ than a frequentist unconditional minimax rule.

3.2. **Capacities and Choquet Integrals.**

The concept of Gamma-posterior expected loss introduced before can be linked to the notion of the Choquet integral for capacities (for a general reference, see Molchanov (2005)) and is a direct adaptation of the framework in Gilboa and Schmeidler (1989) for decisions under ambiguity. This subsection gives a brief discussion of the links to these two literatures and is not essential for the exposition of our computational results.

Definition 3.3. (Upper and Lower Expectation) *Let* $f : \Theta \to \mathbb{R}_+$ *be a nonnegative function. We then define the* lower expectation *of* $f(\theta)$ *with respect to the family of distributions* $\mathcal{Q}$ *over* θ *as*

$$\int f dT(\theta, \mathcal{Q}) := \inf_{Q\in\mathcal{Q}} \int f(\theta) Q(d\theta) = \inf_{Q\in\mathcal{Q}} \mathbb{E}_Q\left[f(\theta)\right]$$

Similarly we define the upper expectation *of* $f(\theta)$ *over* $\mathcal{Q}$

$$\int f dC(\theta, \mathcal{Q}) := \sup_{Q \in \mathcal{Q}} \int f(\theta) Q(d\theta) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[f(\theta) \right]$$

The representation of Gamma-posterior expected loss as a sub-additive monotone capacity functional follows from Lemmata 3.2 and 3.3 in Gilboa and Schmeidler (1989). On a technical note, the lower and upper expectations are Choquet integrals with respect to the capacities T and C, and can be represented as hitting and containment functionals, respectively, of a compact random subset of Θ. Hence, GPEL is a Choquet integral with respect to a convex capacity C that corresponds to the infimum over the class $\mathcal{Q}$.

While a capacity $T(K)$ is defined over all compact subsets $K \subset \Theta$, for the computation of the Choquet integral of a nonnegative function $f(\theta)$, we only need to determine capacity on level sets of the Choquet integrand. Furthermore, by proposition 5.14 in Molchanov (2005) there exists a proper probability measure $p(\theta; f)$ in the core of T_X such that

$$\int f dT_X = \int f(\theta) p(\theta; f) d\theta$$

where the choice of that probability measure depends on the integrand.

Since the core of the capacity in the choice model consists of the posteriors for θ for a family of priors with respect to λ, given a action $a \in \mathcal{A}$ there exists a prior $\pi^*(\theta, \lambda; a)$ such that the Choquet integral of $L(d; \theta)$ can be represented as the expectation of $L(\cdot, a)$ over the posterior $p(\theta; y, \pi^*(\theta, \lambda; a))$.

3.3. **Generic Algorithm.** The main difficulty in simulating the integral in (3.1) consists in that the posterior

$$p^*(\theta; a, y) = \frac{\pi_0(\theta) \int f_Y(y|\theta, \lambda) \pi^*(d\lambda|\theta; y, a)}{\int \int f_Y(y|\tilde{\theta}, \lambda) \pi^*(d\lambda|\tilde{\theta}; y, a) \pi_0(\theta) d\theta}$$

results from a minimization problem and therefore depends on the Choquet integrand $L(\cdot, a)$.

Therefore, instead of sampling directly from $p^*(\theta; a, y)$, we will draw an initial Markov chain from the posterior $p(\theta, \lambda; y, \pi_I)$ where $\pi_I := \pi_I(\theta, \lambda; y)$ is an appropriately chosen (and possibly data-dependent) "instrumental" prior over $\Theta \times \Lambda$ - see e.g. Robert and Casella (2004) on techniques for sampling from a posterior distribution.

We can then rewrite the least favorable posterior of θ for action a and given the sample y in terms of the "instrumental" posterior $p(\theta, \lambda; y, \pi_I)$ using the Radon-Nikodym derivative of the least favorable prior with respect to the instrumental prior as importance weights

$$p^*(\theta; y, a) = \int \frac{\pi^*(\theta, \lambda; a)}{\pi_I(\theta, \lambda; y)} p(\theta, d\lambda; y, \pi_I) \Bigg/ \left(\int \int \frac{\pi^*(\tilde{\theta}, \lambda; a)}{\pi_I(\tilde{\theta}, \lambda; y)} p(\tilde{\theta}, d\lambda; y, \pi_I) d\tilde{\theta} \right)$$

Hence the least favorable prior $\pi^*(\theta,\lambda;a)$ solves

$$\sup_{\tilde{\pi}(\theta,\lambda)} \int L(\theta;a)\frac{\tilde{\pi}(\theta,\lambda)}{\pi_I(\theta,\lambda;y)}p(\theta,d\lambda;y,\pi_I) \quad \text{s.t.} \quad \int \tilde{\pi}(d\lambda|\theta)=1$$
$$\text{and} \quad \tilde{\pi}(\lambda|\theta)\geq 0$$

For every value of $a\in\mathcal{A}$, this is a linear programming problem that can be solved using standard algorithms.

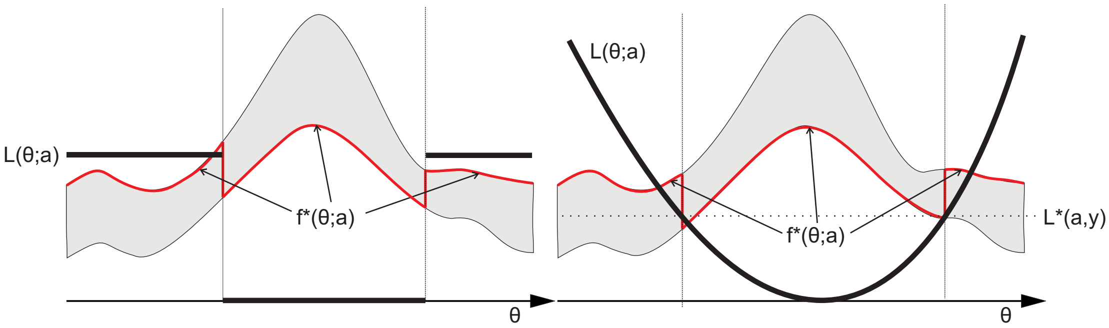

FIGURE 3. Choice of the least favorable model for zero-one loss (left) and quadratic loss (right).

As we will show in section 4, the solution to the problem (3.2) is attained at "extremal points" of the set of priors Γ. More specifically under our assumptions the least favorable prior conditional on θ vanishes on Λ except for the values of λ that maximize or minimize the likelihood $f(y|\theta,\lambda)$. The construction of the least favorable prior will follow a simple cut-off rule on the loss function $L(\theta,a)$, where the prior for θ is updated using the smallest value of the conditional likelihood for values of θ with $L(\theta,a)$ below the cutoff, and according to the largest value when $L(\theta,a)$ is above that cutoff - see figure 1 for a graphical illustration.

4. Simulation Algorithm

In this section, we will discuss an algorithm for computing Gamma-posterior expected loss and optimal actions. The formal results in this section will be proven in the appendix.

4.1. Main Assumptions.

We will first impose conditions on the underlying decision problem in order for Gamma-posterior expected loss to be well-defined and to ensure that simulated integrals converge in probability to their expectations uniformly in the action $a\in\mathcal{A}$ as the size of the simulated sample increases.

In the following, we will take the measure space for θ to be $(\Theta,\mathcal{B},\pi_0)$ for one fixed prior $\pi_0(\theta)$ on Θ, where $\mathcal{B}$ denotes the Borel σ-algebra on Θ. We want to avoid imposing any special structure on the parameter set Λ, but we also assume that we can construct a suitable measure space $(\Lambda,\mathcal{A},\mu)$ depending on the nature of the problem - e.g. if Λ denotes the set of (possibly stochastic) equilibrium selection rules in a finite discrete game, we can take $\mathcal{A}$ to be the Borel algebra on the corresponding probability simplex. For joint distributions over (θ, λ), we consider the product measure space $(\Theta \times \Lambda, \mathcal{B} \otimes \mathcal{A}, \pi_0 \times \mu)$, and we let $\mathcal{M}_{\theta,\lambda}$ and $\mathcal{M}_\lambda$ denote the set of probability measures on $\Theta \times \Lambda$ and Λ, respectively.

Assumption 4.1. ***(Loss Function)*** *The set of actions $\mathcal{A}$ is compact. Furthermore, for every value of a the loss function is bounded, quasi-convex, and measurable in θ, and the set $\mathcal{L} := \{L(a, \cdot) : a \in \mathcal{A}\}$ is a Glivenko-Cantelli class for distributions that are absolutely continuous with respect to the Lebesgue measure on Θ.*

By standard arguments on smooth classes of functions, Assumption 4.1 allows for the quadratic and "one-sided" loss functions with $\mathcal{A} = \Theta$ compact. However, if we consider zero-one loss for the construction of confidence sets, the requirement that the class of loss functions $\mathcal{L}$ is Glivenko-Cantelli limits the complexity of the family of sets $\mathcal{C}$ that we can evaluate. From a result by Eddy and Hartigan (1977), the class of indicator functions for convex subsets of Θ is Glivenko-Cantelli for measures that are absolutely continuous with respect to Lebesgue measure. If in addition we need the class of sets to be Donsker, we either have to restrict our attention to problems for which $\mathcal{C} \subset \mathbb{R}^2$, or all confidence sets satisfy additional smoothness restrictions, e.g. if we only consider ellipsoids whose boundaries have bounded derivatives up to any order - see e.g. Corollary 8.2.25 in Dudley (1999).

The GPEL-optimal decision depends crucially on the set of priors maintained by the decision maker. For our purposes we choose a class of priors that is rather large and does not restrict the possible forms of dependence between θ and λ. This feature is crucial for the "bang-bang" solution to the maximization problem (3.1), but I will argue below that this choice is also justified on normative grounds in many economic applications.

Assumption 4.2. ***(Class of Priors)*** *The marginal prior over θ is fixed at $\pi_0(\theta)$ which is absolutely continuous with respect to Lebesgue measure on Θ. Furthermore, the class of priors over (θ, λ) is given by*

$$\Gamma := \left\{ \pi(\theta, \lambda) \middle| \int \pi(\theta, d\lambda) = \pi_0(\theta) \ \pi_0\text{-a.s. in } \theta \right\}$$

In particular, there are no further restrictions on the conditional prior distribution $\pi_{\lambda|\theta}(\lambda|\theta)$.

This restriction on the class of priors is motivated by the sharp distinction between the roles of the parameters θ and λ in our model: in most problems the parameters θ will correspond to features of preferences, profit or cost functions or other economic fundamentals that the researcher may well have other prior knowledge about from other sources.

In contrast, the lack of restrictions on the conditional prior for λ given θ reflects that we want to be completely agnostic about that nuisance parameter.

As we will show, this lack of restrictions on Γ regarding the conditional priors over λ will also result in a great simplification of the computational problem in settings in which the parameter space Λ is very rich. E.g. for the marriage market problem in the next section, the number of stable matchings increases exponentially in the size of the market, so that it will in general be difficult to determine the full set of stable matchings for a given set of preferences, let alone formulate plausible beliefs about the selection mechanism.

Because of the minimization step in (3.1), it is also necessary to restrict the behavior for the likelihood $f(y|\theta, \lambda)$ as we vary λ in order to guarantee uniform convergence of GPEL across values $a \in \mathcal{A}$.

Assumption 4.3. ***(Likelihood)*** *(i) The parameter set $\Theta \subset \mathbb{R}^k$ is compact, (ii) the p.d.f. $f(y|\theta, \lambda)$ is continuous, bounded from above uniformly in (θ, λ) for all $y \in \mathcal{Y}$, and bounded away from zero on a set $B \subset \Theta$ with $\int_B \pi_0(d\theta) > 0$. (iii) There exists a finite constant $c > 0$ such that $c < \frac{f(y|\theta,\lambda_1)}{f(y|\theta,\lambda_2)} < \frac{1}{c}$ for all $\theta \in \Theta$, $\lambda_1, \lambda_2 \in \Lambda$, and $y \in \mathcal{Y}$, and (iv)* $\inf_{\lambda \in \Lambda} f(y|\theta, \lambda)$ *and* $\sup_{\lambda \in \Lambda} f(y|\theta, \lambda)$ *are attained for π_0-almost all values of $\theta \in \Theta$ and for each $y \in \mathcal{Y}$.*

This condition will be used to ensure that the posterior $p(\theta; y, a)$ resulting from the maximization problem (3.1) is absolutely continuous with respect to the marginal distribution $\int p(\theta, \lambda; y, \pi_I) d\lambda$ for any loss function and instrumental prior $\pi_I(\theta, \lambda)$. If this condition is only satisfied for very small values of c, this will typically result in a small effective sample size (and therefore a large variance for simulated GPEL) for a given number of draws from the distribution.

In a simulation study, Bajari, Hong, Krainer, and Nekipelov (2006) find that in a particular non-cooperative static discrete game the probability of multiplicity of equilibria appears to decrease in the number of players even though the maximal number of equilibria across different payoff draws increases very fast. This would be a very favorable setting for our algorithm to work well, whereas in the marriage market problem considered in our application, the probability of uniqueness vanishes very fast as both sides of the market grow large.

In principle, the bound in Assumption could be weakened somewhat for any given loss function $L(\cdot, \cdot)$, but I prefer to formulate the restriction on the data-generating process independently of the decision problem.

Part (iv) of the assumption holds if Λ is compact, but we can also allow for more general cases. E.g. in Example 2.1 the supremum corresponds to the probability that the latent state satisfies the Nash conditions for y to be an equilibrium, whereas the infimum is given by the probability of y being the unique Nash equilibrium, so that both the infimum and the supremum are attained even though the set Λ of mappings from the observed and unobserved states to probability distributions over three outcomes is not compact.

In the following, we will denote the upper and lower likelihood of the sample y by

$$f_*(y|\theta) := \inf_{\lambda \in \Lambda} f(y|\theta, \lambda), \quad \text{and } f^*(y|\theta) := \inf_{\lambda \in \Lambda} f(y|\theta, \lambda)$$

respectively. By Assumptions 4.3, the minimum over λ always exists. Furthermore let $\lambda_*(\theta;y) \in \arg\min_{\lambda\in\Lambda} f(y|\theta,\lambda)$ and $\lambda^*(\theta;y) \in \arg\max_{\lambda\in\Lambda} f(y|\theta,\lambda)$. If the minimizer $\lambda_*(\cdot)$ (and maximizer $\lambda^*(\cdot)$, respectively) is not unique, we define λ_* and λ^* as any convenient choice from the set of minimizers.

Proposition 4.1. *Under Assumptions 4.1-4.2, for all $a \in \mathcal{A}$, and all $y \in \mathcal{Y}$ satisfying $\int f(y|\theta,\lambda)\pi_0(\theta)d\theta > 0$ for some λ there exists a solution to the maximization problem in (3.1). Furthermore Gamma-Posterior Expected Loss is of the form*

$$\varrho(a,y,\Gamma) = \frac{\int_\Theta L(\theta,a)\mathbb{1}_{\{L(\theta,a)\le L^*\}} f^*(y|\theta)d\pi_0(\theta) + \int_\Theta L(\theta,a)\mathbb{1}_{\{L(\theta,a)>L^*\}} f^*(y|\theta)d\pi_0(\theta)}{\int_\Theta \mathbb{1}_{\{L(\theta,a)\le L^*\}} f^*(y|\theta)d\pi_0(\theta) + \int_\Theta \mathbb{1}_{\{L(\theta,a)>L^*\}} f^*(y|\theta)d\pi_0(\theta)}$$

where $L^* := \varrho(a,y,\Gamma)$.

In particular the maximum over the set of priors Γ is achieved by a simple cut-off rule in $L(\theta,a)$. In order to solve the maximization problem (3.1) using one single initial random sample $(\theta_1,\lambda_1),\dots,(\theta_B,\lambda_B)$ generated by imposing a (possibly data-dependent) instrumental prior $\pi_I(\theta,\lambda;y)$, that prior does not need to have positive density on the full support of every member of the class of priors Γ, but only the priors corresponding to GPEL for the decision problem at hand.

Assumption 4.4. ***(Instrumental Prior)*** *The instrumental prior $\pi_I(\theta,\lambda;y)$ is chosen in a way such that for all $y \in \mathcal{Y}^n$ and actions $a \in \mathcal{A}$, $\pi^*(\theta,\lambda;a,y)$ is absolutely continuous with respect to $\pi_I(\theta,\lambda;y)$.*

Ideally, we would like to choose an instrumental prior for which the variance of the likelihood ratio $\frac{\pi^*(\theta,\lambda;a,y)}{\pi_I(\theta,\lambda;y)}$ is as small as possible, but in practice, such a prior is difficult to construct because we want to "recycle" the Markov chain for different Choquet integrands, and even for a given loss function and values for a and y, the form of the least favorable prior is in general difficult to guess beforehand. The marginal distribution of θ under $\pi_I(\theta,\lambda;y)$ doesn't have to coincide with the actual prior $\pi_0(\theta)$, but we will in fact argue that in practice the instrumental prior should be chosen to be flatter than $\pi_0(\theta)$.

From Proposition 4.1 it follows that an instrumental prior $\pi_I(\theta,\lambda;y)$ can be chosen independently of the loss function, so that in practice it is straightforward to impose Assumption 4.4 as long as Assumption 4.2 holds as well. More specifically, the following result establishes that under Assumption 4.2, this requirement can be satisfied by a prior that conditional on θ puts nonzero probability mass on the two values of λ that minimize and maximize the likelihood of the observed outcome y.

Proposition 4.2. *Suppose Assumptions 4.2 and 4.4 hold, then*

$$\varrho(y,a,\Gamma) = \sup_{\psi\in\Psi} \frac{\int L(\theta;a)\psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)}{\int \psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)}$$

where Ψ is the set of measurable functions $\psi : \Theta \times \Lambda \to \mathbb{R}_+$ such that $\int \frac{\psi(\theta,\lambda)}{\pi_0(\theta)} \pi_I(\theta, d\lambda; y) = 1$ for all $\theta \in \Theta$.

From the previous discussion, the interpretation of $\psi(\theta, \lambda)$ is that of a conditional likelihood ratio between the most pessimistic and the instrumental prior for λ given θ.

Note that we can solve this maximization problem for any Choquet integrand without having to compute the likelihood function $f(y|\theta, \lambda)$ and using one single sample of draws from the posterior corresponding to the instrumental prior $\pi_I(\theta, \lambda; y)$. This representation is therefore a key result for the derivation of the generic simulation algorithm proposed in this paper.

Finally, we assume that we have an algorithm which allows us to draw from the posterior distribution corresponding to the instrumental prior $\pi_I(\theta, \lambda; y)$. For the second version of this assumption recall that a Markov chain $(X_b) = (X_1, X_2, \dots)$ is called *irreducible* if for any non-null Borel set A the return time to A from any given initial state x is finite with strictly positive probability. Furthermore, the chain is called *Harris recurrent* if in addition A is visited by (X_b) infinitely many times with probability 1 from any initial state $x \in A$.[4]

Assumption 4.5. ***(Sampling Procedure under π_I)*** *The sample $(\theta_1, \lambda_1), \dots, (\theta_B, \lambda_B)$ consists of either (i) B i.i.d. draws from the posterior distribution $p(\theta, \lambda; y, \pi_I)$ given the instrumental prior $\pi_I(\theta, \lambda; y)$, or (ii) a Harris recurrent Markov chain with an invariant distribution equal to $p(\theta, \lambda; y, \pi_I)$.*

In a future version of this paper, I am going to state primitive conditions for validity of standard MCMC-based sampling techniques for this problem. The main case of interest will be that of a Markov chain with a transition kernel defined by a multi-stage Gibbs or Metropolis-Hastings sampler (or a hybrid), as e.g. for algorithms based on data augmentation.

4.2. **Algorithm 1.** In a first step, we obtain a sample $(\theta_1, \lambda_1), \dots, (\theta_B, \lambda_B)$ of B draws from the posterior $p(\theta, \lambda; y, \pi_I)$ corresponding to the instrumental prior $\pi_I(\theta, \lambda; y)$. Under Assumption 4.5, it is possible to draw a Markov chain whose marginal distribution is equal to $p(\theta, \lambda; y, \pi_I)$ using the Metropolis-Hastings algorithm or a Gibbs sampler, see Robert and Casella (2004) for a discussion.

For any scalar $\tilde{L} \in \mathbb{R}$ we can now define

$$\hat{J}_B(\tilde{L}, a) := \frac{\sum_{b=1}^{B} L(\theta_b, a) \left[\mathbb{1}_{\{L(\theta_b,a)\leq\tilde{L}, \lambda_b=\lambda_*(\theta_b)\}} + \frac{\pi_I(\lambda_*(\theta_b)|\theta_b)}{\pi_I(\lambda^*(\theta_b)|\theta_b)} \mathbb{1}_{\{L(\theta_b,a)>\tilde{L}, \lambda_b=\lambda^*(\theta_b)\}} \right] \frac{\pi_0(\theta_b)}{\pi_I(\theta_b, \lambda_*(\theta_b); y)}}{\sum_{b=1}^{B} \left[\mathbb{1}_{\{L(\theta_b,a)\leq\tilde{L}, \lambda_b=\lambda_*(\theta_b)\}} + \frac{\pi_I(\lambda_*(\theta_b)|\theta_b)}{\pi_I(\lambda^*(\theta_b)|\theta_b)} \mathbb{1}_{\{L(\theta_b,a)>\tilde{L}, \lambda_b=\lambda^*(\theta_b)\}} \right] \frac{\pi_0(\theta_b)}{\pi_I(\theta_b, \lambda_*(\theta_b); y)}} \tag{4.1}$$

[4] See Robert and Casella (2004) for precise definitions.

In the second step of the algorithm, we compute minmax average risk given the B draws under the instrumental prior by solving

$$\hat{\varrho}_B(y,a,\Gamma) = \sup_{\tilde{L}\in\mathbb{R}} \hat{J}_B(\tilde{L},a) \tag{4.2}$$

Taken together the minimization steps in (4.1) and (4.2) implicitly solve for the most pessimistic prior for evaluating average risk of a given data y by changing the importance weight of each draw relative to the instrumental prior $\pi_I(\theta,\lambda;y)$. The main simplification this approach gives us consists in only manipulating these importance weights as we change the Choquet integrand in (3.1) when comparing different actions $a, a' \in \mathcal{A}$ or solving several different GPEL minimization problems based on the same chain of parameter draws.

Given the regularity conditions assumed in this section, we can establish uniform consistency of simulated GPEL as we increase the number B of draws from the posterior distribution for the parameter:

Proposition 4.3. *Suppose Assumptions 4.1-4.5 hold, then* $\hat{\varrho}_B(y,a,\Gamma) \xrightarrow{p} \varrho(y,a,\Gamma)$ *uniformly in* a*.*

For the practical usefulness of this procedure, it is crucial that convergence in probability is uniform in order to ensure that the minimizer of $\hat{r}_B(y,a,\Gamma)$ converges in probability to one out of possibly several GPEL-optimal statistical decisions as $B \to \infty$.

4.3. **Algorithm 2.** As we will see in the empirical application in the section 5, in some cases it may be substantially harder to draw from the missing data distribution conditional on $\theta, \lambda_*(\theta)$ than under $\theta, \lambda^*(\theta)$, but it is feasible to simulate the likelihood ratio conditional on the augmented data. Suppose that there is a particular choice of $\lambda_0 \in \Lambda$ such that for $y = (s,x)$, we can sample from the (augmented) posterior distribution $p(\theta,u|y,\lambda) \propto \lambda(u;s)g(u|x,\theta)\pi_0(\theta)$, e.g. using a multi-stage Gibbs sampler, and the augmented distribution $f(y|u,\theta,\lambda)$ is absolutely continuous with respect to $f(y|u,\theta,\lambda_0)$ for all $\lambda \in \Lambda$ and u such that $y \in \mathcal{S}(u)$.

Then we can generate a sample of parameter draws and importance weights:

(1) draw a new value θ_b from the posterior arising from the prior $\pi_0(\theta)$ and the likelihood $f(y|\theta,\lambda_0)$.
(2) construct two auxiliary variables containing the likelihood ratios for the augmented sample,

$$\begin{aligned} v_b^* &:= v^*(\theta_b,u_b;y) = \frac{f^*(y|u_b,\theta_b)}{f(y|u_b,\theta_b,\lambda_0)} = \frac{\lambda^*(u;s)}{\lambda_0(u;s)} \\ v_{*b} &:= v_*(\theta_b,u_b;y) = \frac{f_*(y|u_b,\theta_b)}{f(y|u_b,\theta_b,\lambda_0)}\frac{\lambda_*(u;s)}{\lambda_0(u;s)} \end{aligned}$$

In the application in section 4, we will draw θ_b from the upper likelihood $f^*(y|\theta)$ and set the importance weights in step 2 equal to $v_b^* = 1$ and $v_{b,*} = 1$ if u_b satisfies the (deterministic) constraints under $\lambda_*(\theta)$, and let $v_{b,*} = 0$ otherwise. This procedure guarantees that, after marginalizing over the conditional missing data distribution, the acceptance probability is equal to the likelihood ratio without having to calculate or approximate it explicitly.

For any value of the scalar $\tilde{L} \in \mathbb{R}$ we can now define

$$\tilde{J}_B(\tilde{L}, a) := \frac{\sum_{b=1}^B L(\theta_b, a) \left[\mathbb{1}_{\{L(\theta_b,a)\leq\tilde{L}\}} v_*(\theta_b, u_b) + \mathbb{1}_{\{L(\theta_b,a)>\tilde{L}\}} v^*(\theta_b, u_b)\right]}{\sum_{b=1}^B \left[\mathbb{1}_{\{L(\theta_b,a)\leq\tilde{L}\}} v_*(\theta_b, u_b) + \mathbb{1}_{\{L(\theta_b,a)>\tilde{L}\}} v^*(\theta_b, u_b)\right]}$$

as for the standard algorithm, and maximize over $\tilde{L} \in \mathbb{R}$ to obtain

$$\tilde{\varrho}_B(y, a, \Gamma) = \sup_{\tilde{L}\in\mathbb{R}} \tilde{J}_B(\tilde{L}, a)$$

Again, under the same regularity conditions as for Algorithm 1, we obtain the following uniform consistency result:

Proposition 4.4. *Suppose Assumptions 4.1-4.5 hold, then* $\tilde{\varrho}_B(y, a, \Gamma) \xrightarrow{p} \varrho(y, a, \Gamma)$ *uniformly in* a*.*

I do not attempt to compare the two algorithms in terms of their statistical properties, but in many cases one of the two algorithms will be substantially easier to implement than the other.

5. Imposing Independence on Priors

Maximization of expected loss with respect the class of all possible joint priors for θ and λ may in many cases be more conservative than necessary to incorporate our lack of knowledge regarding the equilibrium selection rule into an estimation procedure. Instead we may be willing to restrict our attention to priors that are independent for θ and λ.

More specifically, in this section we are going to replace Assumption 4.2 with

Assumption 5.1. *The marginal prior over* θ *is fixed at* $\pi_0(\theta)$ *which is absolutely continuous with respect to Lebesgue measure on* Θ*. Furthermore, the class of priors over* (θ, λ) *is given by*

$$\Gamma_\perp := \{\pi(\theta, \lambda) = \pi_0(\theta)\nu(\lambda) : \nu \in \mathcal{M}_\lambda\}$$

To see why independence is a substantive restriction for the construction of Gamma posterior expected loss, consider the priors achieving the maximum in (3.1) under the assumptions of theorem: the maximum over Γ is achieved by a prior that put unit conditional mass on the equilibrium selection rules determining the upper and the lower likelihood, respectively, depending on the value of θ, and can in particular not be replicated by a fixed prior distribution over different values of $\lambda \in \Lambda$ since the equilibrium selection rule may not depend on θ. Hence the GEPL criterion with respect to $\Gamma_\perp$ is less conservative than for the prior class Γ.

Solving for GEPL under the prior class $\Gamma_\perp$ is computationally less straightforward in that it requires explicit optimization over the space of selection rules that are fixed across Θ rather than pointwise maximization and minimization of the likelihood function for each value of θ, the parameter of the payoff distribution. But as Proposition 5.1 below shows, the maximum over $\Gamma_\perp$ in (3.1) is attained at a prior which puts unit mass on one particular selection rule $\lambda^*(u)$ which can in turn be characterized by a cut-off rule in u. More specifically, define

$$Z(u, a) := \frac{\int_\Theta L(a, \theta) \prod_{m=1}^M g_m(u_m|x_m, \theta) d\pi_0(\theta)}{\int_\Theta \prod_{m=1}^M g_m(u_m|x_m, \theta) d\pi_0(\theta)} \tag{5.1}$$

for any $u \in \mathcal{U}$ and $a \in \mathcal{A}$. Then we have the following characterization of GPEL with respect to $\Gamma_\perp$:

Proposition 5.1. *Under Assumptions 4.1, 4.3 and 5.1, for all $a \in \mathcal{A}$, and all $y \in \mathcal{Y}$ satisfying $\int f(y|\theta, \lambda)\pi_0(\theta)d\theta > 0$ for some λ there exists a prior $\pi^*(\theta, \lambda; a, y) := \pi_0(\theta)\nu^*(\lambda; a, y) \in \Gamma_\perp$ that solves (3.1), where $\nu^*(\lambda; a, y)$ is degenerate and puts mass only on a deterministic selection rule λ^* for all $i = 1, \dots, n$ which may depend on $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. Furthermore, $\lambda^*(u)$ is characterized by a simple cutoff rule in u where $\prod_{m=1}^M \lambda_m^*(u, s_m) = \min_{\lambda \in \Lambda} \prod_{m=1}^M \lambda_m^*(u, s_m)$ if $Z(u, a) < Z^*$, and $\prod_{m=1}^M \lambda_m^*(u, s_m) = \max_{\lambda \in \Lambda} \prod_{m=1}^M \lambda_m^*(u, s_m)$ otherwise.*

An important consequence of Proposition 5.1 is that the problem of calculating Gamma posterior expected loss over $\Gamma_\perp$ can be solved as a classification problem in the payoff space, $\mathcal{U}$. The structure of this optimization problem and conditions for consistency are similar to the estimation of density contour clusters, see e.g. Polonik (1995) and references herein.

The function $\lambda^*(u)$ is fully determined by the known components of the problem, i.e. the loss function, the likelihood, and the prior distribution for θ. However calculating the least favorable selection rule directly involves solving some of the integration problems our procedure was supposed to sidestep in the first place, and will therefore in general not be practical in many relevant settings. Instead we consider solving the maximization problem in (3.1) under weaker shape restrictions on λ^* that do not require previous knowledge of its exact functional form.

Using the terminology from Polonik (1995), we say that a set $D \subset \mathcal{U}$ is *k-constructible* from the class $\mathcal{C}$ of subsets of $\mathcal{U}$ if it can be constructed from at most k elements of $\mathcal{C}$ using the basic set operations $\cup, \cap$ and complement. In particular the Vapnik-Cervonenkis and Glivenko-Cantelli properties of the class $\mathcal{C}$ are inherited by the class of all set that are k-constructible from $\mathcal{C}$.

Condition 5.1. *(i) For any fixed value of $a \in \mathcal{A}$, the lower contour sets of the function $Z(u,a)$ with respect to u, $LC(\bar{z},a) := \{u \in \mathcal{U} : Z(u,a) \leq \bar{z}\}$ are k-constructible from a class of sets $\mathcal{C}$ for some integer $k < \infty$, where (ii) $\mathcal{C}$ is a class of closed subsets of $\mathcal{U}$ that is Glivenko-Cantelli for all measures that are absolutely continuous with respect to Lebesgue measure on $\mathcal{U}$.*

The restrictions on the shape of the clusters $LC(\bar{z},a)$ should in general also be exploited in the construction of the least favorable equilibrium selection rule, and we will discuss the case of convex lower contour sets in more detail below. We will now state lower level assumptions that imply Condition 5.1 and that hold for a broad range of practically relevant estimation problems. For the following recall that a function $f(x)$ is called *log-concave* if $\log(f(x))$ is concave - in particular any concave function is also log-concave, and any log-concave function is also quasiconcave. Similarly, we call $f(x)$ log-convex if $\log(f(x))$ is convex.

Assumption 5.2. *(i) The loss function $L(\theta,a)$ is either log-convex in θ for all $a \in \mathcal{A}$ or an indicator function for the complement of a convex set, and (ii) the conditional distribution of θ given u and x, $h(\theta|u,x) = \frac{g(u|\theta,x)\pi_0(\theta)}{\int_\Theta g(u|\theta,x)d\pi_0(\theta)}$, is log-concave in (u,θ) for all $x \in \mathcal{X}$. Finally, (iii) for all strategy profiles $s \in \mathcal{S}$, the regions $\{u \in \mathcal{U} : s \in S^*(u)\}$ are convex.*

The first part of this assumptions slightly strengthens the conditions on the loss function compared to Assumption 4.1, but is satisfied by all examples given in section 3. Part (ii) of Assumption 5.2 holds e.g. if the likelihood for u is Gaussian together with a normal prior on θ, as e.g. in the matching market problem in section 5. Part (iii) holds for any discrete action game since best responses are defined by linear inequality conditions on the payoff space. In particular, for any strategy profile $s_m \in \mathcal{S}_m$, the subset of $\mathcal{U}$ such that $u \in S_m^*(u)$ is an intersection of (linear) half spaces, and therefore convex.

Proposition 5.2. *Suppose Assumption 5.2 holds. Then condition 5.1 holds with $k = 1$, where $\mathcal{C}$ is the set of convex subsets of $\mathcal{U}$.*

5.1. **Algorithm.** Using an appropriate sampling procedure, in a first step we obtain a sample $(\theta_1,u_1),\dots,(\theta_B,u_B)$ of B draws from the augmented posterior with respect to the prior $\pi_I(\theta,\lambda;y) := \pi_0(\theta)\delta_{\lambda=\lambda^*(\theta)}$ - note that $\pi_I \notin \Gamma_\perp$. In contrast to the procedures in section 4, we now also keep the auxiliary draws of latent utilities, u_b.

In a second step we calculate

$$\breve{\varrho}(y,a,\Gamma_\perp) = \sup_{C\in\mathcal{C}} \frac{\sum_{b=1}^B \left[\frac{\min_{\lambda\in\Lambda}\prod_{m=1}^M \lambda_m(u_{m,b},s_m)}{\max_{\lambda\in\Lambda}\prod_{m=1}^M \lambda_m(u_{m,b},s_m)} \mathbb{1}\{u_b\in C\} + \mathbb{1}\{u_b\notin C\}\right] L(\theta_b,a)}{\sum_{b=1}^B \frac{\min_{\lambda\in\Lambda}\prod_{m=1}^M \lambda_m(u_{m,b},s_m)}{\max_{\lambda\in\Lambda}\prod_{m=1}^M \lambda_m(u_{m,b},s_m)} \mathbb{1}\{u_b\in C\} + \mathbb{1}\{u_b\notin C\}} \tag{5.2}$$

Maximization over $C \in \mathcal{C}$ is in general computationally demanding, and any efficient algorithm has to exploit the specific structure of the set $\mathcal{C}$. For the case in which $\mathcal{C}$ is a collection of convex subsets of $\mathcal{U}$ maximization in (5.2) can be solved as a linear programming problem in combination with a grid search over the unit interval. Also, since the dimension of $\mathcal{U}$ is typically much larger than that of Θ, there will in general be a curse of dimensionality in the minimization step in (5.2), however the size of the simulated sample (θ_b, u_b) will usually also be very large.

If there are no restrictions on equilibrium selection across the n observations, maximization and minimization of the product $\prod_{m=1}^{M} \lambda_m(u)$ requires in most cases only knowledge over whether s_m is an equilibrium given u_m, and whether it is unique, so that computational demands in this aspect are comparable to the problem for the unrestricted prior class in section 4.

Proposition 5.3. *Suppose Assumptions 4.1 and 4.3-4.5 and Condition 5.1 hold. Then for the empirical risk function $\breve{\varrho}(\cdot)$ defined in (5.2) we have $\breve{\varrho}(y, a, \Gamma_\perp) \xrightarrow{p} \varrho(y, a, \Gamma_\perp)$ uniformly in a.*

6. Empirical Application: Matching Markets with Non-Transferable Utility

We will reconsider the setting in Logan, Hoff, and Newton (2008) of a two-sided marriage market with non-transferable utility. There are n_w women and n_m men in the market, and the model allows only for marriages between women and men, where any individual can also choose to remain single. We assume that with probability one, every individual has strict preferences over all potential spouses on the opposite side of the market, including the option not to get married.

The most commonly applied solution concept for this problem is that of match stability.[5] Under our model assumptions, a stable matching is known to exist for all instances of the matching market, but in general the stable matching is not unique so that an econometric model for mate preferences alone does not define a reduced form for the observed market outcome. Furthermore, the number of stable matchings increases very fast in order in the size of the market, so that it will in general be computationally costly to determine the full set of stable matchings for a given set of preferences.[6] We will see that the approach proposed in this paper circumvents these potential computational difficulties without having to impose any restrictions on which of all possible stable matchings is selected in the observed market.

6.1. Preferences.

We represent woman i's preferences by utilities U_{ij}, where $j = 0, 1, \ldots, n_m$, where U_{i0} is i's utility from remaining single. Similarly, a man j's preferences are given by

[5] See Roth and Sotomayor (1990) for an overview of the theoretical literature on this subject

[6] Gusfield (1985) proposes an algorithm that finds all stable matchings in a market of n men and n women in time $O(n^2 + n|S|)$ where $|S|$ is the number of stable matchings, where for a market of this size there exists an instance of at least $|S| = O(2^{n-1})$ stable matchings (see Theorem 3.19 Roth and Sotomayor (1990)).

utilities V_{ji} for $i = 0, \dots, n_w$, where $i = 0$ stands for the outside option of not getting married. More specifically, we assume that agents have random utilities

$$\begin{array}{rcl} U_{ij} & = & x'_{ij}\beta_w + \varepsilon_{ij} \\ V_{ij} & = & z'_{ji}\beta_m + \zeta_{ji} \end{array}, \quad i = 1, \dots, n_w \text{ and } j = 1, \dots, n_m \tag{6.1}$$

where x_{ij} and z_{ji} are observable pair characteristics, and ε_{ij} and ζ_{ji} are pair-specific disturbances not known to the researcher that reflect unobserved pair-specific characteristics that affect i's preference for j relative to other potential spouses, and j's preference for i, respectively. For each woman $i = 1, \dots, n_w$ and man $j = 1, \dots, n_m$, the utility of remaining single is normalized to zero, i.e. $U_{i0} = V_{j0} = 0$.

In our application, we assume that ε_{ij} and ζ_{ji} are i.i.d. draws from a standard normal distribution and independent of observable characteristics x_{ij}, z_{ji}. This model for preferences is clearly very restrictive - for example, it might be desirable to allow for ε_{ij} to be correlated across $i = 1, \dots, n_w$ but uncorrelated across men in order to allow for unobserved heterogeneity in traits that are perceived as attractive or unattractive by all women, and vice versa. While it is clearly possible to incorporate parametric models for dependence into the model - e.g. by assuming a factor structure for the disturbance, or a random coefficient model for the regression coefficients - we will continue to work with i.i.d. pair-specific taste shocks for expositional purposes.

6.2. **Solution Concept.** We observe one realized matching $\hat{\mu}$ for the marriage market, that is the sample space $\mathcal{Y}$ consists of the joint support for observable individual and pair-level characteristics, and the possible matchings between the two sides. Our analysis is going to focus on match stability as a solution concept for predicting the market outcome.

A man m_i is *acceptable* a woman w_j if w_j would prefer to marry m_i over remaining single. A man m_i is said to *admire* a woman w_j at a matching μ if m_i is acceptable to w_j and prefers her to his mate $\mu(m_i)$ under μ. The definitions for the other side of the market are symmetric. The matching μ is called *stable* if there is no pairing of a man m_i and a woman w_j that admire each other, i.e. block the matching μ.

The matching μ_1 is M-preferred to μ_2, in symbols $\mu_1 >_M \mu_2$ if all men (weakly) prefer their partner under μ_1 to that under μ_2.

Constraint S. (Stable Matching) *Given the observed matching $\hat{\mu}$, random utilities U_{ij} and V_{ji} satisfy the following conditions: (i) if $U_{ij} > U_{i\hat{\mu}(i)}$, then $V_{j\hat{\mu}(j)} > V_{ji}$, and (ii) if $V_{ji} > V_{j\hat{\mu}(j)}$, then $U_{i\hat{\mu}(i)} > U_{ij}$.*

With strict preferences, according to Theorem 2.16 in Roth and Sotomayor (1990) the stable matchings constitute a lattice. I.e. holding preferences fixed, for any two stable matchings μ_1 and μ_2 we can find a stable matching $\lambda := \mu_1 \vee_M \mu_2$ such that all men (weakly) prefer both μ_1 and μ_2 over λ, and there is another stable matching $\nu := \mu_1 \wedge_M \mu_2$ such that all men prefer ν over μ_1 and μ_2. In particular there exists a unique M-optimal stable matching μ_M such that all men prefer μ_M to any other stable match, and a W-optimal match μ_W such that $\mu_W >_W \mu$ for all stable matchings μ. Furthermore, the men and the women have opposite preferences over stable matchings, i.e. if $\mu_1 >_M \mu_2$, then by Theorem 2.13 in Roth and Sotomayor (1990) it follows that $\mu_2 >_W \mu_1$.

Given women's and men's preferences, a stable matching can be constructed using the *deferred acceptance* algorithm by Gale and Shapley (1962):

- In the first round each woman proposes to her most preferred acceptable man.
- In the kth round, each man keeps his most preferred mate among the acceptable women that proposed to him in the $k-1$st round engaged, and rejects all other proposals. Each rejected woman then proposes to their next highest choice.
- If in round K no proposal is rejected, the algorithm stops.

The matching resulting from this procedure is stable, furthermore it is the W-optimal stable matching - see Theorem 2.12 in Roth and Sotomayor (1990). By symmetry, we can also construct the M-optimal stable matching via the deferred acceptance algorithm in which the men propose.

The lower probability for the observed matching $\hat{\mu}$ corresponds to sets of preferences for which it is the unique stable matching:

Constraint U. (Unique Stable Matching) *Given the observed matching $\hat{\mu}$ we have that for any alternative matching $\mu' \neq \hat{\mu}$ there exist woman i and man $j \neq \mu'(i)$ such that either (i) $U_{ij} > U_{i\mu'(i)}$ and $V_{ji} > V_{j\mu'(j)}$ or (ii) $U_{i0} > U_{i\mu'(i)}$, or (iii) $V_{j0} > V_{j\mu'(j)}$.*

This set of constraints is difficult to impose directly when drawing from the joint distribution of $(U_{ij}, V_{ji})_{ij}$, but it turns out that it is much easier to verify ex post by exploiting the relationship between the lattice structure of the set of stable matchings and the deferred acceptance algorithm. I formalize this insight in the following lemma, which is a direct consequence of standard results from the theoretical literature on stable matchings:

Lemma 6.1. *Suppose preferences are strict. Then a matching $\hat{\mu}$ satisfies Constraint U given preferences $(U_{ij}, V_{ji})_{i,j}$ if and only if both the deferred acceptance algorithm with the women proposing, and the deferred acceptance algorithm with the men proposing yield the matching $\hat{\mu}$.*

Proof: Since preferences are strict, the main result of Gale and Shapley (Theorem 2.12 in Roth and Sotomayor (1990)) implies that if $\hat{\mu}$ is produced by the deferred acceptance algorithm with the male side proposing, it is stable and is weakly preferred by all men over any other stable matching. By symmetry, if $\hat{\mu}$ is produced by the deferred acceptance algorithm with the women proposing, it is stable and preferred over any other stable matching by all women.

Hence if the matching $\hat{\mu}$ is unique, it must be produced by any of the two algorithms. For the converse, suppose that there is an alternative stable matching $\mu' \neq \hat{\mu}$. Since preferences are strict, there must be at least one man or one woman strictly preferring his or her spouse under μ' to that under $\hat{\mu}$, contradicting that $\hat{\mu}$ is weakly preferred to any other stable matching by all men and all women at the same time $\square$

In order to see that Assumption 4.3 holds for this model, consider an instance of the market for which $U_{i\hat{\mu}(i)} > 0$ and $V_{j,\hat{\mu}(j)} > 0$, and $U_{ij} < 0$ for all other pairs (i,j) (for notational simplicity suppose that under $\hat{\mu}$ no individual remains single). Since in this case each individual's spouse under $\hat{mu}$ is his or her only acceptable partner and preferences are strict, $\hat{\mu}$ is also the unique stable matching supported by this realization of preferences. Clearly given our assumptions, the probability that the random preferences satisfy these conditions simultaneously is strictly greater than zero, and can in fact be bounded away from zero if the support of x_{ij}, z_{ji} and Θ are compact.

6.3. **Analysis under Exchangeability.** Especially for large matching markets, an analysis based on Conditions S and U will likely not be very informative, but we may want to add the assumption that women and men are exchangeable within their respective side of the market. More specifically, let Assumption 2.1 hold with $\Sigma^* := \Sigma_w \times \Sigma_m$, where Σ_g is the set of all permutations of indices $i = 1, \dots, n_g$ for $g = m, w$.

By Theorem 2.22 in Roth and Sotomayor (1990), for a given realization of (strict) preferences u, the set of agents that remain single is the same for any stable matching. In particular, for every pair $s, s' \in \mathcal{S}(u)$, there exists a permutation $\sigma' \in \Sigma^*$ such that $s' = \sigma'(s)$ which implies that for every $s \in \mathcal{S}(u)$, $\mathcal{S}(u) \subset \{\sigma(s) : \sigma \in \Sigma^*\}$. Hence it is possible to calculate the bounds on the invariant likelihood with respect to Σ^*, $f_{inv,*}(y|\theta), f^*_{inv}(y|\theta)$ by minimization or maximization, respectively, over the permutations of observed pairs. Finding the exact solution to these two combinatorial optimization problem is a computationally challenging task, but in a future version of this paper, we are going to discuss approximate solutions and conservative bounds that can be obtained at a reasonable computational cost.

6.4. **Empirical Analysis.** In order to implement the algorithm from section 3, we assume that the marginal prior over $\theta = (\theta'_w, \theta'_m)'$ is multivariate normal with a given mean $\theta_0 := (\theta'_{w0}, \theta'_{m0})'$ and block-diagonal variance matrix $\Sigma_0 := \text{diag}\,(\Sigma_{w0}, \Sigma_{m0})$,

$$\pi_0(\theta) = N(\theta_0, \Sigma_0) \tag{6.2}$$

We can now implement the first part of Algorithm 2 by iterating over the following steps:

(1) set $\lambda_{k+1} = \lambda^*$ and draw a new parameter vector for women's preferences $\theta_{w,k+1}|\theta_{m,k}, \mathbf{U}_k, \mathbf{V}_k$
(2) draw a new parameter vector for men's preferences, $\theta_{m,k+1}|\theta_{w,k+1}, \mathbf{U}_k, \mathbf{V}_k$
(3) draw $\mathbf{U}_{k+1}|\mathbf{V}_k, \theta_{k+1}, \lambda^*$, i.e. imposing Constraints S
(4) draw $\mathbf{V}_{k+1}|\mathbf{U}_{k+1}, \theta_{k+1}, \lambda^*$, i.e. imposing Constraints S

(5) add a draw $(\theta_{w,k+2}, \theta_{m,k+2}, \lambda_{k+2}) = (\theta_{w,k+1}, \theta_{m,k+1}, \lambda_*)$ if the draw of $\mathbf{V}_{k+1}|\mathbf{U}_{k+1}$ satisfies constraints U, otherwise continue directly with step 1.

Note that in the first step, it is not necessary to condition on λ - i.e. whether the matching is stable - explicitly because by construction of the algorithm, the latent utilities were drawn imposing the stability or uniqueness conditions in the preceding iteration. By the same argument, the conditional distribution of λ depends on the latent utilities but not the value of θ.

Following Logan, Hoff, and Newton (2008), given the normal prior for θ_w in (6.2) the conditional distribution of θ_{k+1} given $\mathbf{U}_k, \mathbf{V}_k$ is given by

$$\theta_{w,k+1}|\mathbf{U}_k, \mathbf{V}_k \sim N(\eta_{wk}, \Omega_{wk})$$

where

$$\Omega_{wk}^{-1} := \sum_{i=1}^{n_w} X_i X_i' + \Sigma_{w0}^{-1} \quad \text{and } \eta_{wk} := \Omega_{wk}\left[\sum_{i=1}^{n_w} X_i U_i + \Sigma_{w0}^{-1}\eta_{w0}\right]$$

where the matrix $X_i = [X_{i0}, X_{i1}, \dots, X_{in_m}]$ is stacking the match-specific observables for woman i, and the column vector $U_i = [U_{i1,k}, \dots, U_{in_m,k}]'$ contains the match-specific utilities from the previous iteration of the algorithm. It is therefore possible to draw $\theta_{w,k+1}$ directly from its conditional distribution. The other component of the parameter, $\theta_{m,k+1}$ can also be drawn from its (multivariate normal) conditional distribution whose first two moments can be obtained in a completely analogous manner.

In order to simulate from a multivariate normal distribution imposing the multiple inequality constraints from constraint sets S and U efficiently, for steps 3 and 4 we modify the procedure by Robert (1995) using an accept/reject algorithm with proposals from translations of mutually independent exponential distributions with suitably chosen hazard rates. Since the blocks in the last two steps are in general highly correlated, the performance of the algorithm will likely improve substantively if we iterate between these steps several times before continuing with a new draw of θ_{k+1}.

Step 5 can in principle be replaced by a proper Metropolis-Hastings step after repeating the Gibbs steps 3 and 4 for the missing data several times, and accept the draw $(\theta_{k+1}, \lambda_*)$ if the proportion of samples for $\mathbf{V}_{k+1}|\mathbf{U}_{k+1}$ from steps 3 and 4 satisfying Constraint U is larger than a draw from the uniform distribution on the unit interval. This procedure should be expected to decrease the variance in approximating the posterior distribution of θ conditional on $\lambda = \lambda_*$ and reduce serial dependence within the Markov chain.

Given the Markov chain $(\theta_1, \lambda_1), \dots, (\theta_B, \lambda_B)$ we can obtain optimal point decisions and credible sets by approximating the minmax risk integrals $r^*(y, a, \Gamma)$ for the corresponding loss functions. By Proposition 4.4, the statistical decision minimizing $r_B(y, a, \Gamma)$ over $a \in \mathcal{A}$ attains the minimum of $r^*(y, a, \Gamma)$ as $B \to \infty$.

6.5. **Data and Estimation Results.** [to be added]

7. Discussion

[to be added]

Appendix A. Proofs

A.1. **Proof of Proposition 4.1:** First, we are going to establish existence of a solution to the problem (3.1): we can rewrite the problem as

$$\varrho(y,a,\Gamma) = \sup_{\omega\in[\omega_*,\omega^*]}\sup_{\pi\in\Gamma}\frac{1}{\omega}\int L(a,\theta)f(y|\theta,\lambda)d\pi(\theta,\lambda) \qquad \text{s.t.} \quad \int f(y|\theta,\lambda)d\pi(\theta,\lambda)=\omega$$

where $\omega_* := \int f(y|\theta,\lambda_*(\theta))\pi_0(\theta)d\theta$ and $\omega^* := \int f(y|\theta,\lambda^*(\theta))\pi_0(\theta)d\theta$.

The set Γ is convex by Assumption 4.2, so that for every fixed value of $\omega\in\mathcal{R}$ a solution to the constrained maximization problem exists by the Hahn-Banach theorem. Since the loss function $L(a,\theta)$ is bounded by Assumption 4.1, the function

$$H^*(\omega) := \sup_{\pi\in\Gamma}\frac{1}{\omega}\int L(a,\theta)f(y|\theta,\lambda)d\pi(\theta,\lambda) \qquad \text{s.t.} \quad \int f(y|\theta,\lambda)d\pi(\theta,\lambda)=\omega$$

is continuous in ω.

By assumption $y\in\mathcal{Y}$ was such that $\int f(y|\theta,\lambda)\pi_0(\theta)d\theta>0$ for some λ, so that by Assumption 4.3 $f_*>0$. Furthermore, since $f(y|\theta,\lambda)$ is uniformly bounded, we also have $f^*<\infty$, so that the interval for ω is compact and bounded away from zero. Hence the function $\frac{1}{\omega}H^*(\omega)$ is also continuous on the interval $[f_*,f^*]$, so that the solution to the problem (3.1) exists by Weierstrass' Theorem.

Next we are going to establish that there is a solution to the problem that meets the requirements in (i) and (ii) in the statement of the Proposition. In the following, denote $P_\theta(B)=\int_B\pi_0(\theta)d\theta$ for any set $B\subset\Theta$. Suppose the prior $\tilde{\pi}(\theta,\lambda)\in\Gamma$ solves the problem (3.1). Now we have to distinguish two cases:

If the set $B\subset\Theta$ on which $\tilde{\pi}(\theta,\lambda)$ does not satisfy properties (i) and (ii) has probability zero under the prior, i.e. $P_\theta(B)=0$, then we can construct a prior $\tilde{\pi}^*(\theta,\lambda)\in\Gamma$ that differs from $\tilde{\pi}(\theta,\lambda)$ only on B and meets the requirements in (i) and (ii). Assumptions 4.1 and 4.3 imply that

$$\frac{\int_{\Theta\times\Lambda}L(\theta,a)f(y|\theta,\lambda)d\tilde{\pi}^*(\theta,\lambda)}{\int_{\Theta\times\Lambda}f(y|\theta,\lambda)d\tilde{\pi}^*(\theta,\lambda)} = \frac{\int_{\Theta\times\Lambda}L(\theta,a)f(y|\theta,\lambda)d\tilde{\pi}(\theta,\lambda)}{\int_{\Theta\times\Lambda}f(y|\theta,\lambda)d\tilde{\pi}(\theta,\lambda)} = \varrho(y,a,\Gamma)$$

so that the prior $\tilde{\pi}^*(\theta,\lambda)$ is also a solution to the problem (3.1).

Now suppose instead that there exists a subset $B'\subset\Theta$ such that $P_\theta(B')>0$ such that for all $\theta\in B$, support$\tilde{\pi}(\lambda|\theta)\not\subseteq\{\lambda_*(\theta),\lambda^*(\theta)\}$. Set $L^*=L^*(a,y):=\varrho(y,a,\Gamma)$, and denote $A:=\{\theta\in\Theta: L(a,\theta)=L^*\}$. If $P_\theta(A\cap B)=0$, then we can modify $\tilde{\pi}$ as in the previous step without changing the value of the objective.

Now consider the case $P_\theta(A\cap B)\neq 0$, where we assume without loss of generality that $L(a,\theta)>L^*$ for all $\theta\in B/A$. Define the prior

$$\breve{\pi}(\theta,\lambda) := \begin{cases} \pi_0(\theta)\delta_{\lambda^*(\theta)} & \text{if } \theta\in B \\ \tilde{\pi}(\theta,\lambda) & \text{otherwise}\end{cases}$$

where $\delta_{\lambda'}$ denotes the (Dirac) delta-function for $\lambda=\lambda'$. Clearly, $\breve{\pi}(\theta,\lambda)\in\Gamma$.

Define $\breve{\omega} := \int_B [f^*(y|\theta) - f(y|\theta,\lambda)]\tilde{\pi}(d\theta, d\lambda)$, and note that by definition of $f^*(y|\theta)$ and $\tilde{\pi} \in \Gamma$, $\breve{\omega} \geq 0$. Now we can bound

$$\begin{aligned}
\breve{\varrho}(y,a) &:= \frac{\int_B L(\theta,a)f^*(y|\theta)\pi_0(d\theta) + \int_{\Theta/B} L(\theta,a)f(y|\theta,\lambda)\tilde{\pi}(d\theta,d\lambda)}{\int_B f^*(y|\theta)\pi_0(d\theta) + \int_{\Theta/B} f(y|\theta,\lambda)\tilde{\pi}(d\theta,d\lambda)} \\
&= \frac{\int_B L(\theta,a)[f^*(y|\theta) - f(y|\theta,\lambda)]\tilde{\pi}(d\theta,d\lambda) + \int_\Theta L(\theta,a)f(y|\theta,\lambda)\tilde{\pi}(d\theta,d\lambda)}{\int_B [f^*(y|\theta) - f(y|\theta,\lambda)]\tilde{\pi}(d\theta,d\lambda) + \int_\Theta f(y|\theta,\lambda)\tilde{\pi}(d\theta,d\lambda)} \\
&> \frac{\left[\breve{\omega} + \int_\Theta f(y|\theta,\lambda)\tilde{\pi}(d\theta,d\lambda)\right] r^*(y,a,\Gamma)}{\breve{\omega} + \int_\Theta f(y|\theta,\lambda)\tilde{\pi}(d\theta,d\lambda)} \\
&= \varrho(y,a,\Gamma)
\end{aligned}$$

contradicting that $\tilde{\pi}(\theta,\lambda)$ attains the supremum in (3.1). Hence we can rule out the case $P_\theta(A \cap B)$, which completes the proof □

A.2. **Proof of Proposition 4.2:** Note that we can write

$$\sup_{\psi\in\Psi} \frac{\int L(\theta;a)\psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)}{\int \psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)} = \sup_{\psi\in\Psi} \frac{\int L(\theta;a)f(y|\theta,\lambda)\psi(\theta,\lambda)\pi_I(d\theta,d\lambda;y)}{\int f(y|\theta,\lambda)\psi(\theta,\lambda)\pi_I(d\theta,d\lambda;y)}$$

Since $\pi^*(\theta,\lambda;a,y)$ is absolutely continuous with respect to $\pi_I(\theta,\lambda;y)$ for (a,y) by Assumption 4.4, it follows from the Radon-Nikodym theorem that there exists a nonnegative measurable function $\psi(\lambda,\theta)$ such that for any function $h(\theta,\lambda)$ that is integrable with respect to π_I, we have $\int h d\pi^* = \int h\psi d\pi_I$. Since $\pi^* \in \Gamma$, it is a probability measure so that $\int \psi d\pi_0 = \int d\pi^* = 1$, so that $\psi \in \Psi$. Hence

$$\sup_{\psi\in\Psi} \frac{\int L(\theta;a)\psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)}{\int \psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)} \geq \sup_{\pi(\theta,\lambda)\in\Gamma} \frac{\int_{\Theta\times\Lambda} L(\theta,a)f(y|\theta;\lambda)\pi(d\lambda,d\theta)}{\int_{\Theta\times\Lambda} f(y|\theta;\lambda)\pi(d\lambda,d\theta)} \tag{A.1}$$

Conversely, since for any function $\psi \in \Psi$, $\int \psi(\theta,\lambda)\pi_I(\theta,d\lambda) = \pi_0(\theta)$ for all $\theta \in \Theta$, there is an element $\tilde{\pi} \in \Gamma$ such that for any measurable function h on $\Theta \times \Lambda$, $\int h d\tilde{\pi} = \int h\psi d\pi_I$. Hence we also have

$$\sup_{\pi(\theta,\lambda)\in\Gamma} \frac{\int_{\Theta\times\Lambda} L(\theta,a)f(y|\theta;\lambda)\pi(d\lambda,d\theta)}{\int_{\Theta\times\Lambda} f(y|\theta;\lambda)\pi(d\lambda,d\theta)} \geq \sup_{\psi\in\Psi} \frac{\int L(\theta;a)\psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)}{\int \psi(\theta,\lambda)p(d\theta,d\lambda;y,\pi_I)}$$

which together with the inequality (A.1) establishes the claim □

A.3. **Proof of Proposition 4.3:** The proof of consistency will proceed in two steps: we will first show pointwise convergence of $\hat{J}_B(\tilde{L},a)$ to its limit, and then strengthen this result to uniform convergence, and finally use Propositions 4.1 and 4.2 show that the limit equals $r(y,a,\Gamma)$.

Fix $a \in \mathcal{A}$, let $C(\tilde{L}) := \left\{\theta \in \Theta : L(\theta,a) < \tilde{L}\right\}$, and define

$$J_0(\tilde{L},a) := \frac{\int_{C(\tilde{L})} L(\theta,a)f_*(y|\theta)\pi(d\theta) + \int_{\Theta/C(\tilde{L})} L(\theta,a)f^*(y|\theta)\pi(d\theta)}{\int_{C(\tilde{L})} f_*(y|\theta)\pi(d\theta) + \int_{\Theta/C(\tilde{L})} f^*(y|\theta)\pi(d\theta)}$$

We will now show that $\hat{J}_B(\tilde{L},a) \xrightarrow{p} J_0(\tilde{L},a)$ for all $\tilde{L} \in \mathbb{R}$.

First note that by Assumption 4.1, $L(\theta,a)$ is bounded, so that both $\hat{J}_B(\tilde{L},a)$ and $J_0(\tilde{L},a)$ are constant for $\tilde{L} < \inf_{a,\theta} L(\theta,a)$ and $\tilde{L} > \sup_{a,\theta} L(\theta,a)$, respectively. Now consider the numerator and denominator of the expression for $\hat{J}_B(\tilde{L},a)$ separately, i.e. let

$$\hat{Q}_B(\tilde{L},a) := \sum_{b=1}^{B} L(\theta_b,a)\left[\mathbb{1}\{L(\theta_b,a) \leq \tilde{L}, \lambda_b = \lambda_*(\theta_b)\} + \frac{\pi_I(\lambda_*(\theta_b)|\theta_b)}{\pi_I(\lambda^*(\theta_b)|\theta_b)}\mathbb{1}\{L(\theta_b,a) > \tilde{L}, \lambda_b = \lambda^*(\theta_b)\}\right] \frac{\pi_0(\theta_b)}{\pi_I(\theta_b,\lambda_*(\theta_b);y)}$$

and

$$\hat{R}_B(\tilde{L},a) := \sum_{b=1}^{B}\left[\mathbb{1}\{L(\theta_b,a)\leq \tilde{L}, \lambda_b=\lambda_*(\theta_b)\} + \frac{\pi_I(\lambda_*(\theta_b)|\theta_b)}{\pi_I(\lambda^*(\theta_b)|\theta_b)}\mathbb{1}\{L(\theta_b,a) > \tilde{L}, \lambda_b=\lambda^*(\theta_b)\}\right]\frac{\pi_0(\theta_b)}{\pi_I(\theta_b,\lambda_*(\theta_b);y)}$$

By Assumption 4.5 (θ_b,λ_b), $b=1,\dots,B$ are (not necessarily independent) draws from the posterior $p(\theta,\lambda;y,\pi_I)$, so that the expectation of $\hat{Q}_B(\tilde{L},a)$ with respect to (θ,λ) can be rewritten as

$$\begin{aligned}\mathbb{E}\left[\hat{Q}_B(\tilde{L},a)\right] &= \int_{C(\tilde{L})} L(\theta,a)p(\theta,\lambda^*(\theta);y,\pi_I)\frac{\pi_0(\theta)}{\pi_I(\theta,\lambda^*(\theta);y)}d\theta \\ &\quad + \int_{\Theta/C(\tilde{L})} L(\theta,a)p(\theta,\lambda_*(\theta);y,\pi_I)\frac{\pi_0(\theta)}{\pi_I(\theta,\lambda_*(\theta_b);y)}d\theta \\ &= \frac{\int_{C(\tilde{L})} L(\theta,a)f_*(y|\theta)\pi_0(\theta)d\theta + \int_{\Theta/C(\tilde{L})} L(\theta,a)f^*(y|\theta)\pi_0(\theta)d\theta}{\int f(y|\theta,\lambda)\pi_I(d\theta,d\lambda;y)} \qquad \text{(A.2)}\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\hat{R}_B(\tilde{L},a)\right] = \frac{\int_{C(\tilde{L})} f_*(y|\theta)\pi_0(\theta)d\theta + \int_{\Theta/C(\tilde{L})} f^*(y|\theta)\pi_0(\theta)d\theta}{\int f(y|\theta,\lambda)\pi_I(d\theta,d\lambda;y)} \qquad \text{(A.3)}$$

Using Assumptions 4.3 and 4.4, we can bound this expectation by

$$\frac{\int_{C(\tilde{L})} f_*(y|\theta)\pi_0(\theta)d\theta + \int_{\Theta/C(\tilde{L})} f^*(y|\theta)\pi_0(\theta)d\theta}{\int f(y|\theta,\lambda)\pi_I(d\theta,d\lambda;y)} \geq \frac{\int f_*(y|\theta)\pi_0(d\theta)}{\int f^*(y|\theta)\pi_I(d\theta)} > \frac{\int f^*(y|\theta)\pi_0(d\theta)}{\int f^*(y|\theta)\pi_I(d\theta)} > 0$$

where $\pi_I(\theta;y) := \int \pi_I(\theta,d\lambda;y)$. Furthermore, by Assumptions 4.1 and 4.3, $L(\theta,a)$ and $f^*(y|\theta)$ are uniformly bounded and therefore also bounded under the $L_1(\pi_0)$ norm.

Hence, in the first case of Assumption 4.5, a law of large numbers for $\hat{Q}_B(\tilde{L},a)$ and $\hat{R}_B(\tilde{L},a)$ together with the continuous mapping theorem implies that

$$\hat{J}_B(\tilde{L},a) = \frac{\hat{Q}_B(\tilde{L},a)}{\hat{R}_B(\tilde{L},a)} \xrightarrow{p} \frac{\mathbb{E}\left[Q_B(\tilde{L},a)\right]}{\mathbb{E}\left[R_B(\tilde{L},a)\right]} = J_0(\tilde{L},a)$$

On the other hand, if the sample (θ_b,λ_b) is from a Harris recurrent Markov chain, the conclusion follows from the Ergodic Theorem (e.g. Theorem 6.63 in Robert and Casella (2004)) using the same line of reasoning as in the independent case.

Next, we establish that convergence is uniform in $\tilde{L}\in\mathbb{R}$ and $a\in\mathcal{A}$: Consider the class of indicator functions for the lower contour sets of $L(\theta,a)$ in Θ,

$$\mathcal{G} := \left\{\theta \mapsto \mathbb{1}\{L(\theta,a) < \tilde{L}\} : a\in\mathcal{A}, \tilde{L}\in\mathbb{R}\right\}$$

. Since by Assumption 4.1, $L(\theta,a)$ the contour sets of $L(\theta,a)$ are convex, and furthermore, $\pi_0(\theta)$ is absolutely continuous with respect to Lebesgue measure on Θ by Assumption 4.2. Hence it follows from a result by Eddy and Hartigan (1977) that $\mathcal{G}$ is Glivenko-Cantelli if Θ is a subset of a Euclidean space.

Since the class $\mathcal{L}$ was also assumed to be Glivenko-Cantelli, we can apply Theorem 3 in van der Vaart and Wellner (2000) on the permanence of the Glivenko-Cantelli property under continuous transformations to the class $\phi(\mathcal{L},\mathcal{G}) := \left\{L(a,\theta)\mathbb{1}\{L(a,\theta) < \tilde{L}\} : a\in\mathcal{A}, \tilde{L}\in\mathbb{R}\right\}$. Since the two summands in $\hat{Q}_B(\tilde{L},a)$ can be represented as averages of functions in $\mathcal{L}$ and $\phi(\mathcal{L},\mathcal{G})$, we get uniform convergence of the numerator. Similarly uniformity of convergence of the denominator $\hat{R}_B(\tilde{L},a)$ follows directly from the Glivenko-Cantelli property of $\mathcal{G}$.

By Assumption 4.3, $\hat{R}_B(\tilde{L},a)$ is bounded away from zero with probability approaching 1, so that $\hat{J}_B(\lambda,a) = \frac{\hat{Q}_B(\tilde{L},a)}{\hat{R}_B(\tilde{L},a)}$ is Lipschitz, and therefore uniformly continuous in $\hat{Q}_B(\tilde{L},a)$ and $\hat{R}_B(\tilde{L},a)$ so that by the continuous mapping theorem convergence of $\hat{J}_B(\tilde{L},a)$ is also uniform in $\tilde{L} \in \mathbb{R}$ and $a \in \mathcal{A}$.

Now, by uniform convergence of $\hat{J}_B(\tilde{L},a)$ in $(\tilde{L},a)$, we have

$$\sup_{a\in\mathcal{A}} \left| \sup_{\tilde{L}\in\mathbb{R}} \hat{J}_B(\tilde{L},a) - \sup_{\tilde{L}\in\mathbb{R}} J_0(\tilde{L},a) \right| \overset{p}{\to} 0$$

Since by Proposition 4.1, $\varrho(y,a,\Gamma) = \sup_{\tilde{L}\in\mathbb{R}} J_0(\tilde{L},a)$, this completes the proof □

A.4. **Proof of Proposition 4.4:** The proof of validity for the second algorithm follows the same steps as the argument in Proposition 4.3. We will therefore only need to show pointwise convergence for $\tilde{L} \in \mathbb{R}$ and $a \in \mathcal{A}$, and the remainder of the proof is completely analogous to that of the previous result. Again, consider numerator and denominator of $\tilde{J}_B(\tilde{L},a)$ separately: for

$$\tilde{Q}_B(\tilde{L},a) := \sum_{b=1}^{B} L(\theta_b,a)\left[\mathbb{1}\{L(\theta_b,a) \le \tilde{L}\}v_{*b} + \mathbb{1}\{L(\theta_b,a) > \tilde{L}\}v_b^*\right]$$

we have

$$\begin{aligned}
\mathbb{E}[\tilde{Q}_B(\tilde{L},a)] &= \int_{C(\tilde{L})} L(\theta,a)\int_{\mathcal{U}} v_*(\theta,u;y)\lambda_0(u;s)g(u|x,\theta)du\pi_0(d\theta) \\
&\quad + \int_{\Theta/C(\tilde{L})} L(\theta,a)\int_{\mathcal{U}} v^*(\theta,u;y)\lambda_0(u;s)g(u|x,\theta)du\pi_0(d\theta) \\
&= \int_{C(\tilde{L})} L(\theta,a)\int_{\mathcal{U}} \lambda_*(u;s)g(u|x,\theta)du\pi_0(d\theta) + \int_{\Theta/C(\tilde{L})} L(\theta,a)\int_{\mathcal{U}} \lambda^*(u;s)g(u|x,\theta)du\pi_0(d\theta) \\
&= \int_{C(\tilde{L})} L(\theta,a)f_*(y|\theta)\pi_0(d\theta) + \int_{\Theta/C(\tilde{L})} L(\theta,a)f^*(y|\theta)\pi_0(d\theta) \\
&= \mathbb{E}[\hat{Q}_B(\tilde{L},a)]
\end{aligned}$$

Similarly, for the denominator

$$\tilde{R}_B(\tilde{L},a) := \sum_{b=1}^{B} \left[\mathbb{1}\{L(\theta_b,a) \le \tilde{L}\}v_{*b} + \mathbb{1}\{L(\theta_b,a) > \tilde{L}\}v_b^*\right]$$

we have $\mathbb{E}[\tilde{R}_B(\tilde{L},a)] = \mathbb{E}[\hat{R}_B(\tilde{L},a)]$, so that the argument can be completed using the same steps as in the proof of Proposition 4.3 □

A.5. **Proof of Proposition 5.1:** Existence of a solution of the problem (3.1) can be shown using the same arguments as in the proof of Proposition 4.1.

For any choice of $\nu(\lambda)$, we can rewrite

$$J(\nu) = \frac{\int_\Lambda \int_\Theta L(a,\theta)f(y|\theta,\lambda)d\pi_0(\theta)d\nu(\lambda)}{\int_\Lambda \int_\Theta f(y|\theta,\lambda)d\pi_0(\theta)d\nu(\lambda)} =: \frac{Q(\nu)}{R(\nu)}$$

By Fubini's theorem,

$$\begin{aligned} Q(\nu) &:= \int_\Lambda \int_\Theta L(a,\theta) f(y|\theta,\lambda) d\pi_0(\theta) d\nu(\lambda) \\ &= \int_\Lambda \int_\Theta L(a,\theta) \int_{\mathcal{U}} \prod_{m=1}^M \lambda_m(u_m,s_m) g_m(u_m|x_m,\theta) du_M \dots du_1 d\pi_0(\theta) d\nu(\lambda) \\ &= \int_\Lambda \int_{\mathcal{U}} \left[\prod_{m=1}^M \lambda_m(u_m,s_m)\right] \left[\int_\Theta L(a,\theta) \prod_{m=1}^M g_m(u_m|x_m,\theta) d\pi_0(\theta)\right] du_M \dots du_1 d\nu(\lambda) \\ &=: \int_\Lambda \int_{\mathcal{U}} \left[\prod_{m=1}^M \lambda_m(u_m,s_m)\right] H(u,a) du d\nu(\lambda) \end{aligned}$$

for any $m = 1, \dots, M$, where we define

$$H(u,a) := \int_\Theta L(a,\theta) \prod_{m=1}^M g_m(u_m|x_m,\theta) d\pi_0(\theta)$$

Similarly, defining

$$h(u) := \int_\Theta \prod_{m=1}^M g_m(u_m|x_m,\theta) d\pi_0(\theta)$$

we can write

$$\begin{aligned} R(\nu) &:= \int_\Lambda \int_\Theta f(y|\theta,\lambda) d\pi_0(\theta) d\nu(\lambda) \\ &= \int_\Lambda \int_{\mathcal{U}} \prod_{m=1}^M \lambda_m(u_m,s_m) h(u) du d\nu(\lambda) \end{aligned}$$

By the same arguments as in the proof of Proposition 4.1, the function $J(\nu)$ is maximized at a prior $\nu^*(\lambda)$ which puts unit mass on $\lambda^* = (\lambda_1^*, \dots, \lambda_M^*)$ with

$$\lambda_m^*(u_m) = \max_{\lambda\in\Lambda} \prod_{m=1}^M \lambda_m(u_m) \mathbb{1}\{Z(u,a) \geq Z^*\} + \min_{\lambda\in\Lambda} \prod_{m=1}^M \lambda_m(u_m) \mathbb{1}\{Z(u,a) < Z^*\}$$

where $Z(u,a) = \frac{H(u,a)}{h(u)}$ is as defined in (5.1), and $Z^* := \varrho(y,a,\Gamma_\perp)$ □

A.6. **Proof of Proposition 5.2:** Consider the ratio

$$Z(u,a) := \frac{\int_\Theta L(\theta,a) g(u|x,\theta) d\pi_0(\theta)}{\int_\Theta g(u|x,\theta) d\pi_0(\theta)}$$

By log-concavity of $h(\theta|u,x)$ and convexity of C, it follows from known properties of log-concave functions (see e.g. Theorem 6 in Prékopa (1973)) that $Z(u,a)$ is log-concave - and therefore in particular quasi-concave - in u for all $a \in \mathcal{A}$. Hence the lower contour set $LC(\bar{z},a) = \{u \in \mathcal{U} : Z(u,a) \leq \bar{z}\}$ is convex for any values of a and $\bar{z} \in \mathbb{R}$. Finally, the set of convex subsets of $\mathcal{U}$ is Glivenko-Cantelli for all measures that are absolutely continuous with respect to Lebesgue measure on $\mathcal{U}$, e.g. by the Theorem of Eddy and Hartigan (1977) so that Condition 5.1 holds with $k = 1$ □

A.7. **Proof of Proposition 5.3:** Consistency of the simulation algorithm can be established using similar arguments as in the proofs for Proposition 4.3.

For any $C \in \mathcal{C}$ denote

$$\begin{aligned}\hat{J}_b(C,a) \quad &:= \quad \frac{\sum_{b=1}^{B}\left[\frac{\min_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_{m,b},s_m)}{\max_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_{m,b},s_m)}\mathbb{1}\left\{u_b \in C\right\} + \mathbb{1}\left\{u_b \notin C\right\}\right]L(\theta_b,a)}{\sum_{b=1}^{B}\frac{\min_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_{m,b},s_m)}{\max_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_{m,b},s_m)}\mathbb{1}\left\{u_b \in C\right\} + \mathbb{1}\left\{u_b \notin C\right\}} \\ &=: \quad \frac{\hat{Q}_b(C,a)}{\hat{R}_b(C,a)}\end{aligned}$$

For a given selection rule $\lambda(u)$, the conditional likelihood ratio between $f(y|u,\theta,\lambda)$ and the distribution corresponding to the most favorable selection rule $f^*(y|u,\theta) := \max_{\lambda\in\Lambda} f(y|u,\theta,\lambda)$ given $u = (u_1',\dots,u_M')'$ is of the form

$$w(u,\lambda) := \frac{f(y|u,\theta,\lambda)}{f^*(y|u,\theta)} = \frac{\prod_{m=1}^{M}\lambda_m(u_m,s_m)}{\max_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_{m,b},s_m)}$$

and let $\lambda^*(u,y) := \max_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_m,s_m)$ and $\lambda_*(u,y) := \min_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_m,s_m)$, respectively.

By Assumption 4.5, the sample $(\theta_b',u_b')'$ is drawn from a distribution with marginal density

$$h(\theta,u|y) := \frac{\mathbb{1}\{s_m \in S_m^*(u), m=1,\dots,M\}g(u|x,\theta)\pi_0(\theta)}{\int_{\mathcal{U}}\int_{\Theta}\mathbb{1}\{s_m \in S_m^*(u), m=1,\dots,M\}g(u|x,\theta)d\pi_0(\theta)du} = \frac{\mathbb{1}\{s_m \in S_m^*(u), m=1,\dots,M\}g(u|x,\theta)\pi_0(\theta)}{\int_{\Theta} f^*(y|\theta)d\theta}$$

Taking expectations,

$$\begin{aligned}\mathbb{E}\left[\hat{Q}_b(C,a)\right] \quad &= \quad \int\Theta\int_C \frac{\lambda_*(u,y)}{\lambda^*(u,y)}L(\theta,a)dh(\theta,u|y) + \int_{\mathcal{U}/C} L(\theta,a)\eth(\theta,u|y) \\ &= \quad \frac{\int_{\Theta} L(\theta,a)\left(\int_C \frac{\lambda_*(u,y)}{\lambda^*(u,y)}\mathbb{1}\{s_m \in S_m^*(u),\forall m\}g(u|x,\theta)du + \int_{\mathcal{U}/C}\mathbb{1}\{s_m \in S_m^*(u),\forall m\}g(u|x,\theta)du\right)d\pi_0(\theta)}{\int_{\Theta}\left(\int_C \frac{\lambda_*(u,y)}{\lambda^*(u,y)}\mathbb{1}\{s_m \in S_m^*(u),\forall m\}g(u|x,\theta)du + \int_{\mathcal{U}/C\mathbf{1}\{s_m\in S_m^*(u),\forall m\}g(u|x,\theta)du} f^*(y|\theta)\right)d\pi_0(\theta)} \\ &= \quad \frac{\int_{\Theta} L(\theta,a)\left(\int_C f(y|u,\theta,\lambda_*(u,y))d\pi_0(\theta)du + \int_{\mathcal{U}/C} f(y|u,\theta,\lambda^*(u,y))du\right)d\pi_0(\theta)}{\int_{\Theta} f^*(y|\theta)d\pi_0(\theta)}\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\hat{R}_b(C,a)\right] \quad = \quad \frac{\int_{\Theta}\left(\int_C f(y|u,\theta,\lambda_*(u,y))du + \int_{\mathcal{U}/C} f(y|u,\theta,\lambda^*(u,y))du\right)d\pi_0(\theta)}{\int_{\Theta} f^*(y|\theta)d\pi_0(\theta)} \tag{A.4}$$

By the same arguments as in the proof of Proposition 4.3,

$$\begin{aligned}\hat{J}_b(C,a) \quad &\overset{p}{\to} \quad \frac{\mathbb{E}\left[\hat{Q}_b(C,a)\right]}{\mathbb{E}\left[\hat{R}_b(C,a)\right]} =: J(C,a) \\ &= \quad \frac{\int_{\Theta} L(\theta,a)\left(\int_C f(y|u,\theta,\lambda_*(u,y))d\pi_0(\theta)du + \int_{\mathcal{U}/C} f(y|u,\theta,\lambda^*(u,y))du\right)d\pi_0(\theta)}{\int_{\Theta}\left(\int_C f(y|u,\theta,\lambda_*(u,y))du + \int_{\mathcal{U}/C} f(y|u,\theta,\lambda^*(u,y))du\right)d\pi_0(\theta)}\end{aligned}$$

for any set $C \in \mathcal{C}$. Since the assumptions of Proposition 5.1 are subsumed under those for this proposition, there is a set $C \subset \mathcal{U}$ such that $J(C,a) = \varrho(y,a,\Gamma_\perp)$. Since by Condition 5.1 $C \in \mathcal{C}$, we have $\varrho(y,a,\Gamma_\perp) \le \sup_{C\in\mathcal{C}} J(C,a)$. Furthermore Assumption 5.1 implies that $\nu_C(\lambda)\pi_0(\theta) \in \Gamma_\perp$ for any set $C \in \mathcal{C}$, where $\nu_C(\lambda)$ is a distribution on Λ which puts unit mass on $\tilde{\lambda}(C)$ with $\prod_{m=1}^{M}\tilde{\lambda}_m(u_m;C) := \mathbb{1}\{u \in C\}\min_{\lambda\in\Lambda}\prod_{m=1} M\lambda_m(u_m,s_m) + \mathbb{1}\{u \notin C\}\max_{\lambda\in\Lambda}\prod_{m=1}^{M}\lambda_m(u_m,s_m)$. Hence we also have $\varrho(y,a,\Gamma_\perp) \ge \sup_{C\in\mathcal{C}} J(C,a)$, so that indeed $\varrho(y,a,\Gamma_\perp) = \sup_{C\in\mathcal{C}} J(C,a)$.

Since by assumption, Condition 5.1 holds, so that in particular the family of sets $\mathcal{C}$ is a Glivenko-Cantelli class for measures that are absolutely continuous with respect to Lebesgue measure. Hence by Assumption

4.1, convergence is also uniform in $C \in \mathcal{C}$ and $a \in \mathcal{A}$, so that $\sup_{C\in\mathcal{C}} \hat{J}_b(C,a) \xrightarrow{p} \sup_{C\in\mathcal{C}} J(C,a) = \varrho(y,a,\Gamma_\perp)$ uniformly in $a \in \mathcal{A}$, and the conclusion of Proposition 5.3 follows □

References

Baccara, M., A. Imrohoroglu, A. Wilson, and L. Yariv (2012): "A Field Study on Matching with Network Externalities," *American Economic Review*, 102(5).

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2006): "Estimating Static Models of Strategic Interaction," NBER Working Paper 12013.

Bajari, P., H. Hong, and S. Ryan (2010): "Identification and Estimation of a Discrete Game of Complete Information," *Econometrica*, 78(5), 1529–1568.

Beresteanu, A., I. Molchanov, and F. Molinari (2009): "Sharp Identification Regions in Models with Convex Predictions: Games, Individual Choice, and Incomplete Data," cemmap working paper CWP27/09.

Berger, J. (1984): *The Robust Bayesian Viewpoint*Robustness of Bayesian Analysis. North Holland.

——— (1985): *Statistical Decision Theory and Bayesian Analysis*. Springer.

Berger, J., D. R. Insua, and F. Ruggeri (2000): *Bayesian Robustness*Robust Bayesian Analysis. Springer.

Bresnahan, T., and P. Reiss (1990): "Entry in Monopoly Markets," *Review of Economic Studies*, 57(4).

——— (1991a): "Empirical Models of Discrete Games," *Journal of Econometrics*, 48, 57–81.

——— (1991b): "Entry and Competition in Concentrated Markets," *Journal of Political Economy*, 99(5), 977–1009.

Brock, W., and S. Durlauf (2001): "Discrete Choice with Social Interactions," *Review of Economic Studies*, 68.

Chen, X., E. Tamer, and A. Torgovitsky (2011): "Sensitivity Analysis in a Semiparametric Likelihood Model: A Partial Identification Approach," working paper, Yale and Northwestern.

Christakis, N., J. Fowler, G. Imbens, and K. Kalyanaraman (2010): "An Empirical Model for Strategic Network Formation," working paper, Harvard University.

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791–1828.

Dudley, R. (1999): *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.

Echenique, F., S. Lee, and M. Shum (2010): "Aggregate Matchings," working paper, Caltech.

Eddy, W., and J. Hartigan (1977): "Uniform Convergence of the Empirical Distribution Function over Convex Sets," *Annals of Statistics*, 5(2), 370–374.

Ferguson, T. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Fox, J. (2010): "Identification in Matching Games," *Quantitative Economics*, 1, 203–254.

Gale, D., and L. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69(1), 9–15.

Galichon, A., and M. Henry (2011): "Set Identification in Models with Multiple Equilibria," *Review of Economic Studies*, forthcoming.

Galichon, A., and B. Salanié (2010): "Matching with Trade-offs: Revealed Preferences over Competing Characteristics," working paper, Columbia and École Polytechnique.

Gilboa, I., and D. Schmeidler (1989): "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18, 141–153.

Gusfield, D. (1985): “Three Fast Algorithms for Four Problems in Stable Marriage,” unpublished manuscript, Yale University.

Hansen, L., and T. Sargent (2008): *Robustness*. Princeton University Press.

Heckman, J. (1978): “Dummy Endogenous Variables in a Simultaneous Equation System,” *Econometrica*, 46(6), 931–959.

Jovanovic, B. (1989): “Observable Implications of Models with Multiple Equilibria,” *Econometrica*, 57(6), 1431–1437.

Kitagawa, T. (2010): “Inference and Decision for Set Identified Parameters Using Posterior Lower and Upper Probabilities,” working paper, UCL.

Kudō, A. (1967): *On Partial Prior Information and the Property of Parametric Sufficiency*vol. 1 of *Proc. Fifth Berkeley Symp. Statist. Probab.* University of California Press.

Liao, Y., and W. Jiang (2010): “Bayesian Analysis in Moment Inequality Models,” *The Annals of Statistics*, 38(1), 275–316.

Loève, M. (1963): *Probability Theory*. van Nostrand, 3 edn.

Logan, J., P. Hoff, and M. Newton (2008): “Two-Sided Estimation of Mate Preferences for Similarities in Age Education, and Religion,” *Journal of the American Statistical Association*, 103(482), 559–569.

Manski, C. (2000): “Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice,” *Journal of Econometrics*, 95, 415–442.

McCulloch, R., and P. Rossi (1994): “An Exact Likelihood Analysis of the Multinomial Probit Model,” *Journal of Econometrics*, 64, 207–240.

Milgrom, P., and J. Roberts (1990): “Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities,” *Econometrica*, 58(6), 1255–1277.

Molchanov, I. (2005): *Theory of Random Sets*. Springer, London.

Moon, H., and F. Schorfheide (2010): “Bayesian and Frequentist Inference in Partially Identified Models,” working paper, USC and University of Pennsylvania.

Neyman, J., and E. Scott (1948): “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, 16(1), 1–32.

Norets, A. (2009): “Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables,” *Econometrica*, 77(5), 1665–1682.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006): “Moment Inequalities and their Application,” working paper, Harvard University.

Polonik, W. (1995): “Measuring Mass Concentrations and Estimating Density Contour Clusters - An Excess Mass Approach,” *Annals of Statistics*, 23(3), 855–881.

Prékopa, A. (1973): “On Logarithmic Concave Measures and Functions,” *Acta Scientiarum Mathematicarum*, 34, 335–343.

Robert, C. (1995): “Simulation of Truncated Normal Variables,” *Statistics and Computing*, 5, 121–125.

Robert, C., and G. Casella (2004): *Monte Carlo Statistical Methods*. Springer.

Roth, A., and M. Sotomayor (1990): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press.

Song, K. (2009): “Point Decisions for Interval-Identified Parameters,” working paper, University of Pennsylvania.

Stoye, J. (2009): “Minimax Regret Treatment Choice with Finite Samples,” *Journal of Econometrics*, 151, 70–81.

Strzalecky, T. (2011): “Axiomatic Foundations of Multiplier Preferences,” *Econometrica*, 79(1), 47–73.

TAMER, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70, 147–165.

TANNER, M., AND W. WONG (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82.

UETAKE, K., AND Y. WATANABE (2012): "Entry by Merger: Estimates from a Two-Sided Matching Model with Externality," working paper, Northwestern University.

VAN DER VAART, A., AND J. WELLNER (2000): "Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes," pp. 115–133, in E. Giné, D. Mason and J. Wellner(eds.): High Dimensional Probability II.