Counterfactual prediction in complete information games: point prediction under partial identification*

Sung Jae Jun† and Joris Pinkse‡

Center for the Study of Auctions, Procurements and Competition Policy
Department of Economics
The Pennsylvania State University

this version: February 2019

We study the problem of counterfactual prediction in discrete decision games with complete information, pure strategies, and Nash equilibria: the presence of multiple equilibria poses unique challenges. We introduce multiple types of counterfactuals to establish sharp identified bounds for their prediction probabilities. We propose and compare various point prediction methods, namely midpoint prediction, an approach using a Dirichlet–based prior, a maximum entropy method, and minmax with an entropy constraint. On balance, we conclude that the maximum–entropy approach is the least of several evils. Our results have implications for counterfactual prediction in other models with partial identification.

Key words: complete information games, counterfactual prediction, partial identification, maximum entropy, Dirichlet process, minmax decisions.

JEL Codes: C01; C10; C57.

*We thank Victor Aguirregabiria for a great discussion at the 2016 Canadian Econometrics Study Group meeting at London, Ontario. We thank Tim Armstrong, Paul Grieco, Ken Hendricks, Rohit Lamba, Andrés Aradillas–López, Richard Blundell, Andrew Chesher, Joachim Freyberger, Ron Gallant, Bruce Hansen, Marc Henry, Nail Kashaev, Michael Leung, Francesca Molinari, Roger Moon, Ulrich Müller, Hashem Pesaran, Jack Porter, Mar Reguant, Bruno Salcedo, Xiaoxia Shi, Elie Tamer, Guofu Tan, conference participants at the Penn State–Cornell econometrics conference at Ithaca, a conference at the Institute for Economics at Academia Sinica, and seminar participants at Cambridge University, the Centre for Microdata and Practice and the Institute for Fiscal Studies at University College London, Erasmus University, the Institute for New Economic Thinking at the University of Southern California, the London School of Economics, the Research Institute of Economics and Management, Simon Fraser University, Syracuse University, Texas A&M, the Tinbergen Institute, the University of British Columbia, the Wang Yanan Institute for Studies in Economics, the University of Wisconsin at Madison, and Western University for helpful comments and suggestions. Finally, we thank the coeditor, the associate editor, and three anonymous referees for their helpful comments.

†619 Kern Bldg, University Park, PA 16802, suj14@psu.edu

‡616 Kern Bldg, University Park, PA 16802, joris@psu.edu

1. Introduction

1.1 Problem: We are interested in prediction problems in which the prediction of interest is only partially identified, particularly ones in which a point prediction is *also* desired. One prominent class of examples of models featuring partial identification consists of discrete decision games with complete information and Nash equilibria in pure strategies, in which *counterfactual prediction probabilities* are desired. There are many applications in which such games arise: examples can be found in Bresnahan and Reiss, 1991b; Kooreman, 1994; Soetevent and Kooreman, 2007; Jia, 2008; Ciliberto and Tamer, 2009; Bajari, Hong, and Ryan, 2010; Grieco, 2014, among others. The models associated with such games are often *incomplete* in the sense that the values of the payoff variables do not always specify a unique outcome. For instance, payoffs can be such that there is room for exactly one player to 'enter,' but it could be either one. The problems associated with identification and estimation of payoffs in such games, which are mostly a consequence of the incompleteness of the model, have been studied extensively and it is well–known how to address them. For instance, Tamer (2003) relies on the assumption that covariates have sufficiently large support and Bajari, Hahn, et al. (2011) use finite mixtures to deal with multiple equilibria. See also Kline and Tamer (2012) and Kline (2015). However, the incompleteness of the model poses unique challenges for the problem of counterfactual prediction even if the payoff structure is known in its entirety.

To facilitate the analysis plus the exposition and to maximize the ratio of intuition to technical detail, we focus on the simplest possible (and somewhat hackneyed) case that has all relevant features: a single–shot game with two players, and binary decisions, albeit that we allow for both strategic substitutability and strategic complementarity (Bulow, Geanakoplos, and Klemperer, 1985). The problems discussed here arise a fortiori in more general scenarios, including ones with more than two players, nonbinary decisions (Aradillas-Lopez, 2011), mixed strategies, dynamics (Aguirregabiria and Mira, 2007; Bajari, Benkard, and Levin, 2007; Pakes, Ostrovsky, and Berry, 2007), incomplete information (Seim, 2006; Liu, Vuong, and Xu, 2013; Xu, 2014), and more general solution concepts (Aradillas-López and Tamer, 2008; Kashaev, 2015; Kashaev and Salcedo, 2015; Magnolfi and Roncoroni, 2016). Although our general approach can be used to address many such problems, the results established here only provide intuition for similar models with more general discrete action spaces.

Player payoffs are functions of unobservables $\boldsymbol{e}$ and observables $\boldsymbol{x}$, both of which are known to the players.[1] The observables $\boldsymbol{x}$ are exogenous in the sense that they are independent of all unobservables in the model. For some combinations of $\boldsymbol{e}, \boldsymbol{x}$ there exists only a single Nash equilibrium in pure strategies and hence a single value y of the outcome variables $\boldsymbol{y}$: the set of such values of $\boldsymbol{e}$ given $\boldsymbol{x} = x$ is denoted by $S_y(x)$. For other combinations of $\boldsymbol{e}, \boldsymbol{x}$ there can exist multiple (in our case two) pure strategy Nash equilibria: $(1,0)$ and $(0,1)$ for the case of strategic substitutability and $(0,0)$, $(1,1)$ in the case of strategic complementarity. The model is hence 'incomplete' (Tamer, 2003) in the sense that the same values of the payoff variables $\boldsymbol{e}, \boldsymbol{x}$ can lead to different outcomes. For given x, $S_m(x)$ denotes the region of values e in which payoffs do not produce a unique

[1] We use bold typeface to denote random variables.

outcome. We therefore call $S_m(x)$ the *multiplicity region.* Most of the formal results of this paper are established for the strategic substitutability case, where the extension to strategic complementarity is natural. The following discussion will assume strategic substitutability unless otherwise specified. For the purpose of intuition, it is adequate to think of the game considered in this paper as an entry game, but the scope is broader than that.

We do not assume that we know why a particular outcome arises in the multiplicity region. Identification of the payoff structure does not require knowledge of such reasons (e.g. Tamer, 2003; Kline, 2015), but such knowledge *is* required for the counterfactual prediction probabilities to be identified. If such knowledge were available then it should be used to obtain point identification and our analysis becomes redundant. Thus, we will seek to *characterize* instead of to *specify* players' behavior in the multiplicity region. For an unknown function p, we *characterize* the behavior of players by a probability

$$\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}), \tag{1}$$

for which

$$\mathbb{P}\{\boldsymbol{y} = (1,0) \mid \boldsymbol{e} = e, \boldsymbol{p} = p, \boldsymbol{x} = x\} = p, \quad e \in S_m(x), \tag{2}$$

where $\boldsymbol{v}$ is unobserved and p belongs to the set $\mathcal{Q}(e,x) = \{p : \exists 0 < v \leq 1 : p(e,v,x) = p\}$. Thus,

$$p(e,v,x) = \mathbb{P}\{\boldsymbol{y} = (1,0) \mid \boldsymbol{e} = e, \boldsymbol{v} = v, \boldsymbol{x} = x\}, \quad e \in S_m(x).$$

The variable $\boldsymbol{v} \sim U(0,1)$ represents potential unobserved 'market' heterogeneity (separate from the payoff shifters $\boldsymbol{e}, \boldsymbol{x}$) that can affect the probability of the outcome $(1,0)$ occurring if $\boldsymbol{e} \in S_m(\boldsymbol{x})$. We shall assume the function p to be weakly increasing in $\boldsymbol{v}$ such that a higher value of $\boldsymbol{v}$ makes the outcome $(1,0)$ no less likely. We can derive results analogous to ours absent this assumption, but bounds obtained without the monotonicity assumption would be wider than the ones that we have and having p weakly increasing in $\boldsymbol{v}$ is attractive from an intuition perspective. Alternatively, one can think of $p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ as a reduced–form representation of $\boldsymbol{p}$ based on the conditional quantile function of $\boldsymbol{p}$ given $\boldsymbol{e}, \boldsymbol{x}$, where we allow for the possibility that $\boldsymbol{p}$ can be 'random' even after controlling for the payoff shifters $\boldsymbol{e}$ and $\boldsymbol{x}$;[2] hence the characterization $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ cannot be tested or refuted. Indeed, there always exists an alternative characterization in which $\boldsymbol{v}$ is absent. However, by using the characterization that includes $\boldsymbol{v}$ the class of counterfactuals we can discuss becomes richer: in section 2, we provide several examples in which payoff shifters alone do not determine the equilibrium selection probability.

Since $\boldsymbol{p}$ is the *probability* that $(1,0)$ is realized in the multiplicity region, we need another (uniformly distributed) unobservable $\boldsymbol{u}$ to complete the description of which outcome is realized: $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}$ are mutually independent, which is not restrictive for reasons discussed in section 2. Thus, if $\boldsymbol{e} \in S_m(\boldsymbol{x})$ then $\boldsymbol{y} = (1,0)$ arises if and only if $\boldsymbol{u} \leq \boldsymbol{p}$. Indeed, we can write $\boldsymbol{y} = y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x})$, where y is an unknown function to be defined formally in section 2. Recall that for $\boldsymbol{e} \notin S_m(\boldsymbol{x})$, the payoff variable values specify a unique outcome and hence

[2] For instance, suppose that the conditional distribution function $F_{\boldsymbol{p}|\boldsymbol{e},\boldsymbol{x}}$ of $\boldsymbol{p}$ given $\boldsymbol{e}, \boldsymbol{x}$ is strictly increasing. Then we can define $\boldsymbol{v} = F_{\boldsymbol{p}|\boldsymbol{e},\boldsymbol{x}}(\boldsymbol{p}|\boldsymbol{e}, \boldsymbol{x})$, which is uniformly distributed and independent of $\boldsymbol{e}, \boldsymbol{x}$ by construction. We can now set $p(e,v,x) = F^{-1}_{\boldsymbol{p}|\boldsymbol{e},\boldsymbol{x}}(v \mid e, x)$. If $\boldsymbol{p}$ is discrete given $\boldsymbol{e}, \boldsymbol{x}$, then the arguments can be extended by using the usual generalized–inverse adjustments for CADLAG functions.

the values of $\boldsymbol{u}, \boldsymbol{v}$ are then irrelevant. For $\boldsymbol{e} \in S_m(\boldsymbol{x})$, however, all four variables $(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x})$ are germane.

Recall from (1) that we use the characterization $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ and that in the multiplicity region $\boldsymbol{u} \leq \boldsymbol{p} \Leftrightarrow \boldsymbol{y} = (1, 0)$. There are examples in the literature (e.g. Bjorn and Vuong, 1984; Jia, 2008; Grieco, 2014) that are nested in our characterization: section 2 contains a discussion. We know of no other papers in the econometrics or applied microeconomics literature that allow for heterogeneity like $\boldsymbol{v}$ in the present context.

Note that we have two random variables, $\boldsymbol{u}$ and $\boldsymbol{v}$, that together represent unobserved market heterogeneity not affecting payoffs: $\boldsymbol{v}$ affects the *probability* $\boldsymbol{p}$ that a particular equilibrium is reached in the multiplicity region and a combination of $\boldsymbol{u}$ and $\boldsymbol{p}$ produces the actual outcome. The reason for having both $\boldsymbol{u}$ and $\boldsymbol{v}$, then, is that we want to allow for counterfactuals in which some, but not necessarily all, of the market heterogeneity is fixed. Further, a model with only $\boldsymbol{v}$ is more general than a model with only $\boldsymbol{u}$ since p need not be invertible in v. Finally, there are examples in the economic theory literature in which $\boldsymbol{u}$ and $\boldsymbol{v}$ have a natural structural interpretation that is lost once either of the two variables is removed from the model; section 2 contains several such examples. Losing the structural interpretation implies losing the ability to conduct a number of interesting thought experiments.

We now turn our attention to the issue of counterfactual prediction. We denote the counterfactual outcomes of $\boldsymbol{x}, \boldsymbol{y}$ by $\boldsymbol{x}^*, \boldsymbol{y}^*$, respectively, and are interested in predicting $\boldsymbol{y}^*$. Since $\boldsymbol{y}^*$ is binary, we focus on the conditional probabilities of the counterfactual outcome given the observables, i.e.

$$\mathbb{P}(\boldsymbol{y}^* = y^* \mid \boldsymbol{x}^* = x^*, \boldsymbol{x} = x, \boldsymbol{y} = y), \tag{3}$$

where $y, y^* \in \mathcal{S} = \{(0,0), (0,1), (1,0), (1,1)\}$. We also condition on $\boldsymbol{x}, \boldsymbol{y}$ in (3) since their values contain information about the values of $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$.

The objective in this paper is to construct predictions of counterfactual outcomes at the population level if parts of the environment are held fixed: we take the distribution of observables as known. In all counterfactuals, $\boldsymbol{x} = x$ is replaced by a new value $\boldsymbol{x}^* = x^*$, where $x^* = x$ is allowed. The definition of the counterfactual outcome of interest $\boldsymbol{y}^*$ will further depend on how the unobservables $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$ are treated. In the simplest (and least interesting) case, $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$ are replaced with independent copies $\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*$ (i.e. $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$ are redrawn), in which case $\boldsymbol{y}^* = y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{x}^*)$. Then, the *regression prediction*

$$q(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) = \mathbb{P}(\boldsymbol{y} = y^* \mid \boldsymbol{x} = x^*) \tag{4}$$

is the same as (3), which is hence conveniently point–identified. However, if at least one of the unobservables $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$ is not redrawn (i.e. is the same in the counterfactual), then the regression prediction (4) differs from the counterfactual prediction probability (3).

We consider various scenarios where different combinations of $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$ are not redrawn. For example, one might be interested in the counterfactual where $\boldsymbol{e}$ is fixed: fixing only $\boldsymbol{e}$ is a different thought experiment than fixing both $\boldsymbol{e}$ and $\boldsymbol{v}$ or fixing $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$. We study all eight scenarios but in the main text we focus on two representative cases, i.e. fixing $\boldsymbol{v}$ and fixing both $\boldsymbol{e}$ and $\boldsymbol{v}$. We provide results for the leading cases in sections 3.1 and 3.2 and for

the other scenarios in section 3.3 and appendix B. We denote the counterfactual prediction probabilities in (3) by respectively

$$q_v(y^* \mid x^*, x, y) \quad \text{and} \quad q_{ev}(y^* \mid x^*, x, y),$$

when $\boldsymbol{v}$ is fixed and when both $\boldsymbol{e}$ and $\boldsymbol{v}$ are fixed: the subscripts denote which variables are fixed. We define similar symbols analogously. Unlike the regression prediction q, neither q_v nor q_{ev} is generally point–identified because they depend on the function p, which is *not* (point–) identified.

In addition to counterfactual prediction probabilities of individual outcomes, we consider the case in which it is the number of competitors instead of their identity that is of interest, as in e.g. Bresnahan and Reiss (1991a). For instance, a government may wish to minimize the occurrence of 'food deserts' and a regulatory agency would typically prefer duopoly to monopoly, but may be less concerned about the identity of the monopolist. We show in section 3 that when fixing $\boldsymbol{v}$, the counterfactual probability of having a monopoly collapses to the corresponding regression prediction and is hence point–identified, although q_v is not point–identified.[3] However if both $\boldsymbol{e}$ and $\boldsymbol{v}$ are fixed then even the probability of monopoly is only partially identified.[4]

There are several solutions to our partial identification problem. One conceptually straightforward approach is to treat the counterfactual prediction problem itself as a partial identification problem and use bounds on the counterfactual probabilities, i.e. to work out the (sharp) *identified set* of counterfactual probability values.[5] Inference under partial identification is now a well–studied topic, also: for instance, a robust Bayes approach (e.g. Kitagawa, 2012) is a possibility. We determine the sharp identified sets for both $q_v(y^* \mid x^*, x, y)$ and $q_{ev}(y^* \mid x^*, x, y)$ in sections 3.1 and 3.2, respectively: the bounds for all other cases are provided in appendix B.

Reguant (2016) considers a related problem; she proposes a general methodology for *computing* valid bounds (via mixed integer programming) to equilibrium counterfactual outcomes, which are functions of equilibrium strategies. Her results, while interesting, are of limited relevance here. We do provide bounds to our counterfactual prediction probabilities which we show (in our proofs) can be characterized as constrained optimization problems, but they are sharp explicit bounds. Further, the constrained optimization problems in our case feature inequality constraints that are different for upper and lower bounds and obtaining bounds is not the main focus of our paper.

While bounds are useful and using point predictions *alone* can be suboptimal (Manski, 2015), point predictions are nevertheless valuable. First, for policy purposes, a 'best guess' may be desired since the optimal policy may depend on the value of the predicted quantity.[6] Second, bounds can be too wide to be practically useful. Third, if bound predictions are used as an input to something else, e.g. a function or an optimization problem, then the identified set of the ultimate object can be wider still or can have the consistency of Swiss

[3] Point identification of the number of competitors does not extend to the case with more than two choices and/or players.

[4] We thank Ken Hendricks for suggesting this case.

[5] An early work on bounds can be found in Fréchet (1935). We thank an anonymous referee for this reference.

[6] See Jun and Pinkse (2017) for an analogous example in the estimation of auction models.

cheese instead of being convex, as is desirable for inference with set–identified models.[7] Fourth, by their very definition, bounds correspond to extreme cases: it is ex ante unclear how much attention one should pay to the most extreme eventualities. Fifth and specific to our prediction problem, the bounds for some counterfactuals are discontinuous in x and the jump in the bound can be as large as 50 percentage points, i.e. half the parameter space: it is unlikely to inspire policy makers with confidence if counterfactual predictions are highly unstable. The reasons for this discontinuity will be explored, e.g. in section 7. In sum, it is useful to have point predictions *in addition to* bounds.

One intuitive point prediction choice is the midpoint θ_m of the identified set. Indeed, Song (2014) has shown that (an efficient estimator of) the midpoint of an identified set can be justified on decision–theoretic grounds by using a 'local asymptotic minimax regret' criterion. But Song's environment is substantially different from ours and he is concerned with estimation, not prediction. Here, because the bounds can be discontinuous in x, so can the midpoints. Indeed, midpoint predictions can have discontinuities in x that are as large as 25 percentage points. Further, midpoint predictions are by definition an average of two extremes. Third, as we discuss in section 3.3 that q_e, q_{ev}, q_{euv} are ranked[8] but their bounds coincide. In other words, even though it is known that e.g. (omitting arguments) $q_e \leq q_{ev} \leq q_{euv}$, the bounds and hence the midpoint predictions are the same. The maximum entropy predictions discussed below do not have this problem. Finally, there is an additional 'inconsistency' problem: as we show in section 7.1, there does not generally exist a single function p that can generate midpoint predictions for different counterfactuals.

We develop several alternatives to simply using bounds, i.e. to taking the midpoint. Our preferred method is to use maximum entropy from information theory. We further develop two additional decision–theoretic methods. Descriptions of and tradeoffs between these three approaches are discussed in section 1.2.

1.2 Alternatives to bounds: Decision–theoretic alternatives to using bounds entail defining a loss function ℓ which measures the distance between the infeasible prediction (which depends on p) and the 'decision' (which does *not* depend on p). We focus on decisions *at the population level* that use the complete knowledge of the model, not a particular set of observations. For example, if θ_0 is the prediction of interest then complete knowledge of the model only provides the bounds $\mathcal{B}_L \leq \mathcal{B}_U$ for θ_0 and these bounds depend only on population objects and they are hence point–identified. Then, the (unconstrained) minmax decision with a quadratic loss is formulated by $\min_{\theta \in [\mathcal{B}_L, \mathcal{B}_U]} \max_{\theta_0 \in [\mathcal{B}_L, \mathcal{B}_U]} (\theta - \theta_0)^2$, which yields the solution $\theta = (\mathcal{B}_L + \mathcal{B}_U)/2$, i.e. the midpoint decision. There is no randomness here and therefore loss and risk are the same thing: so $\max_{\theta_0 \in [\mathcal{B}_L, \mathcal{B}_U]} (\theta - \theta_0)^2$ is the maximum risk.

As mentioned earlier, the unconstrained minmax approach leads trivially to the midpoint prediction, which has downsides that have already been discussed. We consider two alternatives: adding a constraint based on entropy to the minmax formulation and

[7] The nonconvexity problem arises e.g. in Haile and Tamer (2003), which they address by imposing additional technical assumptions on the shape of a 'pseudoprofit function.'

[8] The order can be ascending or descending, depending on the values of y, y^*.

minimizing average risk that uses a 'prior' on p.

The first alternative, which is described in section 5, assumes a symmetric convex loss function, e.g. a quadratic, and minimizes maximum loss/risk, where the maximum is now taken over all functions p whose entropy is not below some specified bound $\mathcal{E}^*$. This is a hybrid method in that it uses both information theory and decision theory. Indeed, if $\mathcal{E}^*$ is chosen to be the maximum attainable value then one obtains the maximum entropy solution. At the other extreme, if one chooses $\mathcal{E}^* = -\infty$ then one obtains the midpoint solution. A different way of looking at this is to note that the entropy restriction reduces the size of the identified set and that the hybrid method then selects the midpoint of this smaller identified set, which also happens to be an interval.

The second alternative is investigated in section 6. There, using Dirichlet processes we specify a new class of probability distributions over p–functions that is consistent with the distribution of observables.

However, whether one uses the approach of section 6 or something else, the decision–theoretic approaches do not solve the fundamental problem at hand. For instance, there exists no *natural* probability measure on the parameter space, i.e. the class of functions p: depending on what measure one assigns to the parameter space any prediction in the identified set can be generated (Aumann, 1961). In other words, the minimum–average–risk approach relocates the problem of choosing a prediction from the identified set to the choice of a measure on the parameter space that is consistent with the distribution of observables.[9] Further, the Dirichlet approach only uses discrete probability distributions, whereas $\boldsymbol{p}$ can — or indeed is more likely to be — continuously distributed. Nevertheless, for the Dirichlet–based probability measures considered in section 6, the decision–theoretic approach generates predictions that are similar to our preferred method, *maximum entropy,* which is discussed in section 4 in detail.

With the maximum entropy method of the information theory literature (Jaynes, 1957a; Jaynes, 1957b; Golan, Judge, and Miller, 1996), one selects the probability distribution that best represents the current state of knowledge as measured by the entropy. The maximum entropy solution complements the information contained in the data with a criterion that is consistent with the (at least) seven hundred year old principle of *Occam's razor:* if there are multiple explanations for the same phenomenon then one should choose the simplest one. In our context, this translates into choosing the function p such that out of all candidates that are consistent with the distribution of observables $(\boldsymbol{x}, \boldsymbol{y})$ and the model assumptions, the random variable $\boldsymbol{p}$ has the distribution that is closest to a uniform (given $\boldsymbol{e}, \boldsymbol{x}$).[10] There are numerous philosophical justifications for the application of Occam's razor; we refer the reader to Baker (2013).

There are justifications for the use of maximum entropy other than Occam's razor. Indeed, Golan (2017) lists sets of axioms for the problem of choosing a probability distribution absent identification. Theorem 2.1 in Golan (2017) establishes that the maximum entropy solution is the only one that satisfies five stated axioms.[11]

Finally, although information theory and decision theory both have their respective

[9] In the other possibility there is an issue of how to choose $\mathcal{E}^*$. Setting $\mathcal{E}^*$ to the two extreme possibilities leads to the maximum entropy solution or the midpoint prediction, which we have already discussed above.

[10] With maximum entropy, the closest choice is unique up to trivial deviations.

[11] Golan (2017) attribute their theorem 2.1 to Shore and Johnson (1980).

merits, one does not need to make a choice of one above another. Indeed, as Topsøe (1979), Harremoës and Topsøe (2001), and Grünwald and Dawid (2004) show, maximum entropy itself has a decision–theoretic interpretation as well. For instance, Grünwald and Dawid (2004) show that maximizing entropy and minimizing worst–case expected loss are each other's dual in a statistical game in which nature reveals values from an unknown distribution, a decision maker specifies a distribution, and loss is measured by a log–loss function.[12]

Maximum entropy is sometimes confused with the Bayesian approach. Although one of Jaynes's intentions behind maximum entropy was to provide Bayesians with a sensible prior, the method itself can equally be used in a classical context. Further, since maximum entropy entails an optimization problem whose constraints correspond to the information available in the data, the maximum entropy solution coincides with the standard classical solution in the case of point identification: the constraints then provide a unique solution. In other cases, maximum entropy provides the minimal amount of additional information needed to provide uniqueness. In other words, maximum entropy can be thought of as providing 'second class' information.[13]

Here, the function p selected by maximum entropy is flat in e in the multiplicity region, which is consistent with equilibrium–selection mechanisms used in the literature (e.g. Bjorn and Vuong, 1984; Jia, 2008; Bajari, Hong, and Ryan, 2010). Note, however, that in our case this is an *outcome* instead of an *assumption.*

One can cook up measures other than the one provided by maximum entropy, so using maximum entropy is not altogether free of arbitrariness. For instance, using a distance criterion other than the one used in maximum entropy (i.e. Kullback–Leibler divergence) yields a different solution. Using maximum entropy on a monotonic transformation of $\boldsymbol{p}$ will generally also lead to a different solution. Indeed, the exercise of making a single guess about the value of a parameter that is only partially identified is inherently arbitrary. But the use of maximum entropy has been justified extensively in the information theory literature: recall the five axioms listed in Golan and see e.g. Cover and Thomas (2012). Second, since the random variable at the center of the problem is the selection probability $\boldsymbol{p}$, it is more natural to deal with the distribution of $\boldsymbol{p}$ than that of e.g. $\boldsymbol{p}^2$. Further, with maximum entropy it is straightforward to introduce additional information (restrictions) to the problem whereas with the decision–theory based Dirichlet–like method doing so is both complicated and cumbersome. As noted above, for the choices made in section 6, the Dirichlet and maximum entropy approaches appear to generate similar predictions while the maximum entropy approach is considerably easier to implement. Finally, as we show in section 8, it is straightforward to address inferential issues for maximum–entropy predictions. Compared to Bayesian alternatives such as pragmatic Bayesian approxima-

[12] If nature draws a value z of $\boldsymbol{z}$ from a density f_0 and a statistician chooses f to specify the distribution, then the log–loss is defined by $-\log f(z)$: i.e. the smaller value f takes at z, the larger the loss is. Then, the expected log–loss is given by $-\mathbb{E}\log f(\boldsymbol{z})$. Grünwald and Dawid (2004) note that maximizing entropy with respect to f_0 can be written as $\max_{f_0}\ \min_f -\mathbb{E}\log f(\boldsymbol{z})$ and they discuss switching the order of max and min to reformulate it as a minmax decision problem.

[13] There are other contexts in which one wishes to select a single function from a set, e.g. theorem A.1 in Chen and Pouzo (2012), but the context plus the considerations and implications of choosing one function over another are entirely different.

tions described in e.g. Murray and Snelson (2006), our approach turns out to be a simple application of the classical Delta method once the payoff parameters are estimated.

1.3 Contribution: In sum, we see the main contributions of this paper as follows. To our knowledge, we are the first to study the problem of counterfactual point–prediction in games of complete information featuring a partially identified infinite–dimensional parameter. We derive formulas for the infeasible counterfactual prediction probabilities $q_v(y^* \mid x^*, x, y)$ and $q_{ev}(y^* \mid x^*, x, y)$ as functions of the unknown, partially identified, function p and construct corresponding identified sets in the form of sharp bounds; see section 3. We further derive bounds for the case in which the object of interest is the number instead of the identity of 'entrants.' We propose, compare, and contrast various point prediction methods. We develop a new decision theory–based point prediction method in the spirit of the Dirichlet–process literature in section 6 and a new point prediction method based on the maximum entropy concept from the information theory literature in section 4.[14] The hybrid method described in section 1.2 is developed in section 5. We demonstrate the virtues of the maximum entropy approach (and to a lesser extent the Dirichlet approach) in a number of examples in section 7. Section 8 demonstrates how inference can be conducted on our counterfactual prediction probabilities. Finally, analogs of our maximum entropy approach can be used to do counterfactual analysis in other models with partial identification, albeit that in our case the partially identified parameter is a function, which is more complicated than if it were finite–dimensional.

2. Setup

2.1 Binary decision games: We consider standard two player binary decision games with complete information and pure strategies. In the normal form in figure 1, the functions ς_1, ς_2 and payoff variables $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{x}$ are known to both players: their respective choices are denoted $\boldsymbol{y}_1, \boldsymbol{y}_2$ and the function ς_i depends on x and y_{3-i} for $i = 1, 2$. We assume that $\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2$ are observable to the econometrician, but $\boldsymbol{e}_1, \boldsymbol{e}_2$ are not. Further, $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2)$ is assumed to be independent of $\boldsymbol{x}$, which can be vector–valued. We assume that both the ς_i–functions and the distribution of $\boldsymbol{e}$ are known or, failing that, are identified, e.g. via a result like that in Tamer (2003): we do this because identification and estimation of payoffs have already been studied extensively.[15]

For given values of e and x, the normal form in figure 1 allows for either strategic complementarity ($\varsigma_i(x, 0) < \varsigma_i(x, 1)$ for $i = 1, 2$; you choosing option 1 increases the payoff for me to choose option 1) or strategic substitutability ($\varsigma_i(x, 0) > \varsigma_i(x, 1)$ for $i = 1, 2$; you choosing option 1 reduces the payoff for me to choose option 1). In many cases there is

[14] Others have used the notion of entropy in the context of partial identification, albeit with a different purpose. Indeed, Schennach (2014) uses entropy as a way of *constructing* the identified set in an estimation problem, whereas we use it to *select* a point from the identified set in a prediction problem. However, her problem and method are much closer substitutes to Galichon and Henry (2011) than to ours.

[15] Our counterfactual analysis does not allow the payoff structure or the distribution of payoff variables to change. However, such changes could in principle be accommodated as long as the new payoff structure and distribution of payoff variables are given or can be identified using other methods.

1 \ 2	0	1
0	$0, 0$	$0, \iota_2(x,0) + e_2$
1	$\iota_1(x,0) + e_1, 0$	$\iota_1(x,1) + e_1, \iota_2(x,1) + e_2$

Figure 1: Normal form

only a single Nash equilibrium in pure strategies, but there are also cases in which there are two pure Nash equilibria.[16]

Indeed, consider a given value x (of $\boldsymbol{x}$) and let $\bar{\iota}_i(x) = \max\{\iota_i(x,0), \iota_i(x,1)\}$ and $\underline{\iota}_i(x) = \min\{\iota_i(x,0), \iota_i(x,1)\}$. Then there are five distinct regions: four regions in which there is a unique Nash equilibrium in pure strategies $(S_{00}(x), S_{01}(x), S_{10}(x), S_{11}(x))$ and one *multiplicity region*, $S_m(x)$; see figure 2 for examples. Thus, $S_{10}(x)$ is the set of $e = (e_1, e_2)$ values for which $(1,0)$ is the unique Nash equilibrium if $\boldsymbol{x} = x$. There are two pure strategy Nash equilibria whenever $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2)$ falls in the multiplicity region $S_m(x) = [-\bar{\iota}_1(x), -\underline{\iota}_1(x)] \times [-\bar{\iota}_2(x), -\underline{\iota}_2(x)]$.

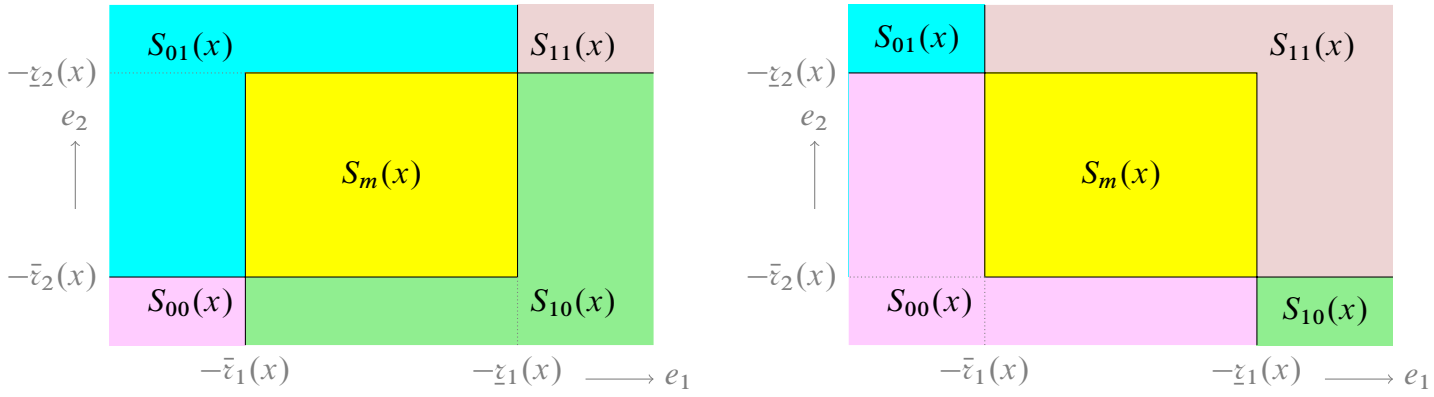

Figure 2: Regions of e–values corresponding to either a specific equilibrium or multiple ones for the cases of strategic substitutability (i.e. $\bar{\iota}_i(x) = \iota_i(x,0)$ for $i = 1,2$, on the left) and strategic complementarity (i.e. $\bar{\iota}_i(x) = \iota_i(x,1)$ for $i = 1,2$, on the right).

If $\boldsymbol{e}$ belongs to $S_m(x)$ and $\iota_i(x,0) < \iota_i(x,1)$ for $i = 1,2$ then one obtains a coordination game with $(0,0)$ and $(1,1)$ as the two Nash equilibria in pure strategies. If instead $\boldsymbol{e}$ belongs to $S_m(x)$ and $\iota_i(x,0) > \iota_i(x,1)$ for $i = 1,2$ then we have an entry game with $(0,1)$ and $(1,0)$ as the two Nash equilibria in pure strategies. We discuss examples of coordination games with two (pure strategy Nash) equilibria that are consistent with the normal form in figure 1 in section 2.3.1 and entry games with two equilibria in section 2.3.2. These examples illustrate the need for the additional random variable $\boldsymbol{v}$ used in the sketch of the counterfactual prediction problem in section 2.2. A discussion of the counterfactuals themselves is postponed until section 3.

[16] Since $\boldsymbol{e}$ is continuously distributed, the probability of more than two pure strategy Nash equilibria is zero. If $\boldsymbol{e}$ were discrete then it is possible that the probability that $\boldsymbol{e}$ were e.g. at the intersection of $S_m, S_{11}, S_{01}, S_{10}$ would be nonzero.

2.2 Prediction: One can relate the outcome $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ to payoff variables $(\boldsymbol{e}, \boldsymbol{x})$ by

$$\begin{cases} \boldsymbol{y}_1 = \mathbb{1}\{\iota_1(\boldsymbol{x}, \boldsymbol{y}_2) + \boldsymbol{e}_1 \geq 0\}, \\ \boldsymbol{y}_2 = \mathbb{1}\{\iota_2(\boldsymbol{x}, \boldsymbol{y}_1) + \boldsymbol{e}_2 \geq 0\}, \end{cases} \tag{5}$$

where $\boldsymbol{x}$ is a vector of exogenous covariates, $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2)$ are errors that are independent of $\boldsymbol{x}$. The model in (5) is *incomplete* in that $\boldsymbol{x}$ and $\boldsymbol{e}$ do not necessarily determine a unique outcome $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ due to the possible presence of multiple Nash equilibria, as noted above (see e.g. Bresnahan and Reiss, 1991a; Tamer, 2003). Incompleteness, here, means that there can be multiple pairs $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ that satisfy (5).

With the exception of section 2.3.1, we assume that there is strategic substitutability, i.e. $\iota_i(x, 1) \leq \iota_i(x, 0)$ for $i = 1, 2$ and all x. We do this merely to avoid duplication of notation and arguments.

The properties of the model in (5) have been studied extensively and conditions under which ι_1, ι_2 and the distribution of $(\boldsymbol{e}, \boldsymbol{x})$ are identified are well–understood (e.g. Tamer, 2003). From hereon, we therefore take the functions ι_1, ι_2 and the distribution of $(\boldsymbol{e}, \boldsymbol{x})$ as given.

Since we are taking the functions ι_i and the distribution of $(\boldsymbol{e}, \boldsymbol{x})$ as given, the S–regions are given, also, and so are their probabilities of occurring, i.e.

$$\pi_y(x) = \mathbb{P}\{\boldsymbol{e} \in S_y(x)\}, \qquad \pi_m(x) = \mathbb{P}\{\boldsymbol{e} \in S_m(x)\}.$$

We further denote

$$\tilde{\pi}_y(x) = \mathbb{P}(\boldsymbol{y} = y \mid \boldsymbol{x} = x).$$

As mentioned in the introduction, we are interested in predicting the counterfactual outcome $\boldsymbol{y}^*$ if the same game is played again under various circumstances. To do this, we have to provide a (partial) description of the equilibrium determination process in the multiplicity region.

Recall from (2) that for some random variable $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ and all x, if $e \in S_m(x)$ then

$$\forall\, p \in \mathcal{Q}(e, x) : \mathbb{P}\big\{\boldsymbol{y} = (1, 0) \,\big|\, \boldsymbol{e} = e, \boldsymbol{p} = p, \boldsymbol{x} = x\big\} = p, \tag{6}$$

where $\boldsymbol{v}$ has a standard uniform distribution and is independent of $\boldsymbol{e}, \boldsymbol{x}$. The function p is unknown and (weakly) monotonic in v for every e, x.

So from the perspective of the researcher, there is a 'probability' $\boldsymbol{p}$ that a particular equilibrium arises in the multiplicity region, where $\boldsymbol{p}$ can depend on (both observable and unobservable) market heterogeneity. This looks similar to the equilibrium selection mechanism used in Grieco (2014), where $p(e, v, x)$ depends only on the payoff shifters e, x, and also nests the structures in Bjorn and Vuong (1984), where $p(e, v, x) = \mu_m(x)$, and Jia (2008), where $p(e, v, x) = 1$. Our specification is a *characterization* instead of an *assumption* and it is consistent with *any* equilibrium refinement.[17] The same can be said

[17] A. Kalai and E. Kalai (2012) (and references therein) provide refinements for two by two games, albeit that in A. Kalai and E. Kalai (2012) utility is transferable, which means that the outcome would not have to correspond to a Nash equilibrium in pure strategies in the corresponding game without transfers. Other references on equilibrium selection include Harsanyi and Selten (1988) and Kandori, Mailath, and Rob (1993), which we will discuss in more detail in section 2.3.

for Grieco (2014) — and hence also for Bjorn and Vuong (1984) — but, as noted in the introduction, we want to allow for the possibility that $\boldsymbol{v}$ is fixed and $\boldsymbol{u}$ changes, i.e. that the unobserved payoff–irrelevant heterogeneity that affects the *equilibrium–selection probability* is fixed but that the equilibrium selection probability itself can be different from zero and one. Indeed, in other contexts it is common to have vector–valued unobservables (e.g. Athey and Imbens, 2007; Briesch, Chintagunta, and Matzkin, 2012): the problem at hand can often be described by a scalar unobservable, but doing so changes the nature of the problem as is argued in Kasy (2011).

Furthermore, as will become apparent in sections 2.3.1 and 2.3.2, there are instances in which $\boldsymbol{u}, \boldsymbol{v}$ do have a structural interpretation that is lost once either of them is dropped and hence having both $\boldsymbol{u}$ and $\boldsymbol{v}$ allows for more interesting counterfactuals than would a single such unobservable. The flipside of this argument is that a characterization in terms of two unobservables $\boldsymbol{u}, \boldsymbol{v}$ does limit the nature of the counterfactual experiments discussed in section 3 if the underlying structural model of players' behavior in the multiplicity region features multiple unobservables that feed into $\boldsymbol{v}$ and that an empiricist is interested in the effects of holding fixed one or more of the structural model unobservables instead of $\boldsymbol{v}$: our paper does not speak to that possibility. However, the presence of two payoff–irrelevant unobservables $(\boldsymbol{u}, \boldsymbol{v})$ instead of one, as in Grieco (2014), increases the likelihood that any explicit model of behavior in the multiplicity region can be mapped to our specification.

Equation (6) is silent about the behavior of $\boldsymbol{p}$ when $\boldsymbol{e} \notin S_m(\boldsymbol{x})$ since the value of $\boldsymbol{p}$ is then immaterial for the selection of $\boldsymbol{y}$. Even if $\boldsymbol{e} \in S_m(\boldsymbol{x})$, however, (6) does not restrict the dependence of $\boldsymbol{e}$ and $\boldsymbol{p}$.

In view of (6) we can represent the outcome $\boldsymbol{y}$ by

$$\boldsymbol{y} = y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}) = \begin{cases} y, & \boldsymbol{e} \in S_y(\boldsymbol{x}), \\ (1,0), & \boldsymbol{e} \in S_m(\boldsymbol{x}) \text{ and } \boldsymbol{u} \leq p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}), \\ (0,1), & \boldsymbol{e} \in S_m(\boldsymbol{x}) \text{ and } \boldsymbol{u} > p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}), \end{cases} \tag{7}$$

where $\boldsymbol{u}$ is uniform and independent of $\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}$.

In view of (7), (6) *completes* the model in that it enables us to determine the outcome as a function of observables and unobservables. However, since p is unknown — and indeed not identified — the structure that (6) imposes is by itself of limited help for the prediction problem studied in this paper.

2.3 Equilibrium selection: As promised, we now discuss examples to illustrate the importance of allowing for an additional unobservable $\boldsymbol{v}$; the fact that $\boldsymbol{p}$ *can* depend on the payoff shifters $\boldsymbol{e}$ and $\boldsymbol{x}$ is intuitive. If $\boldsymbol{e}$ belongs to $S_m(x)$ then there are two pure strategy Nash equilibria. Recall that we *characterize* equilibrium selection by a random variable $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$, where $\boldsymbol{p}$ is the *probability* that a given equilibrium, e.g. $(1,0)$, is selected. The examples in this subsection are divided into two parts, depending on the nature of whether there is strategic complementarity or strategic substitutability.

2.3.1 Coordination games: As noted, if $\tau_i(x,0) < \tau_i(x,1)$ for $i = 1,2$ then one obtains a coordination game in $S_m(x)$ in which $(0,0)$ and $(1,1)$ are (the only) pure strategy Nash

equilibria. Since there are two pure strategy Nash equilibria, the payoffs provide insufficient information to produce a unique outcome. We now discuss equilibrium selection for this case.

As an example, consider the payoff matrix in figure 3.

1 \ 2	0	1
0	0,0	0,-2
1	-2,0	1,1

Figure 3: Simple coordination game

Harsanyi and Selten (1988) refer to $(1,1)$ as the *payoff–dominant equilibrium* and to $(0,0)$ as the *risk–dominant equilibrium:* $(0,0)$ is strictly risk–dominant in this example. Payoff–dominance is equivalent to the traditional definition of efficiency and, as noted in van Damme (1995), is a notion of collective rationality.

In contrast, risk–dominance is about individual rationality. In the example above, $(0,0)$ is risk–dominant because each player will play 0 if she believes that her opponent is equally likely to play either 0 or 1. More generally, i.e. without symmetry, let Λ_i denote the probability with which player $3-i$ plays 0 according to player i's beliefs and let Λ_i^* be the value of Λ_i for which player i is indifferent between playing 0 and 1. Then $(0,0)$ is risk–dominant if it is a Nash equilibrium and $\Lambda_1^* + \Lambda_2^* \leq 1$. In this example, $\Lambda_1^* + \Lambda_2^* = 2/3 < 1$. Note that the same equilibrium can be both risk–dominant and payoff–dominant and need not be either.

There is an extensive literature on equilibrium selection in this context in game theory (e.g. Foster and Young, 1990; Kandori, Mailath, and Rob, 1993). A common question in this literature is whether or not a payoff–dominant equilibrium is more appealing than a risk–dominant one. For instance, Kandori, Mailath, and Rob (1993) propose an evolutionary model and they discuss the probability of a risk–dominant equilibrium being played. There, the same 2×2 stage game is repeated indefinitely, players are chosen at random from a finite set, and a player from the set is replaced with a new uninformed player at random times according to a probability Kandori, Mailath, and Rob call the 'mutation rate.' They show that the probability that the risk–dominant equilibrium is played in the stage game is in the long run independent of initial conditions (the initial strategies) and can be any number between 0 and 1, with the number depending on both the payoffs and the mutation rate.[18] Indeed, the probability that the payoff–dominant equilibrium is played is increasing in the mutation rate. Thus, the payoff values in figure 3 are insufficient to determine which equilibrium is more likely to occur at a given point in time.[19]

In our context, then, $\boldsymbol{v}$ is the mutation rate and $p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ the probability that $(1,1)$ is selected. In order to determine which equilibrium is played in the stage game we need an

[18] We focus on coordination games here, but Kandori, Mailath, and Rob's analysis applies more generally. In particular, it also applies to entry games, as will become apparent in section 2.3.2.

[19] Note that these are unconditional probabilities. The probability that a particular equilibrium occurs at time $t+1$ if the same equilibrium occurred at time t is much higher than the unconditional probability if the mutation rate is low.

additional standard uniform random variable $\boldsymbol{u}$ such that in the multiplicity region,

$$\boldsymbol{u} \leq p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}) \Leftrightarrow \boldsymbol{y} = (1,1).$$

So the mutation rate $\boldsymbol{v}$ and the payoff variables $\boldsymbol{e}, \boldsymbol{x}$ determine the probability that $(1,1)$ is selected in the steady state if $\boldsymbol{e} \in S_m(\boldsymbol{x})$ and $\boldsymbol{u}$ determines which equilibrium is selected in the stage game at the time of 'observation.' We may call $\boldsymbol{u}$ a 'sunspot' in that it is an extrinsic random component that affects the actual outcome of the game (e.g. Cass and Shell, 1983; Aguirregabiria and Mira, 2018).

The example described above is just one of several conceivable examples generating two errors $\boldsymbol{u}, \boldsymbol{v}$ that are unrelated to the payoffs. In any such example, the variables $\boldsymbol{u}, \boldsymbol{v}$ can vary across markets.

The fact that $\boldsymbol{u}$ and $\boldsymbol{v}$ have different interpretations does not mean that both are required to describe equilibrium selection. Indeed, one could feed $\boldsymbol{u}$ and $\boldsymbol{v}$ to a single unobserved variable and represent the selection probability as if the payoff variables were its sole determinants. Put differently, one could define the selection probability without conditioning on $\boldsymbol{v}$. In fact, this is the route chosen by other authors, including Grieco (2014) and Aguirregabiria and Mira (2018).

However, if one only had one of these random variables, say $\boldsymbol{u}$, then one could not attribute meanings to each of $\boldsymbol{u}, \boldsymbol{v}$ as in the Kandori, Mailath, and Rob (1993) example above. This is not necessarily a limitation when the purpose of the analysis is something other than the identification of counterfactuals. However, the focus of our paper *is* on the counterfactuals and the nature of the counterfactual experiments that we consider is therefore important. Indeed, if we only had a single market heterogeneity unobservable then (in the above example) we could no differentiate between holding constant the mutation rate $\boldsymbol{v}$ and holding constant the state $\boldsymbol{u}$ that the market is in: the mutation rate is a long run object whereas $\boldsymbol{u}$ is specific to the stage game.

More generally, without $\boldsymbol{v}$ there are only two counterfactuals that can be considered in $S_m(x)$ if the payoff variables are the same in the counterfactual as in the 'data:' (*1*) the counterfactual outcome is the same as the 'observed' outcome; (*2*) the counterfactual outcome is, conditional on the values of the payoff variables, independent of the 'observed' outcome. The first case arises if $\boldsymbol{u}^*$ ('sunspot' in the counterfactual) is the same as $\boldsymbol{u}$ ('sunspot' in the 'data') and the second case if $\boldsymbol{u}^*$ is a new draw. In other words, by omitting $\boldsymbol{v}$ one rules out the possibility that the payoff–irrelevant unobservables contain *some but not all* information helpful for the counterfactual outcome over and above that which is contained in the payoff variables $\boldsymbol{e}, \boldsymbol{x}$. An explicit example of such a situation can be found at the end of section 2.3.2. While it is true that this limitation can be relaxed by allowing for imperfect dependence between $\boldsymbol{u}^*$ and $\boldsymbol{u}$, doing so would introduce a new random variable, also.[20]

2.3.2 Entry games: If $\imath_i(0,x) > \imath_i(1,x)$ for $i = 1,2$ in figure 1 then we have strategic substitutability, an 'entry' game. In the multiplicity region, the two pure strategy Nash equilibria are now $(1,0)$ and $(0,1)$. In contrast to the case discussed in section 2.3.1, in

[20] For instance, one could write $\boldsymbol{u}^* = \boldsymbol{u} + \upsilon \boldsymbol{v}$.

the entry game version of figure 1 each player prefers a different equilibrium and the two equilibria cannot be ranked.[21] Since there are again two pure strategy Nash equilibria, equilibrium selection is still an issue.

Consider the example depicted in figure 4. This example differs from figure 3 because

1 \ 2	0	1
0	0, 0	0, 1
1	1, 0	−1, −1

Figure 4: An entry game

the two Nash equilibria, i.e. $(1, 0)$ and $(0, 1)$, cannot be ranked: each equilibrium is neither payoff–dominant nor strictly risk–dominant. The same interpretations based on types of market heterogeneity apply here: $\boldsymbol{v}$ is long–run or player–originated heterogeneity such as the mutation rate or a commitment signal from a player and $\boldsymbol{u}$ corresponds to short–run stage idiosyncrasies. Indeed, in the context of Kandori, Mailath, and Rob (1993) $\boldsymbol{u}, \boldsymbol{v}$ play the same role as they did in section 2.3.1.

1 \ 2	0	1
0	0, 0	0, 2
1	1, 0	−1, −1

Figure 5: Another entry game

A second example concerns collusion. Suppose that in the game depicted in figure 5, nonparticipants' payoffs are increasing in the number of 'entrants,' and that it is illegal for players to coordinate. Thus, total surplus can be less if players coordinate. Suppose further that side payments are infeasible due to the risk of discovery. One can think of participants in a procurement withholding a bid to benefit another ring member. Bid rigging on milk contracts for schools is a well–known example; see Porter and Zona (1999).

The players could assign the right to bid randomly, resulting in outcome $(1, 0)$ with probability $\boldsymbol{p}$. If $\boldsymbol{v}$ denotes the bargaining weight of player 1 then $\boldsymbol{p}$ depends on both the conflict payoffs and on $\boldsymbol{v}$.[22]

For example, in figure 5 there are two pure strategy Nash equilibria plus a mixed strategy equilibrium (probabilities $1/3, 2/3$ and $1/2, 1/2$, respectively). Taking the mixed strategy

[21] Note that some of these additional features arise because playing 0 yields 0 regardless of what the opponent chooses to do. For instance, in more general games with strategic substitutability, i.e. ones that do not conform to figure 1, it can easily be the case that the players prefer different equilibria.

[22] In the standard Nash bargaining solution, the conflict payoffs are the payoffs players obtain when they fail to reach agreement. The bargaining weights represent the fraction of the surplus generated by the agreement that each player obtains. For instance, if absent agreement payoffs are (c_1, c_2) and with agreement total surplus is a, then for bargaining weights $v, 1-v$ for players 1 and 2 respectively, player 1 receives $c_1 + (a-c)v$ and player 2 receives $c_2 + (a-c)(1-v)$, where $c = c_1 + c_2$.

payoffs 0,0 to be the 'conflict payoffs,' we have $\boldsymbol{p} = 2\boldsymbol{v} \;/\; (\boldsymbol{v}+1)$.[23] In the normal form in figure 1 with strategic substitutability, we have in the multiplicity region that

$$\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}) = \frac{\{\imath_2(\boldsymbol{x}, 0) + \boldsymbol{e}_2\}\boldsymbol{v}}{\{\imath_1(\boldsymbol{x}, 0) + \boldsymbol{e}_1\}(1-\boldsymbol{v}) + \{\imath_2(\boldsymbol{x}, 0) + \boldsymbol{e}_2\}\boldsymbol{v}},$$

noting that $p(e, v, x)$ is (if $e \in S_m(x)$) increasing in v and between zero and one.

Like in section 2.3.1, in both of the above examples, the structural interpretation is lost and the range of counterfactuals that can be studied is diminished if $\boldsymbol{u}, \boldsymbol{v}$ are collapsed into a single payoffs–independent unobservable random variable. In the collusion example, for instance, it requires two random variables to consider a counterfactual in which the bargaining weights are held fixed but in which different outcomes can arise for the same values of the payoff variables.

2.3.3 Comments: In both the coordination game and entry game examples, if one were willing to make assumptions about the specific equilibrium selection mechanism then such information could be exploited and confidence bands on counterfactual outcome probabilities could be tightened. We have chosen not to go this route to maintain generality.

In view of the similarity between the prediction problems for the coordination and entry games, from here on we only discuss the entry game case, noting that our results equally apply to the coordination game example and indeed to multiplayer examples that possess features of both games.[24]

3. Counterfactuals

We now consider thought experiments in which we consider what will happen if the game is played again under various scenarios and with (potentially) different covariate values. We denote the ex post variables by $(\boldsymbol{e}^*, \boldsymbol{p}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{x}^*, \boldsymbol{y}^*)$ which, except where otherwise noted, will be an independent copy of $(\boldsymbol{e}, \boldsymbol{p}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$.[25] The exceptions are that in different scenarios different combinations of the input variables are assumed to stay unchanged, which we explain in more detail below.

For given e, u, v, recall from (7) that $y(e, u, v, x)$ is the value $\boldsymbol{y}$ would take if $\boldsymbol{e} = e, \boldsymbol{u} = u, \boldsymbol{v} = v, \boldsymbol{x} = x$. Thus, $\boldsymbol{p}^* = p(\boldsymbol{e}^*, \boldsymbol{v}^*, \boldsymbol{x}^*)$ and $\boldsymbol{y}^* = y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{x}^*)$. Using this notation, we can now consider various counterfactual outcomes, which differ depending on which combination of the conditions $\boldsymbol{e}^* = \boldsymbol{e}$, $\boldsymbol{u}^* = \boldsymbol{u}$, $\boldsymbol{v}^* = \boldsymbol{v}$, is applied: e.g. if we keep the unobserved payoff shifter unchanged but redraw all the other unobservables, then the

[23] Indeed, player 1's expected payoff is then $2v \;/\; (v+1)$ and player 2's is $\{1-2v/(v+1)\}2 = (2-2v) \;/\; (v+1)$, such that player 1's share of the pie is v.

[24] For instance, it is conceivable that $(1, 1, 0)$ is a Nash equilibrium in a three player game.

[25] So, $\boldsymbol{x}^*$ has the same support as $\boldsymbol{x}$. If the counterfactual of interest is not in the support of $\boldsymbol{x}$, then we would have to rely on extrapolation.

counterfactual of interest is $\boldsymbol{y}^* = y(\boldsymbol{e}, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{x}^*)$.[26] So there are up to eight different scenarios to consider, but we will focus on two cases, which we believe to be the most meaningful: keeping $\boldsymbol{v}$ unchanged and keeping both $\boldsymbol{e}$ and $\boldsymbol{v}$ unchanged. The rationale for emphasizing these two cases is that we would like to know what will happen in a similar market, or indeed in the same market under different circumstances. In other words, we think of these two cases as representative. In the interest of completeness, we provide results on other counterfactuals, i.e. fixing other combinations of $\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}$ in appendix B,[27] and provide a brief summary thereof in section 3.3.

For each of the counterfactuals, we wish to construct a prediction of $\boldsymbol{y}^*$. We define $q_c(y^* \mid x^*, x, y)$ to be the conditional probability that $\boldsymbol{y}^* = y^*$ given $\boldsymbol{x}^* = x^*, \boldsymbol{x} = x, \boldsymbol{y} = y$, where the subscript c indicates which of $\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*$ are fixed. For instance,

$$\begin{cases} q(y^* \mid x^*, x, y) = \mathbb{P}\big\{y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, x^*) = y^* \;\big|\; \boldsymbol{x}^* = x^*, \boldsymbol{x} = x, \boldsymbol{y} = y\big\}, \\ q_v(y^* \mid x^*, x, y) = \mathbb{P}\big\{y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}, x^*) = y^* \;\big|\; \boldsymbol{x}^* = x^*, \boldsymbol{x} = x, \boldsymbol{y} = y\big\}, \\ q_{ev}(y^* \mid x^*, x, y) = \mathbb{P}\big\{y(\boldsymbol{e}, \boldsymbol{u}^*, \boldsymbol{v}, x^*) = y^* \;\big|\; \boldsymbol{x}^* = x^*, \boldsymbol{x} = x, \boldsymbol{y} = y\big\}, \end{cases}$$

where q represents the case with no constraints and q_v, q_{ev} represent the cases where (some of) the unobservables are unchanged. We use both $\boldsymbol{x}$ and $\boldsymbol{y}$ to predict $\boldsymbol{y}^*$, because the values of both $\boldsymbol{x}$ and $\boldsymbol{y}$ contain information about $\boldsymbol{p}$.

The quantity $q(y^* \mid x^*, x, y)$ is identified because it is equal to

$$\tilde{\pi}_{y^*}(x^*) = \mathbb{P}(\boldsymbol{y}^* = y^* \mid \boldsymbol{x}^* = x^*) = \mathbb{P}(\boldsymbol{y} = y^* \mid \boldsymbol{x} = x^*).$$

The *regression prediction* $\tilde{\pi}_{y^*}(x^*)$ is simple, but it represents the case where none of the unobservables in the environment remains the same and is hence less interesting as a counterfactual exercise.

In sections 3.1 and 3.2 we discuss q_v and q_{ev} in greater detail. Define

$$\mu_m(x) = \mathbb{E}\big\{\boldsymbol{p} \;\big|\; \boldsymbol{e} \in S_m(x), \boldsymbol{x} = x\big\}. \tag{8}$$

Since $\pi_y(x)$ and $\pi_m(x)$ can be recovered from the payoff structure, $\mu_m(x)$ can be identified from

$$\tilde{\pi}_{10}(x) = \pi_{10}(x) + \pi_m(x)\mu_m(x),$$

provided that $\pi_m(x) > 0$: if $\pi_m(x) = 0$ then the value of $\mu_m(x)$ is immaterial. In fact, μ_m is the only identifiable feature of the conditional distribution of $\boldsymbol{p}$ given $\boldsymbol{e}, \boldsymbol{x}$, because the probability mass function $\tilde{\pi}_y(x)$ depends on $\rho(\cdot, \cdot, x)$ only through $\mu_m(x)$.

Throughout the remainder of the paper, we will frequently use the shorthand

$$\delta_y = \mathbb{1}\{y = (1,0)\} - \mathbb{1}\{y = (0,1)\}. \tag{9}$$

[26] Therefore, the 'condition' $\boldsymbol{e}^* = \boldsymbol{e}$ does not mean that we are 'conditioning.' It means that we keep $\boldsymbol{e}$ *unchanged*, i.e. use the *same* random variable, in defining the counterfactual outcome. These two concepts are generally not the same. For instance, suppose that $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ are random variables with a standard exponential distribution. If $\boldsymbol{\xi}_2$ is the same random variable as $\boldsymbol{\xi}_1$ (which is the case we are considering) then $\mathbb{E}\boldsymbol{\xi}_2$ is by definition equal to $\mathbb{E}\boldsymbol{\xi}_1 = 1$. But if e.g. $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ are independent random variables then $\mathbb{E}(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_2 = \boldsymbol{\xi}_1) = 1 \,/\, 2$. So there is a difference between the random variables being the same and conditioning on two different random variables having the same value.

[27] There are eight potential counterfactuals of interest: the regression prediction q; the leading examples q_v and q_{ev}; the predictions considered in appendix B q_e, q_{uv}, and q_{euv}; and predictions that are omitted because of their similarity to ones that are discussed in the paper, namely q_u and q_{eu}.

3.1 Case 1: $\boldsymbol{v}^* = \boldsymbol{v}$: We now study identification of $q_v(y^* \mid x^*, x, y)$. Recall that this corresponds to the case in which the probability $\boldsymbol{p}$ of selecting a particular equilibrium in the multiplicity region only varies because of changes in the values of the payoff variables.

Let

$$\rho(x, x^*) = \text{Cov}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}^*, \boldsymbol{v}, x^*) \,\big|\, \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\big\}$$
$$= \text{Cov}\{\hbar(\boldsymbol{v}, x), \hbar(\boldsymbol{v}, x^*)\}, \quad (10)$$

where

$$\hbar(v, x) = \mathbb{E}\big\{\boldsymbol{p} \,\big|\, \boldsymbol{e} \in S_m(x), \boldsymbol{v} = v, \boldsymbol{x} = x\big\} = \mathbb{E}\big\{p(\boldsymbol{e}, v, x) \,\big|\, \boldsymbol{e} \in S_m(x)\big\}. \quad (11)$$

The function ρ describes the predictability of $\boldsymbol{p}^*$ by $\boldsymbol{p}$: if $\rho(x, x^*)$ (which is nonnegative by construction) is large then we can learn a lot about the value of $\boldsymbol{p}^*$ from the value of $\boldsymbol{p}$. Although we do not observe $\boldsymbol{p}$, the values of $\boldsymbol{x}, \boldsymbol{y}$ (which are observed) provide us with information about $\boldsymbol{p}$. It is intuitive that the only relevance of $\boldsymbol{e}$ in this context is whether or not it belongs to the multiplicity region: since here $\boldsymbol{e}^*$ is an independent copy of $\boldsymbol{e}$ and $\boldsymbol{e}$ is independent of everything else, we cannot learn anything about the effect of $\boldsymbol{e}$ on the value of $\boldsymbol{p}$ other than whether or not $\boldsymbol{e} \in S_m(\boldsymbol{x})$. Hence, $\boldsymbol{e}$ is averaged out. Theorem 1 formalizes this intuition.

Theorem 1. Let

$$\tilde{\rho}_{yy^*}(x, x^*) = \delta_y \delta_{y^*} \frac{\pi_m(x)\pi_m(x^*)}{\tilde{\pi}_y(x)} \rho(x, x^*),$$

where δ_y was defined in (9). Then, $q_v(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \tilde{\rho}_{yy^*}(x, x^*)$. □

The results of theorem 1 are intuitive. The correction term $\tilde{\rho}_{yy^*}(x, x^*)$ reflects what we can learn about $\boldsymbol{v}$ from observing $\boldsymbol{x}, \boldsymbol{y}$. Indeed, if $y = (0, 0)$ or $y = (1, 1)$ then we do not learn anything about the value of $\boldsymbol{v}$ by observing $\boldsymbol{x}, \boldsymbol{y}$ since the value of $\boldsymbol{v}$ is only of relevance in the multiplicity region: hence the correction term is equal to zero. Likewise, if $y^* = (0, 0)$ or $y^* = (1, 1)$ then any knowledge obtained about $\boldsymbol{v}$ by observing $\boldsymbol{x}, \boldsymbol{y}$ is useless. In the remaining cases, if the new multiplicity region is large (and hence $\pi_m(x^*)$ is large) then the correction term is large, also. If the probability that we are in the multiplicity region is large relative to the probability that $\boldsymbol{y} = y$, then again the correction term is large. Finally, if $\rho(\boldsymbol{e}, \boldsymbol{v}, x)$ is highly correlated with $\rho(\boldsymbol{e}^*, \boldsymbol{v}, x^*)$ conditional on $\boldsymbol{e}, \boldsymbol{e}^*$ belonging to their respective multiplicity regions, then again the correction term is large.

If the policy maker is only interested in the number of 'entrants,' not their identity, as in e.g. Bresnahan and Reiss (1991a), then $q_v\{(1, 0) \mid x^*, x, y\} + q_v\{(0, 1) \mid x^*, x, y\}$ is the object of interest. For the present counterfactual, the prediction of the number of entrants is point–identified since theorem 1 implies that

$$q_v\{(1, 0) \mid x^*, x, y\} + q_v\{(0, 1) \mid x^*, x, y\} = \tilde{\pi}_{10}(x^*) + \tilde{\pi}_{01}(x^*). \quad (12)$$

So, if one is concerned whether or not there is a monopoly then the object of interest is point–identified. However, if the identity of the monopolist is of interest then theorem 1 implies that there is a (partially identified) correction term that needs to be dealt with. In

the counterfactual experiment of section 3.2 there will be a correction term even if one is only concerned with the number of entrants.

For the remainder of section 3.1, we will focus on the individual counterfactual $q_v(y^* \mid x^*, x, y)$ rather than the sum (12). The size of the correction term in theorem 1 depends on a number of factors, but it can be bounded. The bound depends on $B(x, x^*)$ given in theorem 2.

Theorem 2. For all x, x^*, $0 \leq \rho(x, x^*) \leq B(x, x^*)$, where $B(x, x^*) = \min\{\mu_m(x), \mu_m(x^*)\} - \mu_m(x)\mu_m(x^*)$. Both bounds are sharp. □

An immediate implication of theorem 2 is that the regression prediction coincides with the lower bound of the sharp identified interval for q_v, which is not true for q_{ev}, as will be shown in section 3.2. Therefore, the maximum difference between the correct counterfactual object and the regression prediction coincides with the length of the sharp identified set of $q_v\{(1,0) \mid x^*, x, (1,0)\}$, which is equal to

$$\tilde{\rho}_{10,10}(x, x^*) = \frac{\pi_m(x)\pi_m(x^*)}{\tilde{\pi}_{10}(x)} B(x, x^*).$$

Note that $\tilde{\rho}_{10,10}(x, x^*)$ can be made arbitrarily close to one by letting $\pi_m(x), \pi_m(x^*) \to 1$ and $\mu_m(x^*) = \mu_m(x) \to 0$. Indeed, $\tilde{\rho}_{10,10}(x, x^*)$ approaches $\min\{1, \mu_m(x^*) \,/\, \mu_m(x)\} - \mu_m(x^*)$ as $\pi_m(x), \pi_m(x^*) \to 1$. More generally, $\tilde{\rho}_{10,10}(x, x^*)$ varies with the ratio $\pi_{10}(x) \,/\, \pi_m(x)$ as depicted in figure 6. Much the same applies to other combinations of $y, y^* \in \{(1,0), (0,1)\}$.

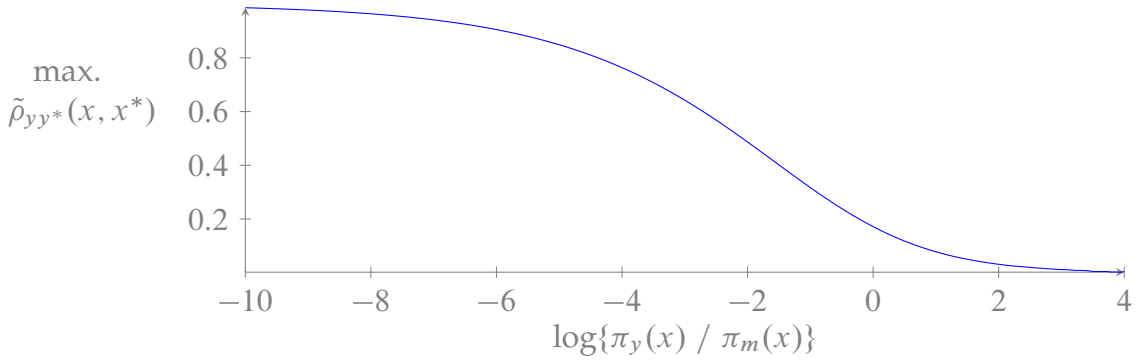

Figure 6: Maximum possible value of $\tilde{\rho}_{yy^*}(x, x^*)$ as a function of $\log\{\pi_y(x) \,/\, \pi_m(x)\}$ if $\mu_m(x^*) = \mu_m(x)$ and $y^* = y = (1,0)$

3.2 Case 2: $e^* = e$ and $v^* = v$: We now turn to the case where both e and v are fixed, i.e. the probability p of selecting a particular equilibrium under multiplicity varies only because of changes in the values of the observable payoff variables. In this case, the value of e is important for prediction.[28] For instance, if $x^* = x$ then the S–regions are unchanged and $p^* = p$.[29] If $x^* \neq x$ then the S–regions will be different and p^* can be different from

[28] The value of e would also be important for prediction if v were not fixed. This case is simpler and is discussed in theorem 10 in appendix B.

[29] By 'S–region' we mean one of of the sets $S_\cdot(\cdot)$.

$\boldsymbol{p}$, but the values of $\boldsymbol{x}, \boldsymbol{y}$ still contain more information to aid in the prediction of $\boldsymbol{y}^*$ than in the case discussed in section 3.1.

In section 3.1 $\boldsymbol{e}^*$ was an independent copy of $\boldsymbol{e}$ and $\pi_m(x), \pi_m(x^*)$ were relevant for prediction. Now, $\boldsymbol{e}^*$ is the same random variable as $\boldsymbol{e}$ and hence the probability that $\boldsymbol{e}$ belongs to the intersection of $S_m(x)$ and $S_m(x^*)$ becomes relevant. We now introduce further notation to facilitate our analysis.

Let $S_{my}(x, x^*) = S_m(x) \cap S_y(x^*)$, $\pi_{my}(x, x^*) = \mathbb{P}\{\boldsymbol{e} \in S_{my}(x, x^*)\}$, and let S_{mm}, S_{yy^*}, π_{mm}, π_{yy^*}, etcetera, be analogously defined.[30] We further define $S_r(x) = S_r^y(x) = \{S_y(x) \cup S_m(x)\}^c$ and $S_{r^*}(x) = S_r^{y^*}(x) = \{S_{y^*}(x) \cup S_m(x)\}^c$. In order to determine the bounds on q_{ev}, it matters where the S–regions intersect. Let

$$\rho(x, x^*) = \mathbb{E}\Big[\mathrm{Cov}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}, x^*) \;\big|\; \boldsymbol{e}\big\} \;\Big|\; \boldsymbol{e} \in S_{mm}(x, x^*)\Big], \tag{13}$$

which is similar to (10) but imposes that $\boldsymbol{e}$ and $\boldsymbol{e}^*$ are the same random variable in both p functions. We use the same symbol ρ in both (10) and (13) for several reasons even though the expressions differ. First, ρ serves the same role of correcting for the information contained in $\boldsymbol{x}, \boldsymbol{y}$ about the value of $\boldsymbol{p}^*$: it is always clear which one is meant from the context. Further, if $p(e, v, x)$ is flat in e, then imposing the constraint of $\boldsymbol{e}^* = \boldsymbol{e}$ makes the two expressions coincide. As will be shown in the following section, the maximum entropy solution for p is indeed flat in e.

But we need an additional correction term to address the fact that $\boldsymbol{x}, \boldsymbol{y}$ contain information about the $S_\cdot(x^*)$–region to which $\boldsymbol{e}$ belongs. This correction term can once again be expressed in terms of a covariance of constructed random variables. Theorem 3 contains the result, for which we need to introduce notation. Let $c_y(e, x) = \mathbb{1}\{e \in S_y(x)\}$, $c_m(x) = \mathbb{1}\{e \in S_m(x)\}$, and

$$b_y(e, v, x) = \begin{cases} p(e, v, x), & y = (1, 0), \\ 1 - p(e, v, x), & y = (0, 1), \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 3. Let $a_y(e, x) = c_y(e, x) + c_m(e, x)\mathbb{E}b_y(e, \boldsymbol{v}, x)$ and

$$\begin{cases} \alpha_{yy^*}(x, x^*) = \dfrac{\mathrm{Cov}\{a_y(\boldsymbol{e}, x), a_{y^*}(\boldsymbol{e}, x^*)\}}{\tilde{\pi}_y(x)}, \\ \tilde{\rho}_{yy^*}(x, x^*) = \delta_y \delta_{y^*} \dfrac{\pi_{mm}(x, x^*)}{\tilde{\pi}_y(x)} \rho(x, x^*). \end{cases}$$

Then, in view of (6), $q_{ev}(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \alpha_{yy^*}(x, x^*) + \tilde{\rho}_{yy^*}(x, x^*)$. □

The intuition for the $\tilde{\rho}$–term in theorem 3 is similar to that in theorem 1: it reflects what we can learn about $\boldsymbol{v}$ from observing $\boldsymbol{y}, \boldsymbol{x}$. Likewise, the α term captures the information for $\boldsymbol{e}$ contained in $\boldsymbol{y}, \boldsymbol{x}$. For example, consider first the case that $y = y^* = (0, 0)$. Then the

[30] Please note that the orders of the input arguments and the subindices are relevant here.

α_{yy^*} term simply increases the probability that $\boldsymbol{y}^* = (0,0)$ because we know that $\boldsymbol{y} = (0,0)$ implies that $\boldsymbol{e}_1, \boldsymbol{e}_2$ are both small. In each of the other cases the intuition is merely a more complicated version of this argument.

Define

$$\gamma_y(x) = \mathbb{E}\big\{\mathcal{b}_y(\boldsymbol{e}, \boldsymbol{v}, x) \,\big|\, \boldsymbol{e} \in S_m(x)\big\} \quad \text{and} \quad \phi_y(x) = \pi_m(x)\gamma_y(x).$$

So, $\gamma_y(x)$ is $\mu_m(x)$, $1 - \mu_m(x)$, or 0, depending on the value of y.

Now, suppose again that one is interested in the number of entrants rather than their identity. Then, theorem 3 implies that

$$\begin{aligned} q_{ev}\big\{(1,0) \,\big|\, x^*, x, y\big\} + q_{ev}\big\{(0,1) \,\big|\, x^*, x, y\big\} \\ = \tilde{\pi}_{10}(x^*) + \tilde{\pi}_{01}(x^*) + \frac{\mathrm{Cov}\big\{a_y(\boldsymbol{e}, x), a_{10}(\boldsymbol{e}, x^*) + a_{01}(\boldsymbol{e}, x^*)\big\}}{\tilde{\pi}_y(x)}. \end{aligned} \quad (14)$$

The covariance on the right hand side in (14) is equal to

$$C_y(x, x^*) + \sum_{\tilde{y} \in \{(1,0),(0,1),m\}} \big[\pi_{y\tilde{y}}(x, x^*) - \pi_{\tilde{y}}(x^*)\{\pi_y(x) + \phi_y(x)\}\big],$$

where

$$C_y(x, x^*) = \sum_{\tilde{y} \in \{(1,0),(0,1),m\}} \pi_{m\tilde{y}}(x, x^*)\mathbb{E}\{\mathcal{b}_y(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_{m\tilde{y}}\}.$$

Therefore, unlike in the case discussed in section 3.1, the counterfactual probability of monopoly is not necessarily identified because $C_y(x, x^*)$ can depend on $p(\boldsymbol{e}, \boldsymbol{v}, x)$. Note however that $p(\boldsymbol{e}, \boldsymbol{v}, x^*)$ is irrelevant for the counterfactual probability of monopoly. The following theorem provides sharp bounds of $C_y(x, x^*)$.

Theorem 4. For all x, x^*, we have

$$\begin{aligned} \max\big\{0, \phi_y(x) - \pi_{m,00}(x, x^*) - \pi_{m,11}(x, x^*)\big\} \le C_y(x, x^*) \\ \le \min\big\{\phi_y(x), \pi_m(x) - \pi_{m,00}(x, x^*) - \pi_{m,11}(x, x^*)\big\}. \end{aligned}$$

The bounds are sharp. □

If y is either $(0,0)$ or $(1,1)$ then point identification obtains: e.g. the probability of monopoly in the counterfactual conditional on no monopoly ex ante is point–identified. Further, if the S–regions do not change (e.g. if $x = x^*$), then we have point identification, also.

However, the situation is less clear if there is a monopoly ex ante. It is then not generally possible to determine the counterfactual probability of monopoly if x is changed to x^* (e.g. by a policy maker). For instance, suppose that $y = (1,0)$ and that $\pi_m(x) = 1$. Suppose further that the policy is such that $\pi_{m,00}(x, x^*) + \pi_{m,11}(x, x^*) = 1 - \upsilon < 1$. Then, the length of the identified interval of $q_{ev}\big\{(1,0) \,\big|\, x^*, x, (1,0)\big\} + q_{ev}\big\{(0,1) \,\big|\, x^*, x, (1,0)\big\}$ equals

$$\min\big\{1, \upsilon \,/\, \mu_m(x)\big\} - \max\big\{0, 1 - (1 - \upsilon) \,/\, \mu_m(x)\big\},$$

which converges to 1 as $\mu_m(x)$ approaches 0.

We now turn to the problem of determining sharp identified bounds for $q_{ev}\left(y^* \,\middle|\, x^*, x, y\right)$ for the balance of this section. Define

$$L_{yy^*}(x,x^*) = \text{max. of } \begin{cases} 0, \\ \phi_y(x) - \pi_{mr^*}(x,x^*) - \pi_{mm}(x,x^*), \\ \phi_{y^*}(x^*) - \pi_{rm}(x,x^*) - \pi_{mm}(x,x^*), \\ \phi_y(x) + \phi_{y^*}(x^*) - \pi_{mr^*}(x,x^*) - \pi_{rm}(x,x^*) - \pi_{mm}(x,x^*), \end{cases}$$

and

$$U_{yy^*}(x,x^*) = \text{min. of } \begin{cases} \phi_y(x) + \phi_{y^*}(x^*), \\ \phi_y(x) + \pi_{ym}(x,x^*), \\ \phi_{y^*}(x^*) + \pi_{my^*}(x,x^*), \\ \pi_{mm}(x,x^*) + \pi_{ym}(x,x^*) + \pi_{my^*}(x,x^*), \end{cases} \tag{15}$$

which will enter into the lower and upper bound formulas for q_{ev}. $L_{yy^*}(x,x^*)$ and $U_{yy^*}(x,x^*)$ depend on identified objects only. Indeed, $L_{yy^*}(x,x^*)$ and $U_{yy^*}(x,x^*)$ are determined by a combination of π–values and the values of $\phi_y(x)$ and $\phi_{y^*}(x^*)$, as is illustrated in figure 7. Depending on the values of $\phi_y(x)$ and $\phi_y(x^*)$, different bounds are binding.

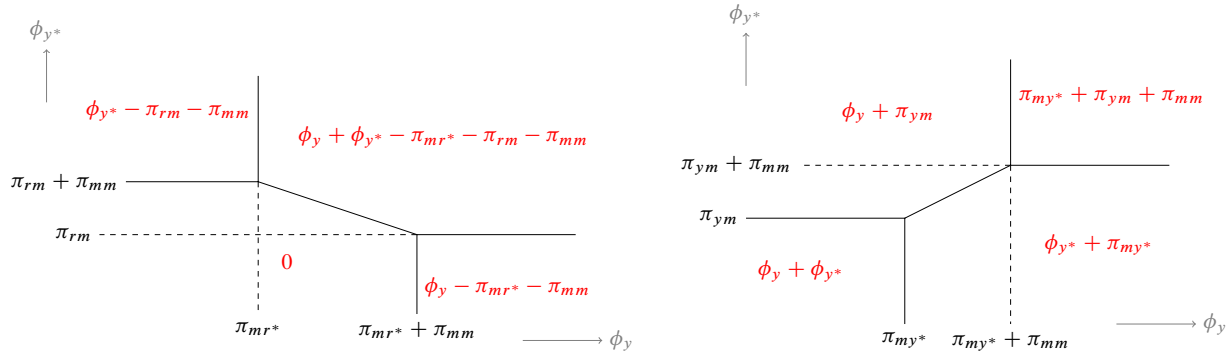

Figure 7: Regions of $\left(\phi_y(x), \phi_{y^*}(x^*)\right)$ and the corresponding values of $L_{yy^*}(x,x^*), U_{yy^*}(x,x^*)$.

We are now ready to describe the sharp identified bounds of $q_{ev}(y^* \mid x^*, x, y)$.

Theorem 5. If $x \neq x^*$, then

$$\pi_{yy^*}(x,x^*) + L_{yy^*}(x,x^*) \leq q_{ev}(y^* \mid x^*, x, y)\tilde{\pi}_y(x) \leq \pi_{yy^*}(x,x^*) + U_{yy^*}(x,x^*).$$

If $x = x^*$, then

$$q_{ev}(y^* \mid x, x^*, y) \in \begin{cases} \{\mathbb{1}(y = y^*)\}, & \text{if } \delta_y \delta_{y^*} = 0, \\ \left[\{\pi_y(x) + \pi_m(x)\gamma_y^2(x)\} \;/\; \tilde{\pi}_y(x), 1\right], & \text{if } \delta_y \delta_{y^*} = 1, \\ \left[0, \{\pi_m(x)\gamma_y(x)\gamma_{y^*}(x)\} \;/\; \tilde{\pi}_y(y)\right], & \text{if } \delta_y \delta_{y^*} = -1. \end{cases}$$

The bounds are sharp in both cases. □

The formulas for the bounds in theorem 5 are different depending on whether or not $x = x^*$ and the bounds for $x = x^*$ do not obtain as a limit if $x \to x^*$: if $\delta_y \delta_{y^*} = 1$ (-1) then the lower (upper) bound is discontinuous at $x = x^*$. This fact can be explained as follows. If $x = x^*$ then $\boldsymbol{p} = \boldsymbol{p}^*$. If $x \neq x^*$ then $\boldsymbol{p}$ and $\boldsymbol{p}^*$ are the outputs of *different* functions of the *same* random variables $\boldsymbol{e}, \boldsymbol{v}$: $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, x)$ and $\boldsymbol{p}^* = p(\boldsymbol{e}, \boldsymbol{v}, x^*)$. Even if $p(e, v, x)$ were continuous in x (which we do not assume) then $\left|p(e, v, x^*) - p(e, v, x)\right| / \|x - x^*\|$ can still be arbitrarily large. Since p need not be monotonic in e, the covariance between $\boldsymbol{p}$ and $\boldsymbol{p}^*$ can be negative, unlike the variance of $\boldsymbol{p}$.

If the function p is flat in e, then $\boldsymbol{p}$ and $\boldsymbol{p}^*$ cannot be negatively correlated, and consequently the formulas for the bounds in theorem 5 for the cases $x = x^*$ and $x \neq x^*$ then coincide. However, even in that case, $q_{ev}(y^* \mid x^*, x, y)$ and $q_v(y^* \mid x^*, x, y)$ generally have different values, because for $q_{ev}(y^* \mid x^*, x, y)$ it matters where the S–regions intersect.

Please note that the bounds in theorem 5 do not generally contain the regression prediction $\tilde{\pi}_{y^*}(x^*)$: see section 7.3 for examples. Therefore, the regression prediction is a poor choice if the object of interest is q_{ev}.

3.3 Other cases: As announced, appendix B contains rigorous results for the remaining counterfactual predictions, namely $q_u, q_{uv}, q_e, q_{eu}, q_{euv}$. The infeasible predictions themselves generally differ from those that we derived in sections 3.1 and 3.2. Indeed, if $y = y^* = (1, 0)$ then

$$\tilde{\pi}_y(x) q_\cdot(y^* \mid x^*, x, y) = \pi_y(x)\pi_{y^*}(x^*) + \pi_y(x)\pi_m(x^*)\mu_m(x^*) + \pi_{y^*}(x^*)\pi_m(x)\mu_m(x) +$$
$$\pi_m(x)\pi_m(x^*) \times \begin{cases} \mu_m(x)\mu_m(x^*), & q_\cdot = q, \\ \mathbb{E}\{\hbar(\boldsymbol{v}, x)\hbar(\boldsymbol{v}, x^*)\}, & q_\cdot = q_v, \\ \mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}^*, \boldsymbol{v}^*, x^*)\} \,\big|\, \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\big], & q_\cdot = q_u, \\ \mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}^*, \boldsymbol{v}, x^*)\} \,\big|\, \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\big], & q_\cdot = q_{uv}, \end{cases}$$

and

$$\tilde{\pi}_y(x) q_\cdot(y^* \mid x^*, x, y) = \pi_{yy^*}(x, x^*) + \pi_{ym}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}^*, x^*) \mid \boldsymbol{e} \in S_{ym}(x, x^*)\} +$$
$$\pi_{my^*}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_{my^*}(x, x^*)\} +$$
$$\pi_{mm}(x, x^*) \times \begin{cases} \mathbb{E}\{\bar{p}(\boldsymbol{e}, x)\bar{p}(\boldsymbol{e}, x^*) \mid \boldsymbol{e} \in S_{mm}(x, x^*)\}, & q_\cdot = q_e, \\ \mathbb{E}\big[p(\boldsymbol{e}, \boldsymbol{v}, x)p(\boldsymbol{e}, \boldsymbol{v}, x^*) \,\big|\, \boldsymbol{e} \in S_{mm^*}(x, x^*)\big], & q_\cdot = q_{ev}, \\ \mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}^*, x^*)\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big], & q_\cdot = q_{eu}, \\ \mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}, x^*)\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big], & q_\cdot = q_{euv}, \end{cases}$$

where $\bar{p}(e, x) = \mathbb{E}p(e, \boldsymbol{v}, x)$. The reason for the dichotomy between the cases in which $\boldsymbol{e}$ is or is not fixed is that $\boldsymbol{e}$ is vector–valued, that p is not necessarily monotonic in e, and that the multiplicity regions live in $\boldsymbol{e}$–space.

One notable conclusion from the above two displayed equations is that the prediction

probabilities can be ordered, e.g.

$$\begin{cases} q(y^* \mid x^*, x, y) & \leq & q_v(y^* \mid x^*, x, y) & \leq & q_{uv}(y^* \mid x^*, x, y), \\ q_e(y^* \mid x^*, x, y) & \leq & q_{ev}(y^* \mid x^*, x, y) & \leq & q_{euv}(y^* \mid x^*, x, y), \end{cases} \tag{16}$$

which is intuitive. However, if $x \neq x^*$, as is shown in theorems 10 to 14 in appendix B, then the bounds on q_u, q_{uv} coincide with those on q_v and the bounds on q_e, q_{eu}, q_{euv} coincide with those on q_{ev}.[31] Since the bounds coincide, so do the midpoint predictions. Thus, despite the fact that the q–functions can be ordered as indicated in (16), the midpoint predictions are the same in each case.[32] The maximum entropy solution proposed in section 4 does not have this unfortunate feature and neither does the minmax solution with entropy constraint in section 5 for most constraint values.

4. Maximum entropy

The principle of maximum entropy is the notion that among the probability distributions that satisfy all testable restrictions available the probability distribution that best represents the current state of knowledge is the one that maximizes the entropy (Jaynes, 1957a; Jaynes, 1957b). The entropy measures the amount of uncertainty that a probability distribution represents.[33]

In our case, this entails finding the joint density of $\boldsymbol{e}, \boldsymbol{p}, \boldsymbol{x}$ which maximizes the entropy: recall from (6) that p is the quantile function corresponding to the conditional density f of $\boldsymbol{p}$ given $\boldsymbol{e}, \boldsymbol{x}$. Our notation suggests that $\boldsymbol{p}$ is continuously distributed, but discrete distributions obtain as limit cases: we do *not* assume any regularity on f such as smoothness or boundedness. For ease of exposition, we will further take $\boldsymbol{x}$ to be continuously distributed but the nature of the distribution of $\boldsymbol{x}$ is immaterial.

We continue using the same environment as before, i.e. we take the distributions of $\boldsymbol{e}, \boldsymbol{x}$ as given and take $\boldsymbol{e}, \boldsymbol{x}$ to be independent. Further, recall that $\mu_m(x)$, defined in (8), is identified for all x, and hence it imposes another constraint on f. Thus, using all information available, we consider maximizing the *conditional entropy*[34]

$$\begin{aligned} f^* = \operatorname*{argmax}_f & - \iiint_0^1 f(p \mid e, x) \log f(p \mid e, x) \, \mathrm{d}p \, f_e(e) f_x(x) \, \mathrm{d}e \, \mathrm{d}x \\ \text{subject to} & \begin{cases} \forall e, x : \int_0^1 f(p \mid e, x) \, \mathrm{d}p = 1, \\ \forall x : \int_{S_m(x)} \int_0^1 p f(p \mid e, x) \, \mathrm{d}p \, f_e(e) \, \mathrm{d}e = \mu_m(x) \pi_m(x). \end{cases} \end{aligned} \tag{17}$$

[31] We focus on the case $x \neq x^*$ because $q_{euv}(y^* \mid x, x, y) = \mathbb{1}(y^* = y)$.

[32] The midpoint predictions are different if one compares fixing $\boldsymbol{e}$ and not fixing $\boldsymbol{e}$, however.

[33] Admittedly, one could define alternative such measures; our use of entropy is motivated by it being the dominant choice in information theory.

[34] The problem can be equivalently formulated by the unconditional entropy of the joint density of $(\boldsymbol{e}, \boldsymbol{p}, \boldsymbol{x})$.

The optimization problem in (17) can be solved using standard constrained optimization techniques (see e.g. Cover and Thomas, 2012). The result is provided in theorem 6 below. Let

$$I(\lambda) = \int_0^1 \exp(p\lambda)\,\mathrm{d}p, \qquad \mathscr{L}(\lambda) = \log I(\lambda). \tag{18}$$

Theorem 6. The solution to (17) is given by

$$f^*(p \mid e, x) = \begin{cases} \mathbb{1}(0 \leq p \leq 1), & e \notin S_m(x), \\ A\{p, \lambda_m(x)\}\mathbb{1}(0 \leq p \leq 1), & e \in S_m(x), \end{cases} \tag{19}$$

where

$$\lambda_m(x) = \operatorname*{argmin}_{\lambda}\{\mathscr{L}(\lambda) - \mu_m(x)\lambda\}, \qquad A(p, \lambda) = \exp(p\lambda) \;/\; I(\lambda). \tag{20}$$

□

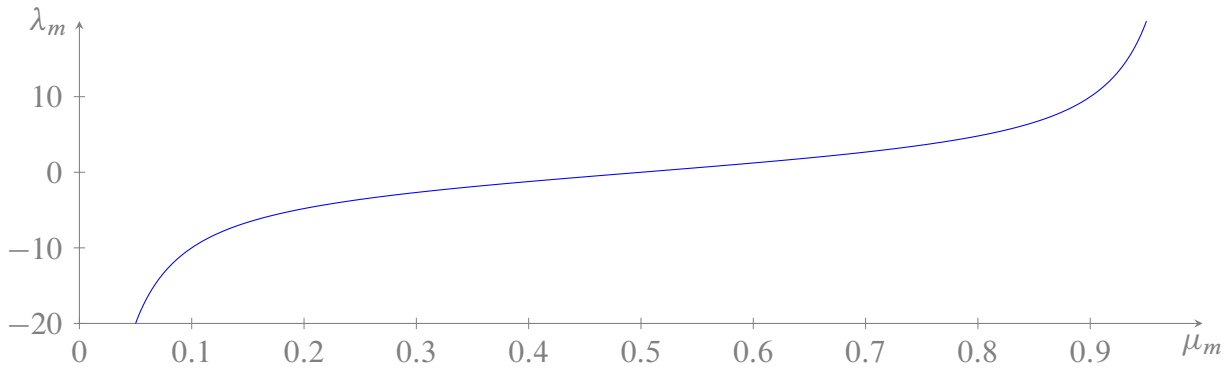

Figure 8: λ_m as a function of μ_m.

The relationship between λ_m and μ_m is depicted in figure 8. We can see from the first order condition of the minimization problem in (20) that the solution λ_m satisfies

$$\mu_m = \mathscr{L}'(\lambda_m) = \begin{cases} \dfrac{1}{2}, & \lambda_m = 0, \\ \dfrac{1}{1 - \exp(-\lambda_m)} - \dfrac{1}{\lambda_m}, & \lambda_m \neq 0, \end{cases}$$

where it should be noted that $\mathscr{L}'$ is positive and well–behaved at zero.[35]

Theorem 6 implies that $f^*(p \mid e, x)$ is different depending on whether or not $e \in S_m(x)$, but does not otherwise depend on the value of e. For $e \notin S_m(x)$, we have no information about $f^*(p \mid e, x)$ and hence maximum entropy produces a uniform distribution. For $e \in S_m(x)$, we only have information about the value of $\mu_m(x)$. If $\mu_m(x) = 0.5$ then there is nothing to suggest that the conditional distribution is not uniform. Otherwise, the density must be adjusted to accommodate the value of $\mu_m(x)$, as depicted in figure 9. In all cases,

[35] For instance, $\mathscr{L}'$ is continuous and differentiable at zero.

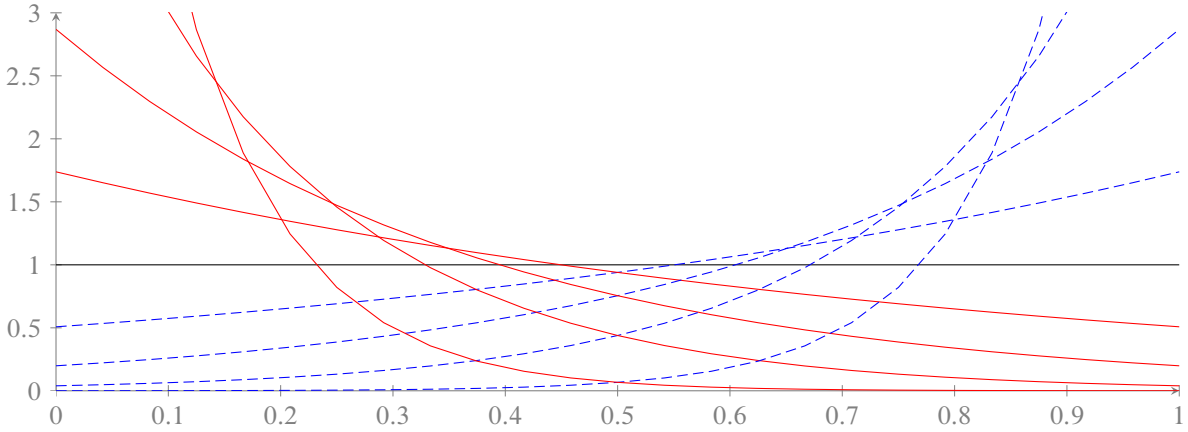

Figure 9: Conditional densities $f^*(p \mid e, x)$ for $e \in S_m(x)$ for values of $\mu_m = 0.1, \ldots, 0.9$, with blue dashed curves corresponding to $\mu_m > 0.5$ and red solid curves to $\mu_m < 0.5$.

the maximum entropy solution picks f^* such that the conditional density of $\boldsymbol{p}$ given $\boldsymbol{e}, \boldsymbol{x}$ is closest to a uniform as measured by the entropy. The conditional density function becomes steeper as $|\mu_m(x) - 0.5|$ increases.

Recall that p is the quantile function corresponding to $f_{\boldsymbol{p}|\boldsymbol{e},\boldsymbol{x}}$.

Theorem 7. The function p corresponding to $f = f^*$ is the function p^* given by

$$p^*(e, v, x) = \begin{cases} v, & e \notin S_m(x), \\ \hbar^*(v, x), & e \in S_m(x), \end{cases}$$

where the function $\hbar^*$ corresponds to $\hbar$ defined in (11) and is given by

$$\hbar^*(v, x) = \begin{cases} \dfrac{\log\big(1 + v[\exp\{\lambda_m(x)\} - 1]\big)}{\lambda_m(x)}, & \lambda_m(x) \neq 0, \\ v, & \lambda_m(x) = 0. \end{cases}$$ □

It is straightforward to calculate the correction terms in theorems 1 and 3 that are implied by p^* (and $\hbar^*$). Indeed, since p^* is flat in e when $e \in S_m(x)$, the maximum entropy correction terms corresponding to $\tilde{\rho}_{yy^*}(x, x^*)$ in theorems 1 and 3 coincide. We provide the general formulas in appendix C.

The function p^* is not generally continuous in e or x since the behavior of p^* is different inside and outside of $S_m(x)$. However, for e in the interior of $S_m(x)$, $p^*(e, v, x)$ is continuous in x and flat in e.

To see how the maximum entropy predictions differ across counterfactual experiments, we now provide the maximum entropy analogs to the counterfactual prediction formulas presented in section 3.3. If again $y = y^* = (1, 0)$ and $x \neq x^*$ then

$$\tilde{\pi}_y(x) q.(y^* \mid x^*, x, y) = \pi_y(x)\pi_{y^*}(x^*) + \pi_y(x)\pi_m(x^*)\mu_m(x^*) + \pi_m(x)\pi_{y^*}(x^*)\mu_m(x) +$$

$$\pi_m(x)\pi_m(x^*) \times \begin{cases} \mu_m(x)\mu_m(x^*), & q_\cdot = q, \\ \mathbb{E}\{\hbar^*(\boldsymbol{v}, x)\hbar^*(\boldsymbol{v}, x^*)\}, & q_\cdot = q_v, \\ \mathbb{E}\big[\min\{\hbar^*(\boldsymbol{v}, x), \hbar^*(\boldsymbol{v}^*, x^*)\}\big], & q_\cdot = q_u, \\ \min\{\mu_m(x), \mu_m(x^*)\}, & q_\cdot = q_{uv}. \end{cases}$$

and

$$\tilde{\pi}_y(x)q_\cdot(y^* \mid x^*, x, y) = \pi_{yy^*}(x, x^*) + \pi_{ym}(x, x^*)\mu_m(x^*) + \pi_{my^*}(x, x^*)\mu_m(x)$$
$$+ \pi_{mm}(x, x^*) \times \begin{cases} \mu_m(x)\mu_m(x^*), & q_\cdot = q_e, \\ \mathbb{E}\{\hbar^*(\boldsymbol{v}, x)\hbar^*(\boldsymbol{v}, x^*)\}, & q_\cdot = q_{ev}, \\ \mathbb{E}\big[\min\{\hbar^*(\boldsymbol{v}, x), \hbar^*(\boldsymbol{v}^*, x^*)\}\big], & q_\cdot = q_{eu}, \\ \min\{\mu_m(x), \mu_m(x^*)\}, & q_\cdot = q_{euv}. \end{cases}$$

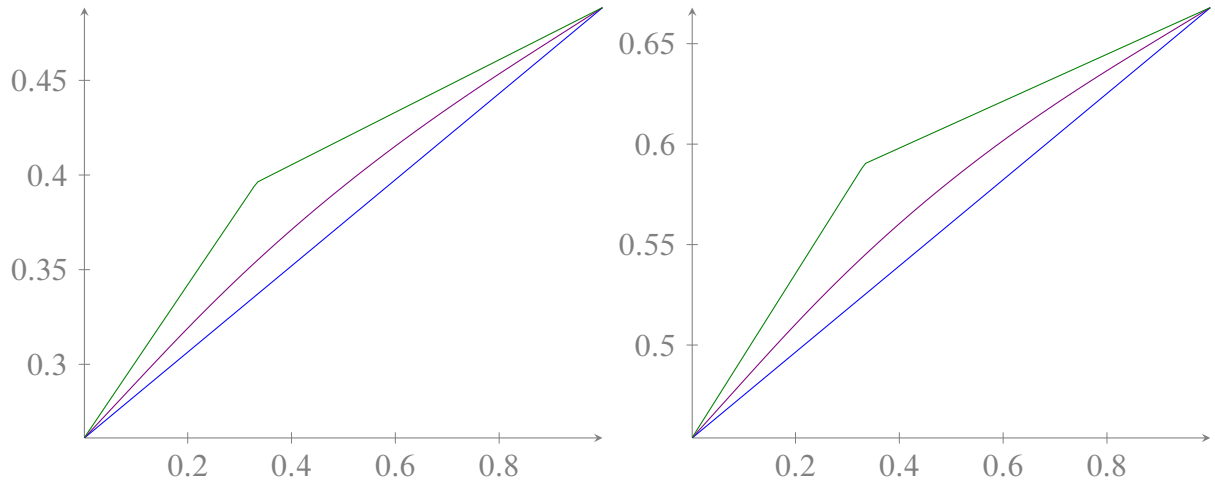

Figure 10: Maximum entropy solutions for $q_\cdot, q_v, q_{uv}$ (left graph, bottom to top) and q_e, q_{ev}, q_{euv} (right graph, bottom to top) as a function of $\mu_m(x^*)$ if $y = y^* = (1, 0)$, $\mu_m(x) = 1 \,/\, 3$, $\boldsymbol{e}_1, \boldsymbol{e}_2$ independent $N(0, 1)$, shift in x value moves multiplicity region from $[-1, -1]^2$ to $[0, 2]^2$.

For one simple design, the maximum entropy predictions are depicted in figure 10. The predictions are ordered the way they should be (see section 3.3), unlike the midpoint solutions. The kink in both graphs is due to the minimum function and located where $\mu_m(x^*) = \mu_m(x)$.

In section 7, we will compare the maximum entropy solutions with various alternatives using a number of stylized examples.

5. Minmax with entropy

As mentioned in the introduction, the usual minmax approach with a symmetric loss function yields an uninteresting solution which is moreover difficult to defend. We now

consider a generalized minmax problem, where we minimize the maximum risk subject to the corresponding entropy value being bounded below by a specified value. It will become clear that the midpoint and maximum entropy predictions are special cases.

We focus our attention on $q_v(y^* | x, x^*, y; p)$, where our notation makes it explicit that the counterfactual prediction depends on p which is only partially identified.[36] The problem is then to choose a function of 'identifiable objects' from the identified set of p. This function can then be used to produce $q_v(y^* \mid x, x^*, y; p)$. The exercise is at the population level in that the objective is to find a function of (population) distributions, not a sample. So there is no difference between risk and loss here.

Fix $\chi = (x, x^*, y, y^*)$. The aggregate entropy of p at x, x^* is

$$\mathcal{E}(p) = \sum_{\tilde{x} \in \{x, x^*\}} \iint_0^1 \log \partial_v p(e, v, \tilde{x}) \, \mathrm{d}v f_{\boldsymbol{e}}(e) \, \mathrm{d}e. \tag{21}$$

Since $p(e, v, \tilde{x})$ is the v–conditional quantile of $\boldsymbol{p}$ given $\boldsymbol{e} = e$ and $\boldsymbol{x} = \tilde{x}$, $\mathcal{E}(p)$ is simply sum of two conditional entropy values of $\boldsymbol{p}$, fixing x and x^*, i.e.

$$\mathcal{E}(p) = -\sum_{\tilde{x} \in \{x, x^*\}} \iint_0^1 f_{\boldsymbol{p}|\boldsymbol{e},\boldsymbol{x}}(p \mid e, \tilde{x}) \log f_{\boldsymbol{p}|\boldsymbol{e},\boldsymbol{x}}(p \mid e, \tilde{x}) \, \mathrm{d}p f_{\boldsymbol{e}}(e) \, \mathrm{d}e. \tag{22}$$

The expression in (21) is more convenient for us than the one in (22) because the risk is more easily expressed as a function of p. Let $\mathcal{P}$ be the class of functions p that is nondecreasing in v and satisfies the mean restrictions

$$\int_{S_m(\tilde{x})} \int_0^1 p(e, v, \tilde{x}) \, \mathrm{d}v f_{\boldsymbol{e}}(e) \, \mathrm{d}e = \mu_m(\tilde{x}) \pi_m(\tilde{x}), \quad \tilde{x} = x, x^*$$

as well as $0 \leq p(e, v, \tilde{x}) \leq 1$ for all e, v and $\tilde{x} = x, x^*$.

Now let ℓ be a symmetric loss function: $\ell(d, \tilde{d}) = (d - \tilde{d})^2$ is a common choice. Consider the following optimization problem: for some chosen value $\mathcal{E}^*$,

$$\min_d \max_{\substack{p \in \mathcal{P} \\ \mathcal{E}(p) \geq \mathcal{E}^*}} \ell\{d, q_v(p)\}, \tag{23}$$

where $q_v(p)$ is shorthand for $q_v(y^* \mid x, x^*, y; p)$. As noted, if $\mathcal{E}^* = -\infty$, then the solution to (23) is simply the midpoint of the bounds in theorems 1 and 2. If $\mathcal{E}^*$ is chosen to be the maximum achievable value, then the solution to (23) is equal to the maximum entropy prediction, i.e. the counterfactual prediction using p given in theorem 7. Thus, the midpoint and the maximum entropy predictions are special cases of the present approach.

Since the loss function is symmetric, the solution to (23) will be the midpoint of the identified set for $q_v(p)$ restricted to the class of functions $p \in \mathcal{P}$ for which $\mathcal{E}(p) \geq \mathcal{E}^*$. We therefore present the functions p_{-1} and p_1 that correspond to the maximum and minimum attainable values for q_v, respectively: we can then use p_σ, $\sigma = \pm 1$, together with (10) and theorem 1 to compute the solution. We now describe the solutions p_{-1}, p_1.

[36] We provide results for q_v: results for the other counterfactual predictions are messier.

Define

$$\epsilon_\sigma(z) = \int_0^z \exp(-\sigma t^2)\,\mathrm{d}t.$$

Further, for $j = 0, 1, 2$ and $\kappa_0 < \kappa_1$, let

$$I_{\sigma j}(\kappa_0, \kappa_1) = \int_{\kappa_0}^{\kappa_1} t^j \exp(-\sigma t^2)\,\mathrm{d}t \;/\; \big\{\epsilon_\sigma(\kappa_1) - \epsilon_\sigma(\kappa_0)\big\}.$$

Finally, let $\kappa^\circ_{\sigma 0}, \kappa^\circ_{\sigma 1}$ be the maximizers ($\sigma = -1$) or the minimizers ($\sigma = 1$) of

$$\big\{I_{\sigma 2}(\kappa_0, \kappa_1) - I^2_{\sigma 1}(\kappa_0, \kappa_1)\big\} \exp\big\{\mathcal{E}^* - 2\sigma I_{\sigma 2}(\kappa_0, \kappa_1)\big\} \;/\; \big\{\epsilon_\sigma(\kappa_1) - \epsilon_\sigma(\kappa_0)\big\}^2$$

subject to

$$\begin{cases} 2\sigma I_{\sigma 2}(\kappa_0, \kappa_1) + 2\log\big[\big\{\epsilon_\sigma(\kappa_1) - \epsilon_\sigma(\kappa_0)\big\} \;/\; \big\{\kappa_1 - 2\sigma I_{\sigma 2}(\kappa_0, \kappa_1)\big\}\big] = \mathcal{E}^* - \log(1-\mu_m)(1-\mu^*_m), \\ \qquad \big\{\kappa_1 - I_{\sigma 1}(\kappa_0, \kappa_1)\big\} \;/\; \big\{I_{\sigma 1}(\kappa_0, \kappa_1) - \kappa_0\big\} \geq \max\big\{(1-\mu_m)/\mu_m, (1-\mu^*_m)/\mu^*_m\big\}, \end{cases}$$

where $\mu = \mu_m(x)$ and $\mu^* = \mu_m(x^*)$.

Theorem 8. If a solution to (23) exists and $\mathcal{E}^* > -\infty$, then the solution is given by

$$d^* = \tilde{\pi}_y(x^*) + \delta_y \delta_{y^*} \pi_m(x)\pi_m(x^*)\big\{\rho^\circ_{-1}(x, x^*) + \rho^\circ_1(x, x^*)\big\} \;/\; \{2\tilde{\pi}_y(x)\},$$

where $\rho^\circ_\sigma(x, x^*) = \int_0^1 p_\sigma(v) p^*_\sigma(v)\,\mathrm{d}v - \mu_m(x)\mu_m(x^*)$ with

$$\begin{cases} p_\sigma(v) = \mu_m(x) + \big\{1 - \mu_m(x)\big\}\big\{\kappa^\circ_\sigma(v) - I^\circ_{\sigma 1}\big\}/(\kappa^\circ_{\sigma 1} - I^\circ_{\sigma 1}), \\ p^*_\sigma(v) = \mu_m(x^*) + \big\{1 - \mu_m(x^*)\big\}\big\{\kappa^\circ_\sigma(v) - I^\circ_{\sigma 1}\big\}/(\kappa^\circ_{\sigma 1} - I^\circ_{\sigma 1}), \end{cases}$$

where $I^\circ_{\sigma 1} = I_{\sigma 1}(\kappa^\circ_{\sigma 0}, \kappa^\circ_{\sigma 1})$ and $\kappa^\circ_\sigma(v) = \epsilon^{-1}_\sigma(c^\circ_{\sigma 1} v + c^\circ_{\sigma 2})$ with $c^\circ_{\sigma 1} = \epsilon_\sigma(\kappa^\circ_{\sigma 1}) - \epsilon_\sigma(\kappa^\circ_{\sigma 0})$ and $c^\circ_{\sigma 2} = \epsilon_\sigma(\kappa^\circ_{\sigma 0})$. □

Theorem 8 provides a general solution to (23) but it does not establish the uniqueness of the function p_σ that delivers the solution. Note further that we have fixed x, x^* and take them to be different.

Figure 11 depicts the results of a numerical example and shows $\rho^\circ_\sigma(x, x^*)$ as a function of $\mathcal{E}^*$. The entropy constraint is not binding if $\mathcal{E}^* = 0$ and the solution is then simply the midpoint prediction. Otherwise the entropy constrained minmax procedure yields the midpoint of the entropy–constrained identified set, which is an interval whose length shrinks as $\mathcal{E}^*$ increases. Note that the upper bound adjusts faster than the lower bound; this is consistent with our finding in section 7 that the midpoint predictions tend to exceed the maximum entropy ones.

6. Dirichlet approach

6.1 Overview: In this section we discuss yet another possibility involving Dirichlet processes. The approach has a decision–theoretic foundation, but its implementation is considerably more complicated than maximum entropy and it does not do much to mitigate the

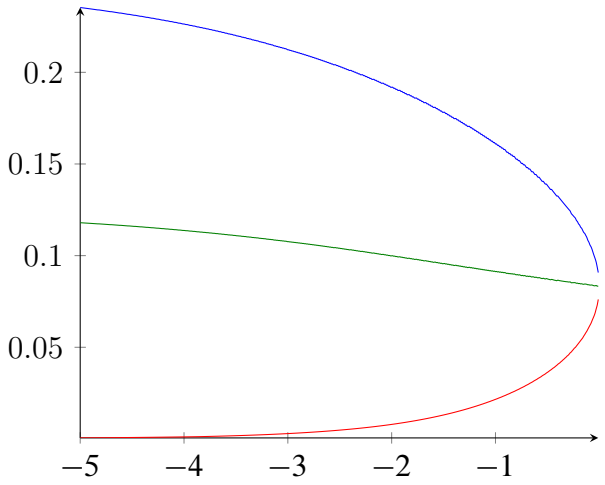

Figure 11: $\rho^{\circ}_{\sigma}(x, x^*)$ as a function of $\mathcal{E}^*$ with $\mu(x) = 0.5$ and $\mu(x^*) = 0.25$. The red line corresponds to $\sigma = 1$ and the blue to $\sigma = -1$. The green line is the midpoint between the blue and red lines.

inherent arbitrariness of the problem at hand. We will compare the results to the maximum entropy ones in section 7.

Recall that the prediction of interest is a deterministic function of p; we will be explicit about this fact in our discussion below. We will treat p as a parameter, consider a prior distribution for p, and look for the prediction that minimizes average risk. Since $p(e, \cdot, x)$ is a conditional quantile function (satisfying a mean constraint), we need to assign a probability measure on the (constrained) space of (conditional) distributions; Dirichlet processes will be used for this purpose.

Let $T_p(x, x^*, y, y^*)$ be the object of interest, i.e. either $q_v(y^* \mid x^*, x, y)$ or $q_{ev}(y^* \mid x^*, x, y)$. Suppose that we have a prior on the parameter space that p^{-1} belongs to: we will say more about the parameter space later.[37] Suppose further that $\boldsymbol{G}$ is a draw from the prior. Then, our prediction based on the draw $\boldsymbol{G}$ will be

$$T_{\boldsymbol{G}^{-1}}(x, x^*, y, y^*), \tag{24}$$

which is a random object. The approach considered in this section is based on the probability distribution of (24). We show in section 6.2 that the mean of (24) is the average risk optimal prediction when a quadratic loss function is used.

We emphasize that there is no Bayesian updating here, because the only information available about the parameter p in the distribution of observables is the regression function μ_m, which we assume to be known (throughout this paper).

We now obtain the likelihood function (in terms of the parameter p). Note from (6) that for $e \in S_m(x)$,

$$\mathbb{P}\{\boldsymbol{y} = (1, 0) \mid \boldsymbol{e} = e, \boldsymbol{v} = v, \boldsymbol{x} = x; p\} = p(e, v, x),$$

[37] p^{-1} is the (generalized) inverse of the quantile function p with respect to the argument v. For a quantile function Q, its generalized inverse G^{-1} is defined by $Q^{-1}(p) = \sup\{p : \ Q(v) \leq p\}$, which is a distribution function. Similarly, for a distribution function G, the generalized inverse G^{-1} is defined by $G^{-1}(v) = \inf\{x : G(x) \geq v\}$, which is a quantile function.

which implies

$$\frac{1}{\pi_m(x)} \int_{S_m(x)} \mathbb{P}\{\boldsymbol{y} = (1,0) \mid \boldsymbol{e} = e, \boldsymbol{x} = x; p\} f_e(e)\, \mathrm{d}e = \mu_m(x). \tag{25}$$

Equation (25) shows that the likelihood function depends on p only through μ_m, which is known. Therefore, since μ_m is known and fixed, the likelihood is also known and it does not depend on p other than via μ_m, which is a different way of saying (as we have before) that μ_m is the only identifiable object about behavior in the multiplicity region.

6.2 Average risk optimality: We now consider an alternative approach, where we formulate the problem as a decision problem of making the 'best' guess for $T_p(x, x^*, y, y^*)$ as a function of x, x^*, y, y^*. To be consistent with our previous discussion, we study the decision problem at the 'population level,' i.e. in terms of the 'population' distribution: there is no sample in the problem considered.

Fixing x, x^*, y, y^*, denote our decision by $d(x, x^*, y, y^*)$. For a given loss function ℓ, the loss (risk) of the decision d is

$$R(d, p) = \ell\big\{d(x, x^*, y, y^*), T_p(x, x^*, y, y^*)\big\}. \tag{26}$$

Like in section 5 our analysis is at the population level as: the loss is not random and therefore the loss and the risk are the same thing.[38] However, since the prediction of interest depends on p, the risk in (26) depends on p, also. Therefore, we consider average risk, i.e. risk averaged over p, for which we need a prior for p.[39]

Now consider the parameter space that the function p^{-1} belongs to. For given values of x, x^*, the function p^{-1} belongs to $\mathcal{F}_{xx^*} = \mathcal{F}_x \cup \mathcal{F}_{x^*}$, where

$$\mathcal{F}_x = \left\{p^{-1}(\cdot,\cdot,x) : \ \frac{1}{\pi_m(x)} \int_{S_m(x)} \int_0^1 p(e, v, x) f_e(e)\, \mathrm{d}e\, \mathrm{d}v = \mu_m(x)\right\}. \tag{27}$$

Therefore, once we have priors ω_x, ω_{x^*} on $\mathcal{F}_x, \mathcal{F}_{x^*}$, we can induce a prior ω_{xx^*} on $\mathcal{F}_{xx^*}$. Since p^{-1} is a (conditional) distribution function, Dirichlet processes provide a natural way of choosing ω_x: Dirichlet processes are the infinite–dimensional extension of the Dirichlet distribution, which is a conjugate prior for the multinomial distribution. We propose an algorithm to do this in section 6.3.

Once we have ω_{xx^*}, the average risk is given by

$$\bar{R}(d) = \int_{\mathcal{F}_{xx^*}} R(d, p)\, \mathrm{d}\omega_{xx^*}(p^{-1}) \geq \min_s \int_{\mathcal{F}_{xx^*}} \ell\big\{s, p(x, x^*, y, y^*)\big\}\, \mathrm{d}\omega_{xx^*}(p^{-1}),$$

where the inequality shows that the function d^* that minimizes $\bar{R}$ can be found by separate minimization for each combination of x, x^*, y, y^*. Indeed, the decision d^* that minimizes

[38] We could average y, y^*, x, x^* by using the likelihood function, which is independent of p (other than via μ_m) and hence is known. But it does not make any difference from simply fixing them as we do here.

[39] As we discussed earlier, a simple (unconstrained) minmax approach leads to the midpoint prediction if ℓ is symmetric. But the midpoint prediction has flaws as will become apparent in section 7.1.

$\bar{R}$ is

$$d^*(x, x^*, y, y^*) = \operatorname*{argmin}_{s\in[0,1]} \int_{\mathcal{F}_{xx^*}} \ell\{s, T_p(x, x^*, y, y^*)\}\, \mathrm{d}\omega_{xx^*}(p^{-1}).$$

Thus, when the loss function is quadratic, $d^*(x, x^*, y, y^*)$ is simply a mean as we explained below (24).

6.3 Dirichlet prior: We now describe a way of drawing probability distributions from ω_{xx^*}, for which it suffices that we can draw probability distributions from ω_x on support $\mathcal{F}_x$. To keep things simple, we will restrict $\mathcal{F}_x$ by considering elements p^{-1} that do not depend on e for all $e \in S_m(x)$. With maximum entropy, the fact that the p used is flat in e is a *result* of the maximum entropy procedure, but here it is a *restriction*. Dropping the argument e from our notation, we have

$$\mathcal{F}_x = \left\{p^{-1}(\cdot, x) : \int_0^1 p(v, x)\, \mathrm{d}v = \mu_m(x)\right\}, \tag{28}$$

which is a collection of distribution functions constrained to have mean $\mu_m(x)$.

A Dirichlet process $\mathcal{D}$ is commonly used as a probability distribution for probability distributions. It is a stochastic process whose finite marginals are described by the Dirichlet distribution. It is characeterized by a (Dirichlet process) prior H, not to be confused with ω_x, and a hyperparameter $\zeta > 0$. H and ζ correspond to the mean and the precision of the Dirichlet process, respectively, in the following sense. If $\boldsymbol{G} \sim \mathcal{D}(\zeta, H)$, then for an arbitrary event A we have

$$\mathbb{E}\boldsymbol{G}(A) = H(A), \qquad \mathbb{V}\boldsymbol{G}(A) = H(A)\{1 - H(A)\} \,/\, (1 + \zeta).$$

In the discussion below we will use the uniform distribution $U[0, 1]$ for H and let $\zeta = 1$.

A probability distribution $\boldsymbol{G}$ drawn from a Dirichlet process is always a discrete distribution with infinitely many mass points. A standard procedure for drawing $\boldsymbol{G}$ from e.g. $\mathcal{D}(1, U[0, 1])$ is the following. Let $\boldsymbol{\beta}_1, \boldsymbol{\tau}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}_2, \ldots$ be independent draws from $U[0, 1]$, let $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j \prod_{t=1}^{j-1}(1 - \boldsymbol{\beta}_t)$ for $j > 1$. Then, the distribution $\boldsymbol{G}$ assigning probability masses $\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \cdots$ to the mass points $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \cdots$, respectively, can be shown to be a draw from $\mathcal{D}(1, U[0, 1])$. This procedure is known as the 'stick–breaking' construction. See e.g. Teh (2010) for details.

In our case, however, the standard stick–breaking construction does not provide what we want because the parameter space $\mathcal{F}_x$ has a constraint on the mean. Below, we explain how we modify the standard procedure to draw a probability distribution from $\mathcal{F}_x$. Since $\mathcal{F}_x$ depends on x only through $\mu_m(x)$, we will drop the dependence on x in our notation and simply use μ_m in lieu of $\mu_m(x)$ for the remainder of this section.

Let the $\boldsymbol{\beta}_j$'s and $\boldsymbol{\beta}_j^*$'s be defined as above. Suppose that the $\boldsymbol{\tau}_j$'s are i.i.d. $U[0, 1]$ and (unconditionally) independent of the $\boldsymbol{\beta}_j$'s. So, if we would not impose a mean condition then our procedure below would be equivalent to standard stick–breaking. We now determine the locations of the mass points $\boldsymbol{\tau}_j$'s to ensure that

$$\boldsymbol{W}_1 = \sum_{j=1}^{\infty} \boldsymbol{\beta}_j^* \boldsymbol{\tau}_j = \mu_m,$$

for which we will draw $\boldsymbol{\tau}_j$'s from appropriate conditional distributions.

First, note that the distribution of

$$\boldsymbol{W}_j = \frac{\boldsymbol{\beta}_j}{\boldsymbol{\beta}_j^*} \sum_{i=j}^{\infty} \boldsymbol{\beta}_i^* \boldsymbol{\tau}_i,$$

does not depend on j and that

$$\forall\, j : \boldsymbol{W}_j = \boldsymbol{\beta}_j \boldsymbol{\tau}_j + (1 - \boldsymbol{\beta}_j) \boldsymbol{W}_{j+1}. \tag{29}$$

Further, the Bayes rule says that

$$f_{\boldsymbol{\tau}_j | \boldsymbol{\beta}_j, \boldsymbol{W}_j}(\tau \mid \beta, \mu) = \frac{f_{\boldsymbol{W}_j | \boldsymbol{\tau}_j, \boldsymbol{\beta}_j}(\mu \mid \tau, \beta)}{f_{\boldsymbol{W}_j | \boldsymbol{\beta}_j}(\mu \mid \beta)}. \tag{30}$$

Here, by (29) and the fact that the distribution of $\boldsymbol{W}_j$ does not depend on j, it follows that

$$\mathbb{P}(\boldsymbol{W}_j \le w \mid \boldsymbol{\tau}_j = \tau_j, \boldsymbol{\beta}_j = \beta_j) = \mathbb{P}\big\{\beta_j \tau_j + (1 - \beta_j)\tau_j \boldsymbol{W}_{j+1} \le w\big\} = F^{\circ}\bigg(\frac{w - \beta_j \tau_j}{1 - \beta_j}\bigg), \tag{31}$$

where F° is the distribution function of $\boldsymbol{W}_1$. We discuss further down how to compute F° efficiently.

Finally, note from (30) and (31) that $F_{\boldsymbol{\tau}_j | \boldsymbol{\beta}_j, \boldsymbol{W}_j}(\tau \mid \beta, \mu) = \tilde{F}^{\circ}(\tau; \beta, \mu)$, where

$$\tilde{F}^{\circ}(t; \tilde{\beta}, \tilde{\mu}) = \frac{F^{\circ}\Big(\frac{\tilde{\mu}}{1 - \tilde{\beta}}\Big) - F^{\circ}\Big(\frac{\tilde{\mu} - \tilde{\beta} t}{1 - \tilde{\beta}}\Big)}{F^{\circ}\Big(\frac{\tilde{\mu}}{1 - \tilde{\beta}}\Big) - F^{\circ}\Big(\frac{\tilde{\mu} - \tilde{\beta}}{1 - \tilde{\beta}}\Big)}, \quad \max\Big\{0, \frac{\tilde{\mu} - (1 - \tilde{\beta})}{\tilde{\beta}}\Big\} \le t \le \min\Big(1, \frac{\tilde{\mu}}{\tilde{\beta}}\Big). \tag{32}$$

This motivates the following procedure.

Procedure 1. Do the following:

1. Draw $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots \sim U(0, 1)$;
2. Draw $\boldsymbol{\tau}_1$ from $\tilde{F}^{\circ}(\cdot; \boldsymbol{\beta}_1, \mu_m)$;
3. Draw $\boldsymbol{\tau}_2$ from $\tilde{F}^{\circ}\big\{\cdot; \boldsymbol{\beta}_2, (\mu_m - \boldsymbol{\beta}_1 \boldsymbol{\tau}_1) \,/\, (1 - \boldsymbol{\beta}_1)\big\}$;
4. Draw $\boldsymbol{\tau}_3$ from $\tilde{F}^{\circ}\big[\cdot; \boldsymbol{\beta}_3, \{\mu_m - \boldsymbol{\beta}_1 \boldsymbol{\tau}_1 - (1 - \boldsymbol{\beta}_1)\boldsymbol{\beta}_2 \boldsymbol{\tau}_2\} \,/\, \{(1 - \boldsymbol{\beta}_1)(1 - \boldsymbol{\beta}_2)\}\big]$;
5. Continue ad nauseam.
6. Let $\boldsymbol{G}$ be the probability distribution with mass points $\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \ldots$ at $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \ldots$, where $\boldsymbol{\beta}_j^*$ is as defined in the text.

□

Then, $\boldsymbol{G}$ is a draw from $\mathcal{D}(1, U[0,1])$ unconditionally. But once we fix $\boldsymbol{G}$'s mean at μ_m, its distribution is different for which we use the notation $\boldsymbol{G} \sim \mathcal{D}^*_{\mu_m} = \mathcal{D}^*_{\mu_m}(1, U[0,1])$.

From (32) it follows that for any $j > 1$,

$$\mu_m - \prod_{t=1}^{j-1}(1-\boldsymbol{\beta}_t) \leq \sum_{t=1}^{j} \boldsymbol{\beta}^*_t \boldsymbol{\tau}_t \leq \mu_m,$$

which ensures that $\boldsymbol{G}$ has mean μ_m, as required: taking $j \to \infty$ leads to $\boldsymbol{W}_1 = \mu_m$.

Our procedure requires us to implement a draw from $\tilde{F}^\circ(\cdot; \tilde{\beta}, \tilde{\mu})$. All one has to do for this purpose is to compute

$$\boldsymbol{\tau} = \frac{\tilde{\mu} - (1-\tilde{\beta}) F^{\circ -1}\Big\{(1-\boldsymbol{\tau}^*) F^\circ\Big(\frac{\tilde{\mu}}{1-\tilde{\beta}}\Big) + \boldsymbol{\tau}^* F^\circ\Big(\frac{\tilde{\mu}-\tilde{\beta}}{1-\tilde{\beta}}\Big)\Big\}}{\tilde{\beta}},$$

where $\boldsymbol{\tau}^*$ is a draw from a $U(0,1)$.

It now remains to be shown how to compute F°. From (29) it follows that

$$\forall w : F^\circ(w) = \int_0^1 \int_0^1 F^\circ\Big(\frac{w - \beta\tau}{1-\beta}\Big)\, \mathrm{d}\beta\, \mathrm{d}\tau,$$

which allows us to solve for F° numerically.

7. Comparison of methods

In this section we compare the maximum entropy approach with other possibilities. Section 7.1 shows that the midpoint prediction method is inconsistent in the sense that there generally is no single function p that produces the midpoint predictions in all cases. Section 7.2 uses two simple designs to compare the sharp bounds on q_v to regression predictions, midpoint predictions, and maximum entropy predictions. Section 7.3 does the same for q_{ev}. Finally, section 7.4 compares and contrasts all methods to the Dirichlet–process–based idea of section 6.

7.1 Inconsistency of the midpoint method: As promised, we now show that there generally exists no single function p that is consistent with the midpoint prediction for all values of x^*. In other words, we can construct examples in which any function p that rationalizes the midpoint prediction for one combination of x, x^* does not rationalize the midpoint prediction for a different combination, as evidenced by example 1.

Example 1. Suppose that $x \neq x^*$, $y = y^* = (1,0)$, $\pi_m(x) = \pi_m(x^*) = 1$, and $\mu_m(x) = \mu_m(x^*) = \tilde{\mu}$ for some $0 < \tilde{\mu} \leq 1$. Suppose that there exists a function p° such that the predictions based on p° are the midpoint predictions. Then, for any $\bar{x}$ in the support of $\boldsymbol{x}$, the predictions based on p° must satisfy

$$q^\circ_v(y^* \mid \bar{x}, \bar{x}, y) = \frac{\mathbb{E}\{p^\circ(\boldsymbol{e}, \boldsymbol{v}, \bar{x}) p^\circ(\boldsymbol{e}^*, \boldsymbol{v}, \bar{x})\}}{\mu_m(\bar{x})} = q^\circ_{ev}(y^* \mid \bar{x}, \bar{x}, y) = \frac{\mathbb{E}p^{\circ 2}(\boldsymbol{e}, \boldsymbol{v}, \bar{x})}{\mu_m(\bar{x})},$$

because the midpoint predictions of $q_v(y^*|\bar{x}, \bar{x}, y)$ and $q_{ev}(y^*|\bar{x}, \bar{x}, y)$ coincide by theorems 2 and 5. Therefore, we have $\mathbb{E}\,\mathbb{V}\{p^\circ(\boldsymbol{e}, \boldsymbol{v}, x) \,|\, \boldsymbol{v}\} = 0$ a.s., from which we conclude that p° is flat in e. From hereon we omit the argument e without loss of generality, i.e. $p^\circ(e, \cdot, \cdot) = p^\circ(\cdot, \cdot)$.

Now, suppose that $\tilde{\mu} = 1\ /\ 2$, in which case the midpoint prediction of $q_{ev}(y^* \mid x^*, x, y)$ equals $1/2$, and the midpoint predictions of $q_{ev}(y^* \mid x, x, y)$ and $q_{ev}(y^* \mid x^*, x^*, y)$ are both equal to $3\ /\ 4$ by theorem 5. Therefore, if p° is to be consistent with the midpoint predictions we must have

$$q^\circ_{ev}(y^* \mid x^*, x, y) = \frac{\mathbb{E}\{p^\circ(\boldsymbol{v}, x)p^\circ(\boldsymbol{v}, x^*)\}}{\tilde{\mu}} = \frac{1}{2}, \tag{33}$$

$$q^\circ_{ev}(y^* \mid x, x, y) = \frac{\mathbb{E}p^{\circ 2}(\boldsymbol{v}, x)}{\tilde{\mu}} = \frac{3}{4}, \tag{34}$$

$$q^\circ_{ev}(y^* \mid x^*, x^*, y) = \frac{\mathbb{E}p^{\circ 2}(\boldsymbol{v}, x^*)}{\tilde{\mu}} = \frac{3}{4}. \tag{35}$$

Equation (33) implies that $\mathrm{Cov}\{p^\circ(\boldsymbol{v}, x), p^\circ(\boldsymbol{v}, x^*)\} = 0$, which means that either $p^\circ(\boldsymbol{v}, x)$ or $p^\circ(\boldsymbol{v}, x^*)$ should be equal to $\tilde{\mu} = 1\ /\ 2$ with probability one, because p° is monotonic in v. But that conclusion contradicts either (34) or (35). Therefore, there exists no function p° that rationalizes the midpoint predictions. □

There are further arguments against the use of midpoint predictions, as will become apparent in the remainder of section 7.

7.2 Case 1: $\boldsymbol{v}^* = \boldsymbol{v}$: We now turn to a comparison of the prediction methods for the case discussed in section 3.1, i.e. if $q_v(y^* \mid x^*, x, y)$ is the object of interest. Both here and in section 7.3 we focus on the case in which $\mu_m(x^*) = \mu_m(x)$, which simplifies the comparison while still allowing us to convey the main issues at hand.

Recall from theorem 7 that the maximum entropy method produces a solution p^* which (in $S_m(x)$) depends on x only through $\mu_m(x)$. Therefore, $\mu_m(x) = \mu_m(x^*)$ implies that $p^*(e, v, x) = p^*(e, v, x^*)$ for all $e \in S_m(x)$, v, and hence it follows that

$$\begin{aligned}\rho^*(x, x^*) &= \mathbb{V}\left\{\hbar^*(\boldsymbol{v}, \boldsymbol{x}) \,\middle|\, \boldsymbol{x} = x\right\} \\ &= \int_0^1 p^2 A\{p, \lambda_m(x)\}\,\mathrm{d}p - \left\{\int_0^1 pA\{p, \lambda_m(x)\}\,\mathrm{d}p\right\}^2 = \mathcal{L}''\{\lambda_m(x)\},\end{aligned} \tag{36}$$

where A is defined in (20). Example 2 is based on (36).

Example 2. Suppose that $\pi_{10}(x) = \pi_{10}(x^*) = \pi_{01}(x) = \pi_{01}(x^*)$, $\pi_m(x) = \pi_m(x^*)$, $\pi_{00}(x) = \pi_{00}(x^*)$, and $\mu_m(x) = \mu_m(x^*) = \tilde{\mu}$ for some $0 < \tilde{\mu} \leq 1$. We consider two cases, namely $\pi_{10}(x) = 1/4$, $\pi_m(x) = 4/9$ and $\pi_{10}(x) = 0$, $\pi_m(x) = 1$. Figure 12 depicts prediction probabilities as a function of $\tilde{\mu}$ in each case.

In both cases, the midpoint prediction method yields higher predictions than the maximum entropy method. In the extreme case $\pi_m(x) = 1$ (depicted in the right panel), the difference is especially pronounced when $\tilde{\mu}$ is close to zero. The upper bound there

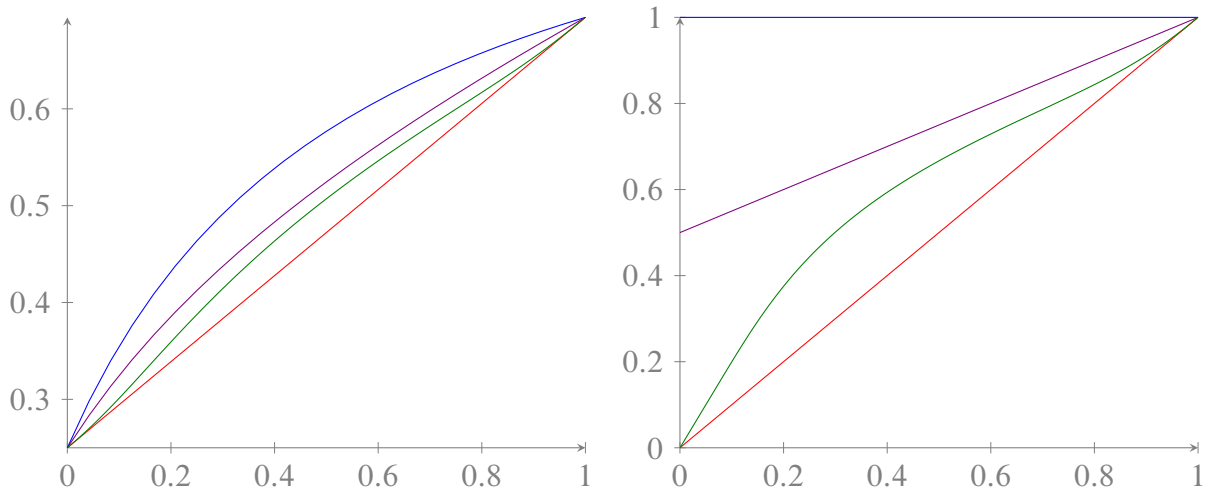

Figure 12: Infeasible predictions as a function of $\mu_m(x) = \mu_m(x^*)$ if $y = y^* = (1, 0)$ using regression (red; also smallest), largest (blue), midpoint (purple), and maximum entropy (green) methods. In both panels $\pi_m(x) = \pi_m(x^*)$ and $\pi_y(x) = \pi_y(x^*)$. In the left panel, $\pi_m(x) = 4/9$ and $\pi_y(x) = 1/4$, whereas in the right panel $\pi_m(x) = 1$ and $\pi_y(x) = 0$.

equals one for all $\tilde{\mu} \in (0, 1]$ since the largest possible value of $\rho(x, x^*)$ is $\tilde{\mu}(1 - \tilde{\mu})$, which corresponds to

$$p(e, v, x) = \mathbb{1}(v > 1 - \tilde{\mu}), \quad e \in S_m(x), \tag{37}$$

From theorem 1, we know that (in the right panel case)

$$q_v\{(1, 0) \mid x^*, x, (1, 0)\} = \frac{1}{\tilde{\mu}} \int_0^1 \hbar(v, x)\hbar(v, x^*)\, \mathrm{d}v. \tag{38}$$

For most functions p, (38) will be close to zero if $\tilde{\mu}$ is close to zero. However, for p in (37), the value of q_v is one for any value of $\tilde{\mu}$, which is an extreme possibility. In a way, then, the midpoint method is overly conservative since it puts an inordinate amount of weight on an extreme choice of p in (37); see also example 5. The maximum entropy method, in contrast, does not have this problem. □

7.3 Case 2: $e^* = e$ and $v^* = v$: Since $p^*(e, v, x)$ defined in theorem 7 does not vary with e over $S_m(x)$, the function ρ in theorem 3 using p^* is given by

$$\rho^*(x, x^*) = \mathbb{E}\big[\mathrm{Cov}\{p^*(\boldsymbol{e}, \boldsymbol{v}, x)p^*(\boldsymbol{e}, \boldsymbol{v}, x^*) \mid \boldsymbol{e}\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big] = \mathrm{Cov}\big\{\hbar^*(\boldsymbol{v}, x), \hbar^*(\boldsymbol{v}, x^*)\big\}.$$

We continue to focus on the case in which $\mu_m(x) = \mu_m(x^*)$ and hence $\hbar^*(v, x) = \hbar^*(v, x^*)$ for all v, as explained in section 7.2. Consequently, the formula for $\rho^*(x, x^*)$ is again given by (36), i.e. $\rho^*(x, x^*) = \mathscr{L}''\{\lambda_m(x^*)\}$. The function a_y in theorem 3 using p^* is provided in (54) in appendix C, which can be used to determine the maximum entropy value of $\alpha_{yy}(x, x^*)$. Indeed, if the S regions at x coincide with those at x^* and $\mu_m(x) = \mu_m(x^*)$, then the maximum entropy choice of $\alpha_{yy^*}(x, x^*)$ is for $y^* = y = (1, 0)$ given by

$$\big[\pi_y(x)\{1 - \pi_y(x)\} + \mu_m^2(x)\pi_m(x)\{1 - \pi_m(x)\} - 2\pi_y(x)\pi_m(x)\mu_m(x)\big] \,/\, \tilde{\pi}_y(x). \tag{39}$$

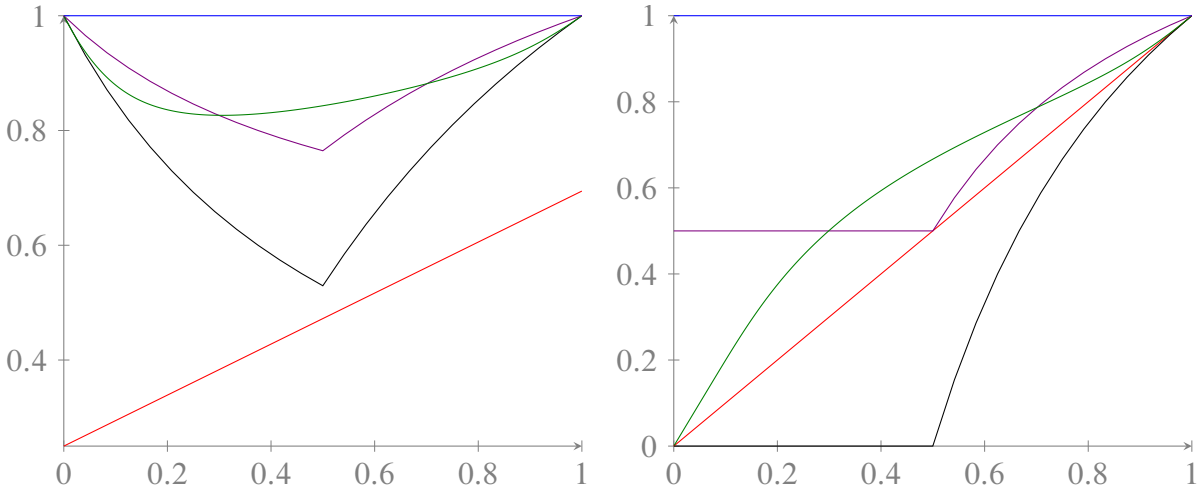

Figure 13: Infeasible predictions as a function of $\mu_m(x) = \mu_m(x^*)$ if x and x^* correspond to the same S–regions with $x \neq x^*$, and $y = y^* = (1, 0)$. The predictions are produced using the regression (red), smallest (black), midpoint (purple), largest (blue), and maximum entropy (green) methods. In the left panel, $\pi_m(x) = 4 \ / \ 9$ and $\pi_y(x) = 1 \ / \ 4$, whereas in the right panel $\pi_m(x) = 1$ and $\pi_y(x) = 0$.

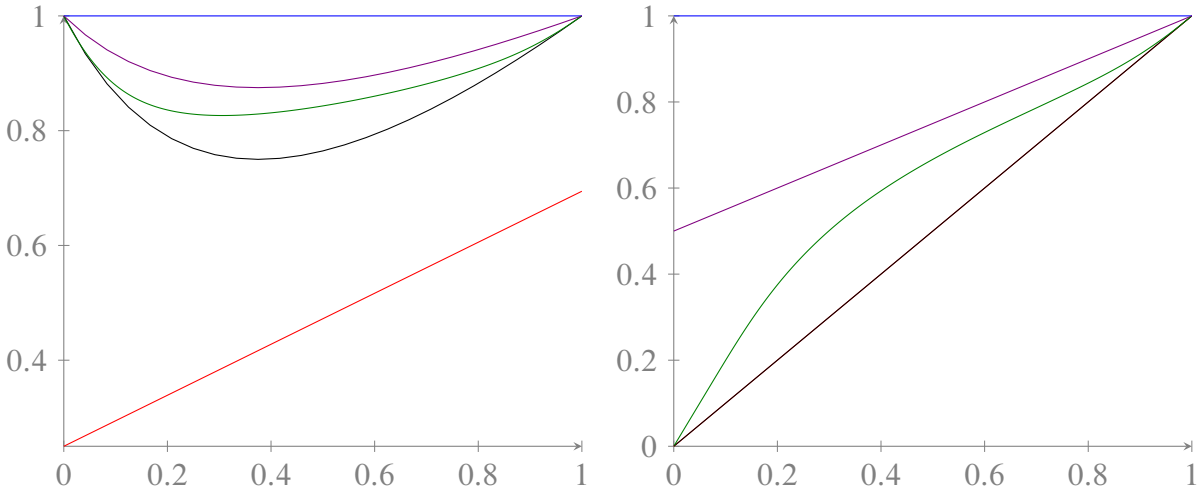

Figure 14: Infeasible predictions as a function of $\mu_m(x)$ if $x = x^*$ and $y = y^* = (1, 0)$, using regression (red), smallest (black), midpoint (purple), largest (blue), and maximum entropy (green) methods. In the left panel, $\pi_m(x) = 4 \ / \ 9$ and $\pi_y(x) = 1 \ / \ 4$, whereas in the right panel $\pi_m(x) = 1$ and $\pi_y(x) = 0$.

Example 3 below is based on (36) and (39).

Example 3. Consider x, x^* such that the S–regions coincide and $\mu_m(x) = \mu_m(x^*) = \tilde{\mu}$ for some $0 < \tilde{\mu} \leq 1$. We consider four cases: each of the two cases considered in example 2 where we now distinguish between $x \neq x^*$ and $x = x^*$. Holding $\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}$ fixed does not necessarily yield the same outcome because of the incompleteness of the model: a different equilibrium can be selected in the multiplicity region. Figures 13 and 14 depict the cases

$x^* \neq x$ and $x^* = x$, respectively, where in both figures $y^* = y = (1, 0)$. Recall that the regression prediction $\tilde{\pi}_{10}(x^*)$ need not belong to the identified set for q_{ev}. Therefore, the regression prediction is a poor choice if the object of interest is $q_{ev}(y^* \mid x, x^*, y)$.

First consider the right panel in figure 13. The midpoint method yields 0.5 for all $\tilde{\mu} < 0.5$. So the 'conservative prediction' problem mentioned in example 2 arises here, also. Further, the maximum entropy predictions behave more smoothly than e.g. the midpoint predictions: the midpoint predictions have a noticeable kink at $\tilde{\mu} = 0.5$.

Comparing figures 13 and 14 shows that the midpoint method yields drastically different predictions depending on whether or not $x^* = x$. For instance, in the extreme case $\pi_m(x) = \pi_m(x^*) = 1$, if $\tilde{\mu} = 1/2$ then theorem 5 implies that the sharp identified interval is given by

$$\big[\mathbb{1}(x^* = x) \;/\; 2,\; 1\big].$$

So a minute change in x^* can result in a 25 percentage point jump in the midpoint prediction, which is undesirable. □

Intuition for the discontinuity problem illustrated in example 3 was provided below theorem 5. Recall that the sharp bounds change *continuously* if $p(\boldsymbol{e}, \boldsymbol{v}, x)$ and $p(\boldsymbol{e}, \boldsymbol{v}, x^*)$ are restricted to have nonnegative correlation. For instance, if the function p is flat in e, then the discontinuity issue disappears.

Since all available information about the function p is contained in $\mu_m(x)$ and $S_m(x)$ (and hence in $\pi_m(x)$), restricting p to be flat in e does not lead to a loss of 'information' in the maximum entropy sense: the restriction that p is flat in e is not binding in the maximum entropy optimization problem and hence the maximum entropy solution is flat in e. With the Dirichlet–based procedure of section 6, we restricted p to be flat in e. However, since the parameter space $\mathcal{F}_x$ in (27) only depends on x via $\mu_m(x)$ and $S_m(x)$, any prior that depends on x only through the identifiable objects $\mu_m(x)$ and $S_m(x)$ will not impose 'extra' information about how p depends on e, and therefore will produce predictions that are continuous in x.

7.4 Dirichlet–based predictions, midpoints, and maximum entropy: We now consider the average risk optimal predictions using the prior we proposed in section 6. Instead of choosing a particular loss function, we consider the distribution of the prediction when p is drawn by the algorithm described in section 6.3. For instance, the mean and median are the average risk optimal predictions using a quadratic and an absolute deviation loss function, respectively.

Example 4. Consider again the scenario of examples 2 and 3, specifically the cases depicted in the right panels of figures 12 and 13, i.e. $\pi_m(x) = \pi_m(x^*) = 1$ and $y = y^* = (1, 0)$. Recall that by theorem 1 the infeasible prediction $q_v(y^* \mid x^*, x, y)$ in this case is given by

$$\frac{\mathbb{E}\big\{\hbar(\boldsymbol{v}, x)\hbar(\boldsymbol{v}, x^*)\big\}}{\mu_m(x)}. \tag{40}$$

Likewise, the infeasible prediction for $q_{ev}(y^* \mid y, x, x^*)$ is given by

$$\frac{\mathbb{E}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x)p(\boldsymbol{e}, \boldsymbol{v}, x^*)\big\}}{\mu_m(x)}. \tag{41}$$

Now, consider the restricted parameter space defined in (28). Then, in our current design, (40) and (41) coincide and equal

$$T_p(x, x^*, y, y^*) = \frac{\mathbb{E}\{p(\boldsymbol{v}, x)p(\boldsymbol{v}, x^*)\}}{\mu_m(x)}. \tag{42}$$

Note that the S–regions do not change in the current setup. If the parameter space for p is not restricted to $\mathcal{F}_x$ then (42) is an upper bound of both (40) and (41).

Figure 15 depicts the distribution of (42) when the function p is a random draw from the Dirichlet–like–process $\mathcal{D}^*_{\mu_m}$ described in section 6.3. Recall from section 6 that this entails pretending that each probability distribution is a possible distribution of $\boldsymbol{p}$ for given μ_m–value with the caveat that the draws from $\mathcal{D}^*_{\mu_m}$ are discrete distributions whereas the distribution of $\boldsymbol{p}$ is in most cases continuous.

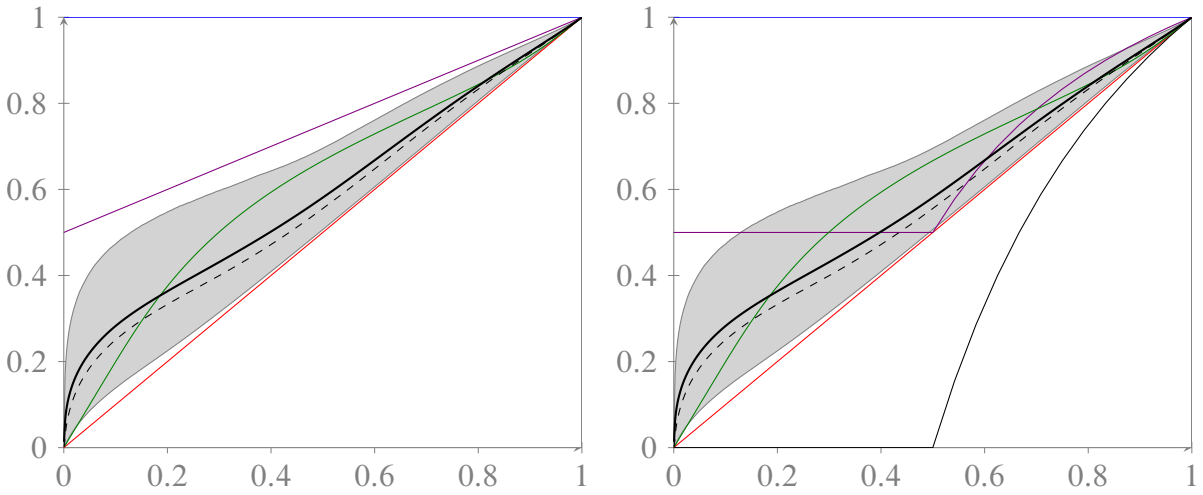

Figure 15: Infeasible predictions as a function of $\mu_m(x) = \mu_m(x^*)$ if $y = y^* = (1, 0)$ using regression (red), largest (blue), midpoint (purple), and maximum entropy (green) methods. The grey regions depict the area between 10% and 90% quantiles of the Dirichlet process–based experiment described in section 6.3 as a function of its mean with the dashed line indicating the median. In both panels $\pi_m(x) = \pi_m(x^*) = 1$ and $x \neq x^*$. The left panel depicts q_v and the right panel depicts q_{ev}. The maximum entropy prediction and the Dirichlet–based predictions are the same in both panels while the bounds (and hence the midpoints) are different.

For each value of μ_m, each draw from $\mathcal{D}^*_{\mu_m}$ produces a probability distribution for $\boldsymbol{p}$, whose quantile function corresponds to (a draw of) the function p. Therefore, each draw from $\mathcal{D}^*_{\mu_m}$ produces the prediction $T_p(x, x^*, y, y^*)$ displayed in (42) for a different function p.

The results of our experiments are depicted in figure 15. The left panel in figure 15 is identical to the right panel in figure 12 and the right panel in figure 15 is identical to the right panel in figure 13, except that there are some additions. The additions in the two panels are identical. The thick solid black line represents the mean of the distribution of

$T_p(x, x^*, y, y^*)$ as a function of μ_m, the dashed line the median, and the bottom and top of the gray shaded area the bottom and top decile, respectively.

The Dirichlet graphs are consistent with the maximum entropy predictions (green curves). But the Dirichlet graphs are inconsistent with the midpoint predictions. As we explained below (42), the Dirichlet graphs are generally upper bounds to the infeasible predictions, with equality if p is flat in e. They moreover depend on x only through $\mu_m(x)$. The fact that the midpoint prediction in the left panel of figure 15 is for essentially all μ_m above the 90% quantile of the distribution of $T_p(x, x^*, y, y^*)$ produced by the Dirichlet experiment is troubling. The results in the right panel are less problematic, but the midpoint predictions nevertheless look at odds with the Dirichlet draws.

The Dirichlet experiment has limitations. First, Dirichlet process draws are, as noted, discrete probability distributions whereas $\boldsymbol{p}$ is usually (though not necessarily) continuously distributed. Further, we chose a uniform prior and pseudocount hyperparameter equal to one for the Dirichlet process for convenience.[40] Different choices for the input parameters produce different results. Finally, we only draw distributions for $\boldsymbol{p}$ (and hence functions p) that do not depend on the value of $e \in S_m(x)$ and only depend on x via $\mu_m(x)$. But that actually helps the midpoint prediction because (42) is only an upper bound to the infeasible prediction: the infeasible prediction can in fact be lower still. □

While acknowledging its limitations, example 4 provides further evidence against the use of the midpoint method. We conclude with an example that emphasizes the fact that the midpoint prediction puts too much weight on extreme possibilities.

Example 5. Consider the case in which $x \neq x^*$, $y = y^* = (1, 0)$, $\pi_m(x) = \pi_m(x^*) = 1$, but where $\mu_m(x^*)$ may or may not equal $\mu_m(x)$. Then, by theorem 1,

$$q_v(y^* \mid x^*, x, y) = \mu_m(x^*) + \frac{\mathrm{Cov}\{\hbar(\boldsymbol{v}, x), \hbar(\boldsymbol{v}, x^*)\}}{\mu_m(x)} = \frac{\mathbb{E}\{\hbar(\boldsymbol{v}, x)\hbar(\boldsymbol{v}, x^*)\}}{\mu_m(x)},$$

where $\hbar(v, x) = \mathbb{E}p(\boldsymbol{e}, v, x)$, as before. The highest attainable value of $q_v(y^* \mid x^*, x, y)$ is one. Here is what needs to happen to get $q_v(y^* \mid x^*, x, y)$ equal to one. Since $\hbar$ cannot exceed one, it must be true that $\mu_m(x^*) \geq \mu_m(x)$ and that $\hbar(v, x^*) = 1$ whenever $\hbar(v, x) > 0$. Since $\hbar$ is nondecreasing in v, it must be true that $\hbar(v, x) = \mathbb{1}\{v \geq 1 - S_m(x)\}$ and that $\hbar(v, x^*) = \hbar(v, x)$ for all $v \geq 1 - S_m(x)$.

The midpoint solution assigns weight 0.5 to this case and weight 0.5 to the case in which $\hbar(v, x) = \mu_m(x)$ for all values of v. □

8. Inference

So far we have focused on counterfactuals in games, for which we assumed that both the structure and probability distributions were known. However, in practice payoffs need

[40] The prior determines how likely it is that the mass points of a probability distribution drawn are in particular locations. A uniform prior means that those mass points can be anywhere in the unit interval with equal probability: the distributions generated by the Dirichlet process are themselves not uniform. Moreover, recall that we generate distributions conditional on the distributions having mean μ_m, not unconditionally. The pseudocount hyperparameter, determines the relative size of mass points: the value chosen by us (one) implies that the probability mass at the first mass point is on average $1/2$, at the second $1/4$, etcetera.

to be estimated which produces estimation error. In this section we discuss inferential issues for maximum entropy (ME) counterfactual predictions, focusing on q_v. Like before, payoffs are not the main focus of this paper, so we make the following assumption.

Assumption A. *For some* $0 < r \leq 1/2$ *and some* $\boldsymbol{X}$, $n^r(\hat{\boldsymbol{\mu}} - \mu) \xrightarrow{d} \boldsymbol{X}$, *where* n *is the sample size,* $\mu = [\mu_m(x), \mu_m(x^*)]^\intercal$, *and* $\hat{\boldsymbol{\mu}}$ *is an estimator of* μ. □

As noted in the introduction, finding $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{X}$ has been discussed by many authors with their exact form depending on the assumptions made. In Tamer (2003), $r = 1/2$ and $\boldsymbol{X}$ is normally distributed.

In the interest of space conservation we assume that $x \neq x^*$: the case $x = x^*$ is analogous. Further, we focus on doing inference on the nontrivial portion of the counterfactual prediction, i.e. the object $\rho(x, x^*)$ (or more precisely, the one that corresponds to the ME solution of p), which only depends on μ. We now translate knowledge of the large sample behavior of $\hat{\boldsymbol{\mu}}$ into a result on the limit properties on $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\rho}}(x, x^*)$, an estimator of $\rho = \rho(x, x^*)$ which is defined in (10) and used in theorem 1; theorems 6 and 7 provide the ME solution to be used in theorem 1.

Indeed, theorem 7 tells us the explicit functional form of $\hbar^*(\cdot, x)$ and $\hbar^*(\cdot, x^*)$: they depend on $\lambda = [\lambda_m(x), \lambda_m(x^*)]^\intercal$, which is related to μ via (20). The function $\mathcal{L}$ linking λ and μ is defined in (18), which is well–behaved as will be apparent below. Define $\hat{\boldsymbol{\lambda}} = [\hat{\boldsymbol{\lambda}}_m(x), \hat{\boldsymbol{\lambda}}_m(x^*)]^\intercal$ with $\hat{\boldsymbol{\lambda}}_m(\tilde{x}) = \mathcal{L}'^{-1}\{\hat{\boldsymbol{\mu}}_m(\tilde{x})\}$ for $\tilde{x} = x, x^*$.

Theorem 9. Under assumption A, $n^r(\hat{\boldsymbol{\lambda}} - \lambda) \xrightarrow{d} \mathbb{D}_{\mathcal{L}}^{-1}\boldsymbol{X}$, where $\mathbb{D}_{\mathcal{L}}$ is a diagonal matrix with diagonal elements $\mathcal{L}'\{\lambda_m(x)\}$ and $\mathcal{L}'\{\lambda_m(x^*)\}$. □

In theorem 9 invertibility of $\mathbb{D}_{\mathcal{L}}$ is implicit. Theorem 9 establishes that the limit distribution of $\hat{\boldsymbol{\lambda}}$ follows from that of $\hat{\boldsymbol{\mu}}$, which was assumed known in assumption A. Since the counterfactual predictions are simple functions of $\hbar^*$, conducting inference on them is just a matter of applying the Delta method. The above results do not depend on which counterfactual one considers but, as mentioned, we focus on q_v from here on.

Making the dependence of the ME solution of ρ on λ explicit in our notation, we obtain

$$\rho(x, x^*; \lambda) = \int_0^1 \hbar^*\{v, x; \lambda_m(x)\}\hbar^*\{v, x^*; \lambda_m(x^*)\}\,\mathrm{d}v - \mathcal{L}'\{\lambda_m(x)\}\mathcal{L}'\{\lambda_m(x^*)\},$$

where $\hat{\boldsymbol{\rho}}$ is defined by replacing λ with $\hat{\boldsymbol{\lambda}}$. The Delta method then produces

$$n^r\{\hat{\boldsymbol{\rho}}(x, x^*) - \rho(x, x^*)\} = \mathbb{D}_\rho(\lambda)n^r(\hat{\boldsymbol{\lambda}} - \lambda) + o_p(1),$$

where $\mathbb{D}_\rho$ is the Jacobian matrix of ρ.

9. Conclusions

We have shown in the paper that the problem of counterfactual prediction in an incomplete model has issues that are distinct from the familiar problems with identification and

estimation of model parameters. We have considered five approaches in the specific context of a complete information binary decision game with pure strategies and Nash equilibria: bounding the counterfactual prediction probabilities, using the midpoint prediction, taking a minmax approach with an entropy constraint, minimizing the average risk, and maximum entropy prediction. On balance, we prefer the maximum entropy approach since it does not have some of the unattractive features that other approaches have and is comparatively straightforward to implement.

Our results can be applied to other (counterfactual) prediction problems with partially identified parameters. It may further be of interest to study the relative merits of maximum entropy for the purpose of selecting a single point from the identified set in a partially identified world.

A. Proofs

Proof of Theorem 1. First,

$$q_v(y^* \mid x^*, x, y) = \frac{\mathbb{P}\big\{y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, x^*) = y^*, y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}^*, x) = y\big\}}{\tilde{\pi}_y(x)} =$$

$$\tilde{\pi}_{y^*}(x^*) + \frac{\pi_m(x)\pi_m(x^*)}{\tilde{\pi}_y(x)} \times$$

$$\mathrm{Cov}\big[\mathbb{1}\{y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, x^*) = y^*\}, \mathbb{1}\{y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}^*, x) = y\} \,\big|\, \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\big]. \quad (43)$$

The covariance in (43) equals zero unless both y and y^* belong to $\{(1,0),(0,1)\}$. For $y = y^* = (1,0)$ the covariance in (43) equals (by the law of iterated expectations)

$$\mathrm{Cov}\{\hbar(\boldsymbol{v}, x^*), \hbar(\boldsymbol{v}, x)\} = \rho(x, x^*).$$

The remaining three cases follow analogously. □

Proof of Theorem 2. Since $\hbar(v, x)$ is for all x nondecreasing in v, it follows that $\hbar(\boldsymbol{v}, x)$ and $\hbar(\boldsymbol{v}, x^*)$ have nonnegative covariance, which can be made to equal zero, e.g. by making $\hbar(v, x) = \mu_m(x)$ for all values of v. This establishes the lower bound.

For the upper bound, note that the covariance is maximized if $\hbar(v, x^*) = \hbar(v, x)\mu_m(x^*) \,/\, \mu_m(x)$, resulting in the stated upper bound. This upper bound is attained by setting $\hbar(v, x) = \mathbb{1}\{v > 1 - \mu_m(x)\}$. □

Proof of Theorem 3. We have

$$\begin{aligned}
&\mathbb{P}\big\{y(\boldsymbol{u}^*, \boldsymbol{e}, \boldsymbol{v}, x^*) = y^*, y(\boldsymbol{u}, \boldsymbol{e}, \boldsymbol{v}, x) = y\big\} \\
&\qquad = \mathbb{E}\big\{c_y(\boldsymbol{e}, x)c_{y^*}(\boldsymbol{e}, x^*)\big\} + \mathbb{E}\big\{c_y(\boldsymbol{e}, x)c_m(\boldsymbol{e}, x^*)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*)\big\} \\
&\quad + \mathbb{E}\big\{c_m(\boldsymbol{e}, x)c_{y^*}(\boldsymbol{e}, x^*)b_y(\boldsymbol{e}, \boldsymbol{v}, x)\big\} + \mathbb{E}\big\{c_m(\boldsymbol{e}, x)c_m(\boldsymbol{e}, x^*)b_y(\boldsymbol{e}, \boldsymbol{v}, x)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*)\big\} \\
&\qquad = \pi_{mm}\mathrm{Cov}\big\{b_y(\boldsymbol{e}, \boldsymbol{v}, x), b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*) \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big\} \\
&\qquad\qquad + \mathrm{Cov}\big\{a_y(\boldsymbol{e}, x), a_{y^*}(\boldsymbol{e}, x^*)\big\} + \mathbb{E}a_y(\boldsymbol{e}, x)\mathbb{E}a_{y^*}(\boldsymbol{e}, x^*). \quad (44)
\end{aligned}$$

The stated result follows if one divides the right hand side in (44) by $\tilde{\pi}_y(x) = \mathbb{E}a_y(\boldsymbol{e}, x)$. □

Proof of Theorem 4. Partition $S_m(x)$ into four disjoint regions: $S_A = S_m(x) \cap S_{10}(x^*)$, $S_B = S_m(x) \cap S_{01}(x^*)$, $S_C = S_m(x) \cap S_m(x^*)$, and $S_D = S_m(x) \cap \{S_{00}(x^*) \cup S_{11}(x^*)\}$. Let $\pi_A = \mathbb{P}(\boldsymbol{e} \in S_A)$ and let $\mu_A = \mathbb{E}\{b_y(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_A\}$. Let π_B, μ_C, etcetera, be similarly defined. We then solve

$$\min_{\mu_A,\mu_B,\mu_C\mu_D} (\pi_A\mu_A + \pi_B\mu_B + \pi_C\mu_C)$$
$$\text{subject to } \begin{cases} 0 \le \mu_A, \mu_B, \mu_C, \mu_D \le 1, \\ \pi_A\mu_A + \pi_B\mu_B + \pi_C\mu_C + \pi_D\mu_D = \phi_y(x), \end{cases}$$

to obtain the lower bound, noting that the solution has

$$\pi_A\mu_A + \pi_B\mu_B + \pi_C\mu_C = \max\{0, \phi_y(x) - \pi_D\},$$

because we are trying to make μ_D as large as possible. Maximizing the same objective function subject to the same constraints yields the upper bound. Since $\mu_A, \mu_B, \mu_C, \mu_D$ are otherwise unconstrained, the bounds are sharp. □

Proof of Theorem 5. When $x = x^*$, we have

$$\mathbb{P}\{y(\boldsymbol{e}, \boldsymbol{u}^*, \boldsymbol{v}, x^*) = y^*, y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, x) = y\} = \pi_{yy^*}(x, x)$$
$$+ \pi_m(x)\mathbb{E}\{\bar{b}_y(\boldsymbol{e}, x)\bar{b}_{y^*}(\boldsymbol{e}, x) \mid \boldsymbol{e} \in S_m(x)\} + \delta_y\delta_{y^*}\pi_m(x)\mathbb{E}\{\mathbb{V}(p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e}) \mid \boldsymbol{e} \in S_m(x)\},$$

where $\bar{b}_y(e, x) = \mathbb{E}b_y(e, \boldsymbol{v}, x)$. The result follows from the fact that

$$0 \le \mathbb{V}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e}\} \le \mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e}\}\big[1 - \mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e}\}\big],$$

where the upper bound is attained when $p(e, \boldsymbol{v}, x)$ is binary taking one with probability $\mathbb{E}p(e, \boldsymbol{v}, x)$, and the lower bound is attained when $p(e, \boldsymbol{v}, x) = \mu_m(x)$ with probability one.

So, we focus on the case $x \neq x^*$. Some tedious but simple mathematical manipulations show that $\tilde{\pi}_y(x)q_{ev}(y^* \mid x^*, x, y)$ can alternatively be expressed as

$$\mathbb{E}\Big[\{c_y(\boldsymbol{e}, x) + c_m(\boldsymbol{e}, x)b_y(\boldsymbol{e}, \boldsymbol{v}, x)\}\{c_{y^*}(\boldsymbol{e}, x^*) + c_m(\boldsymbol{e}, x^*)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*)\}\Big]$$
$$= \pi_{yy^*}(x, x^*) + \pi_{my^*}(x, x^*)\mathbb{E}\{\bar{b}_y(\boldsymbol{e}, x) \mid \boldsymbol{e} \in S_{my^*}(x, x^*)\}$$
$$+ \pi_{ym}(x, x^*)\mathbb{E}\{\bar{b}_{y^*}(\boldsymbol{e}, x^*) \mid \boldsymbol{e} \in S_{ym}(x, x^*)\}$$
$$+ \pi_{mm}(x, x^*)\mathbb{E}\{b_y(\boldsymbol{e}, \boldsymbol{v}, x)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*) \mid \boldsymbol{e} \in S_{mm}(x, x^*)\}. \quad (45)$$

Note that $b_y(e, v, x)$ is unrestricted in $e \in S_m(x)$ but is restricted (i.e. monotonic) in v, albeit that we have the condition $\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_m(x)\} = \mu_m(x)$ and $b_y(e, v, x)$ must belong to $[0, 1]$. Therefore, the bounds we seek can be attained for functions p that are flat in v, which means that there is no loss of generality in dropping v from the notation, which we do from hereon.

Now, let $S_r^y(x) = [0, 1]^2 - S_m(x) - S_y(x)$. Further, let $S_{mr^*}(x, x^*) = S_m(x) \cap S_r^{y^*}(x^*)$ and $S_{rm}(x, x^*) = S_r^y(x) \cap S_m(x^*)$. We then define

$$\bar{z}_{my^*}^{xy}(x, x^*) = \mathbb{E}\{b_y(\boldsymbol{e}, x) \mid \boldsymbol{e} \in S_{my^*}(x, x^*)\},$$

and let $\mathcal{Z}$'s with different subscript and superscript combinations be defined analogously. Then the right hand side in (45) reduces to

$$\pi_{yy^*}(x,x^*) + \pi_{my^*}(x,x^*)\mathcal{Z}^{xy}_{my^*}(x,x^*) + \pi_{ym}(x,x^*)\mathcal{Z}^{x^*y^*}_{ym}(x,x^*) \\ + \pi_{mm}(x,x^*)\mathbb{E}\big\{\mathcal{b}_y(\boldsymbol{e},x)\mathcal{b}_{y^*}(\boldsymbol{e},x^*) \,\big|\, \boldsymbol{e} \in S_{mm}(x,x^*)\big\}. \quad (46)$$

Therefore, we seek to minimize/maximize (46) subject to the (mean) restrictions, namely (dropping the (x,x^*) arguments)

$$\begin{cases} \pi_{ym}\mathcal{Z}^{x^*y^*}_{ym} + \pi_{mm}\mathcal{Z}^{x^*y^*}_{mm} + \pi_{rm}\mathcal{Z}^{x^*y^*}_{rm} = \phi_{y^*}(x^*), \\ \pi_{my^*}\mathcal{Z}^{xy}_{my^*} + \pi_{mm}\mathcal{Z}^{xy}_{mm} + \pi_{mr^*}\mathcal{Z}^{xy}_{mr^*} = \phi_y(x). \end{cases} \quad (47)$$

Since $\mathcal{b}_y(e,x)$ must belong to $[0,1]$, (46) implies that for the lower bound we want to strive to make $\mathcal{b}_y(e,x)\mathcal{b}_{y^*}(e,x^*)$ equal to zero for $e \in S_{mm}(x,x^*)$ and equal to one for the upper bound. This implies that there is no loss of generality in assuming that $\mathcal{b}_y(e,x)$ is binary for all e,x,y. So, we assume that $\mathcal{b}_y(e,x)$ is either one or zero hereafter. Define

$$S^y_+(x) = \{e \in S_m(x) :\ \mathcal{b}_y(e,x) = 1\} \quad \text{and} \quad S^y_-(x) = \{e \in S_m(x) :\ \mathcal{b}_y(e,x) = 0\},$$

which forms a partition of $S_m(x)$.

The optimization of (46) subject to (47) is illustrated in figure 16: minimization at the top and maximization at the bottom.

Consider the lower bound first. Define a,b,c,d to be the probabilities illustrated at the top of figure 16 so that $\mathbb{P}(\boldsymbol{y} = y, \boldsymbol{y}^* = y^* | \boldsymbol{x} = x, \boldsymbol{x}^* = x^*) = a + b + c + c$. Then, note that a,b,c,d correspond to the four terms in (46), (e.g. $\pi_{my^*}\mathcal{Z}^{xy}_{my^*} = b$), because by construction $\mathcal{b}_y(e,x)$ is equal to one when $e \in S^y_+(x)$ and equal to zero, otherwise. The constraints in (47) can likewise be expressed in terms of the probabilities illustrated at the top of figure 16. So, using the shorthand $\phi = \phi_y(x)$ and $\phi^* = \phi_{y^*}(x^*)$, minimizing (46) subject to (47) can be formulated as

$$\begin{cases} \min\limits_{a,b,c,d} (a+b+c+d) \\ \quad \text{s.t. } c + d + f \geq \max(\phi^* - \pi_{rm}, 0) \\ \qquad b + d + e \geq \max(\phi - \pi_{mr^*}, 0) \\ \qquad d + e + f \leq \pi_{mm} \\ \qquad a = \pi_{yy^*} \\ \qquad a,b,c,d,e,f \geq 0 \end{cases}$$

Now,

$$\begin{aligned} b + c + d &= (b + d + e) + (c + d + f) - (d + e + f) \\ &\geq \max\big\{\max(\phi^* - \pi_{rm}, 0) + \max(\phi - \pi_{mr^*}, 0) - \pi_{mm}, 0\big\} \\ &= \max\big(\phi^* - \pi_{rm} - \pi_{mm}, \phi - \pi_{mr^*} - \pi_{mm}, \phi + \phi^* - \pi_{mr^*} - \pi_{rm} - \pi_{mm}, 0\big). \end{aligned}$$

Finding the upper bound can likewise be formulated as a maximization problem. Define a, b, c, d as the probabilities illustrated at the bottom of figure 16. Then, maximizing (46) subject to (47) can be formulated as

$$\begin{cases} \max\limits_{a,b,c,d} (a+b+c+d) \\ \quad \text{s.t. } b \le \min(\phi - d, \pi_{ym}) \\ \qquad c \le \min(\phi^* - d, \pi_{my^*}) \\ \qquad 0 \le d \le \pi_{mm}, \\ \qquad a = \pi_{yy^*}. \end{cases}$$

Now,

$$\begin{aligned} b + c + d &\le \min(\phi - d, \pi_{my^*}) + \min(\phi^* - d, \pi_{ym}) + d \\ &\le \min(\phi + \phi^*, \phi + \pi_{ym}, \phi^* + \pi_{my^*}, \pi_{my^*} + \pi_{ym} + \pi_{mm}). \end{aligned}$$

Since the proof is constructive, the bounds are sharp. □

Proof of Theorem 6. This is a Lagrangean optimization problem with a function–valued parameter. The first order conditions are

$$\begin{cases} f_e(e) f_x(x) \{1 + \log f^*(p \mid e, x)\} - \nu_1(e,x) - \nu_2(x) p f_e(e) \mathbb{1}\{e \in S_m(x)\} = 0, \\ \qquad\qquad \int f^*(p \mid e, x) \, \mathrm{d}p = 1, \\ \int_{S_m(x)} \int_0^1 p f^*(p \mid e, x) \, \mathrm{d}p f_e(e) \, \mathrm{d}e = \mu_m(x) \pi_m(x), \end{cases}$$

almost everywhere, where ν_1, ν_2 are Lagrangean parameters. Thus, taking $\lambda(x) = \nu_2(x) \,/\, f_x(x)$, it follows that for all $p \in [0,1]$,

$$f^*(p \mid e, x) \propto \begin{cases} \exp\{\lambda(x) p\}, & e \in S_m(x), \\ 1, & e \notin S_m(x), \end{cases}$$

which yields (19).

Further, the second condition in (20) imposes that f^* integrate to one and the first that f^* has the correct mean since $\mathcal{L}'(\lambda) = I'(\lambda) \,/\, I(\lambda) = \int_0^1 p \exp(p\lambda) \, \mathrm{d}p \,\big/\, \int_0^1 \exp(p\lambda) \, \mathrm{d}p$. □

Proof of Theorem 7. From (19) it follows that

$$f^*(p \mid e, x) = \begin{cases} 1, & e \notin S_m(x), \\ A\{p, \lambda_m(x)\}, & e \in S_m(x). \end{cases}$$

Inverting the conditional distribution function $\int_0^p f^*(s \mid e, x) \, \mathrm{d}s$ yields the stated result. □

Lemma 1. $\mathcal{L}'$ is continuous, increasing, and nonzero everywhere.

Proof. Continuity is trivial by L'Hôpital. Further,

$$\mathscr{L}''(\lambda) = \begin{cases} 1/12, & \text{if } \lambda = 0, \\ 1/\lambda^2 - \exp(\lambda)/\{\exp(\lambda) - 1\}^2, & \text{if } \lambda \neq 0, \end{cases}$$

which is continuous by L'Hôpital, also. Indeed, for $\lambda \neq 0$,

$$\begin{aligned} \lambda^2\{\exp(\lambda) - 1\}^2 \mathscr{L}''(\lambda) &= \exp(2\lambda) - (2 + \lambda^2)\exp(\lambda) + 1 \\ &= \exp(\lambda)\{\exp(\lambda) + \exp(-\lambda) - 2 - \lambda^2\} = 2\exp(\lambda)\sum_{j=2}^{\infty} \frac{\lambda^{2j}}{(2j)!} > 0. \end{aligned}$$

Finally, for the last assertion, simply note that $\lim_{\lambda \to -\infty} \mathscr{L}'(\lambda) = 0$ and that $\mathscr{L}'$ is increasing. □

Proof of Theorem 9. By lemma 1, we know that $\mathscr{L}'^{-1}$ is differentiable everywhere. So,

$$n^r\{\hat{\boldsymbol{\lambda}}_m(\tilde{x}) - \lambda_m(\tilde{x})\} = \frac{n^r\{\hat{\boldsymbol{\mu}}_m(\tilde{x}) - \mu_m(\tilde{x})\}}{\mathscr{L}'\big[\mathscr{L}'^{-1}\{\tilde{\boldsymbol{\mu}}_m(\tilde{x})\}\big]}, \qquad \tilde{x} = x, x^*,$$

where $\tilde{\boldsymbol{\mu}}_m(\tilde{x})$ is between $\hat{\boldsymbol{\mu}}_m(\tilde{x})$ and $\mu_m(\tilde{x})$. Then, the result follows from lemma 1 and assumption A. □

B. Other cases

Below, we present some results on prediction probabilities other than q, q_{ev}, q_v. The proofs refer back to details in earlier proofs, especially that of theorem 5.

Theorem 10. $q_e(y^*|x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \alpha_{yy^*}(x, x^*)$. Further, the bounds on $q_e(y^*|x^*, x, y)$ are identical to those on $q_{ev}(y^* \mid x^*, x, y)$ given in theorem 5.

Proof. Recall that $\bar{b}_y(e, x) = \mathbb{E}\, b_y(e, \boldsymbol{v}, x)$. Since $b_y(\boldsymbol{e}, \boldsymbol{v}, x)$ and $b_{y^*}(\boldsymbol{e}, \boldsymbol{v}^*, x^*)$ are independent conditional on $\boldsymbol{e}$, we have

$$\begin{aligned} &\mathbb{P}\{y(\boldsymbol{e}, \boldsymbol{u}^*, \boldsymbol{v}^*, x^*) = y^*, y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, x) = y\} \\ &\quad = \mathbb{E}\{c_y(\boldsymbol{e}, x)c_{y^*}(\boldsymbol{e}, x^*)\} + \mathbb{E}\{c_y(\boldsymbol{e}, x)c_m(\boldsymbol{e}, x^*)\bar{b}_{y^*}(\boldsymbol{e}, x^*)\} \\ &\qquad + \mathbb{E}\{c_m(\boldsymbol{e}, x)c_{y^*}(\boldsymbol{e}, x^*)\bar{b}_y(\boldsymbol{e}, x)\} + \mathbb{E}\{c_m(\boldsymbol{e}, x)c_m(\boldsymbol{e}, x^*)\bar{b}_y(\boldsymbol{e}, x)\bar{b}_{y^*}(\boldsymbol{e}, x^*)\} \\ &\qquad\quad = \tilde{\pi}_y(x)\tilde{\pi}_{y^*}(x^*) + \mathrm{Cov}\{a_y(\boldsymbol{e}, x), a_{y^*}(\boldsymbol{e}, x^*)\} \\ &\qquad\quad = \pi_{yy^*}(x, x^*) + \pi_{ym}(x, x^*)\mathbb{E}\{\bar{b}_{y^*}(\boldsymbol{e}, x^*) \,\big|\, \boldsymbol{e} \in S_{ym}(x, x^*)\} \\ &\qquad\qquad + \pi_{my^*}(x, x^*)\mathbb{E}\{\bar{b}_y(\boldsymbol{e}, x) \,\big|\, \boldsymbol{e} \in S_{my^*}(x, x^*)\} \\ &\qquad\qquad\quad + \pi_{mm}(x, x^*)\mathbb{E}\{\bar{b}_y(\boldsymbol{e}, x)\bar{b}_{y^*}(\boldsymbol{e}, x^*) \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\}, \end{aligned}$$

which coincides with (45), except that b in the last term of (45) is replaced with $\bar{b}$ here. Since the bounds in theorem 5 obtain for functions p that are flat in v, as noted in the proof of theorem 5, the bounds must coincide. □

Theorem 11.

$$q_u(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \delta_y \delta_{y^*} \frac{\pi_m(x)\pi_m(x^*)}{\tilde{\pi}_y(x)} \rho_u(x, x^*),$$

where $\rho_u(x, x^*) = \mathbb{E}\big\{\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}^*, \boldsymbol{v}^*, x^*)\} \big| \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\big\} - \mu_m(x)\mu_m(x^*)$. Further, the bounds on q_u are identical to the bounds of q_v given in theorem 2.

Proof. The expression for q_u follows from simple algebra. For the bounds, we focus on the case $y = y^* = (1, 0)$: the other cases are analogous. Since $\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \bigm| \cdot\big\} \leq \min\big\{\mathbb{E}(\boldsymbol{p} \mid \cdot), \mathbb{E}(\boldsymbol{p}^* \mid \cdot)\big\}$, the upper bound of $\rho_u(x, x^*)$ is given by $\min\big\{\mu_m(x), \mu_m(x^*)\big\} - \mu_m(x)\mu_m(x^*)$. This upper bound is attained when $\boldsymbol{p}$ and $\boldsymbol{p}^*$ are binary. For the lower bound, note that $\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \bigm| \cdot\big\} \geq \mathbb{E}(\boldsymbol{p}\boldsymbol{p}^* \mid \cdot)$, where $\boldsymbol{p}, \boldsymbol{p}^*$ are independent given $\boldsymbol{e}, \boldsymbol{e}^*$. □

Theorem 12.

$$q_{uv}(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \delta_y \delta_{y^*} \frac{\pi_m(x)\pi_m(x^*)}{\tilde{\pi}_y(x)} \rho_{uv}(x, x^*), \tag{48}$$

where $\rho_{uv}(x, x^*) = \mathbb{E}\big[\min\big\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}^*, \boldsymbol{v}, x^*)\big\} \big| \boldsymbol{e} \in S_m(x),\ \boldsymbol{e}^* \in S_m(x^*)\big] - \mu_m(x)\mu_m(x^*)$. Further, the bounds of q_{uv} are identical to those on q_v given in theorem 2.

Proof. Define

$$\hbar_y(e, u, v, x) = \begin{cases} \mathbb{1}\big\{u \leq p(e, v, x)\big\}, & y = (1, 0), \\ \mathbb{1}\big\{u > p(e, v, x)\big\}, & y = (0, 1), \\ 0, & y \in \big\{(0, 0), (1, 1)\big\}. \end{cases}$$

Note that

$$\begin{aligned} &\mathbb{P}\big\{y(\boldsymbol{e}^*, \boldsymbol{u}, \boldsymbol{v}, x^*) = y^*, y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, x) = y\big\} \\ &\quad = \tilde{\pi}_y(x)\tilde{\pi}_{y^*}(x, x^*) + \\ &\qquad \pi_m(x)\pi_m(x^*)\mathrm{Cov}\big\{\hbar_y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, x), \hbar_{y^*}(\boldsymbol{e}^*, \boldsymbol{u}, \boldsymbol{v}, x^*) \bigm| \boldsymbol{e} \in S_m(x),\ \boldsymbol{e}^* \in S_m(x^*)\big\}. \end{aligned} \tag{49}$$

Tedious but simple algebra shows that the covariance in (49) equals $\delta_y \delta_{y^*} \rho_{uv}(x, x^*)$, which yields (48).

Now we establish the bounds. We focus on the case $y = y^* = (1, 0)$; the other cases follow analogously. For the upper bound, note that $\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \bigm| \cdot\big\} \leq \min\big\{\mathbb{E}\big(\boldsymbol{p} \bigm| \cdot\big), \mathbb{E}\big(\boldsymbol{p}^* \bigm| \cdot\big)\big\}$, with equality for $p(e, v, x) = \mathbb{1}\big\{v \geq 1 - \mu_m(x)\big\}$. For the lower bound,

$$\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \bigm| \boldsymbol{e}, \boldsymbol{e}^*\big\} \geq \mathbb{E}(\boldsymbol{p}\boldsymbol{p}^* \mid \boldsymbol{e}, \boldsymbol{e}^*) \geq \mathbb{E}(\boldsymbol{p} \mid \boldsymbol{e})\mathbb{E}(\boldsymbol{p}^* \mid \boldsymbol{e}^*) \tag{50}$$

where the inequalities hold with equality for $p(e, v, x) = \mathbb{1}\{e \in S_m(x)\}$. Take the expectations in (50) over $S_m(x)$ and $S_m(x^*)$ to obtain the sharp lower bound. □

For the remaining two cases (q_{eu} and q_{euv}), we define

$$g_y(e, v, x) = c_y(e, x) + c_m(e, x)\hbar_y(e, v, x).$$

Theorem 13.

$$q_{eu}(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \frac{\mathrm{Cov}\{g_y(\boldsymbol{e}, \boldsymbol{v}, x), g_{y^*}(\boldsymbol{e}, \boldsymbol{v}^*, x^*)\}}{\tilde{\pi}_y(x)} + \delta_y \delta_{y^*} \frac{\pi_{mm}(x, x^*)}{\tilde{\pi}_y(x)} \rho_{eu}(x, x^*),$$

where $\rho_{eu}(x, x^*) = \mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}^*, x^*)\} - p(\boldsymbol{e}, \boldsymbol{v}, x)p(\boldsymbol{e}, \boldsymbol{v}^*, x^*) \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big]$. Further, the bounds on $q_{eu}(y^* \mid x^*, x, y)$ are identical to those of $q_{ev}(y^* \mid x^*, x, y)$ given in theorem 5.

Proof. The formula for q_{eu} follows from tedious algebra, so we focus on the bounds. We consider the case $y = y^*(1, 0)$: the other combinations follow analogously. Now, $\tilde{\pi}_y(x) q_{eu}(y^* \mid x^*, x, y)$ equals

$$\begin{aligned} \pi_{yy^*}(x, x^*) + \pi_{ym}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}^*, x^*) \mid \boldsymbol{e} \in S_{ym}(x, x^*)\} \\ + \pi_{my^*}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_{my^*}(x, x^*)\} \\ + \pi_{mm}(x, x^*)\mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}^*, x^*)\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big]. \quad (51) \end{aligned}$$

If $x = x^*$, the second and third terms in (51) are equal to zero, so we only need to consider the last term. The upper bound follows from the fact that

$$\mathbb{E}\big\{\min\{\boldsymbol{p}, \boldsymbol{p}^*\} \,\big|\, \cdot\big\} \leq \min\big\{\mathbb{E}(\boldsymbol{p} \mid \cdot), \mathbb{E}(\boldsymbol{p}^* \mid \cdot)\big\}. \quad (52)$$

For the lower bound, note that by the Jensen inequality

$$\begin{aligned} \mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \,\big|\, \boldsymbol{e} \in S_m(x)\big\} &= \mathbb{E}\left[\int_0^1 \mathbb{P}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) > t \,\big|\, \boldsymbol{e}\big\}\, \mathrm{d}t \,\middle|\, \boldsymbol{e} \in S_m(x)\right] \\ &\geq \left[\mathbb{E}\left\{\int_0^1 \mathbb{P}(\boldsymbol{p} > t \mid \boldsymbol{e})\, \mathrm{d}t \,\middle|\, \boldsymbol{e} \in S_m(x)\right\}\right]^2 = \mu_m^2(x). \end{aligned}$$

If $x \neq x^*$ then (52) and $\mathbb{E}\big\{p(\boldsymbol{e}, \boldsymbol{v}^*, x^*) \big| \boldsymbol{e} \in S_m(x^*)\big\} = \mathbb{E}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x^*) \big| \boldsymbol{e} \in S_m(x^*)\big\} = \mu_m(x^*)$ imply that (51) is bounded above by $\pi_{yy^*}(x, x^*) + U_{yy^*}(x, x^*)$, where U_{yy^*} is as defined in (15). Sharpness follows from the fact that (51) is bounded below by (45). For the lower bound, note that $\mathbb{E}\big\{\min\{\boldsymbol{p}, \boldsymbol{p}^*\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big\} \geq \mathbb{E}\big\{\bar{b}_y(\boldsymbol{e}, x)\bar{b}_{y^*}(\boldsymbol{e}, x^*) \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big\}$, which is identical to the second factor of the last term in (45), except that b_y is replaced with $\bar{b}_y$. □

Theorem 14.

(*i*) If $x^* = x$ then $q_{euv}(y^* \mid x^*, x, y) = \mathbb{1}(y^* = y)$;

(*ii*) If $x^* \neq x$ then

$$\begin{aligned} q_{euv}(y^* \mid x^*, x, y) = \\ \tilde{\pi}_{y^*}(x^*) + \frac{\mathrm{Cov}\{g_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*), g_y(\boldsymbol{e}, \boldsymbol{v}, x)\}}{\tilde{\pi}_y(x)} + \delta_y \delta_{y^*} \frac{\pi_{mm}(x, x^*)}{\tilde{\pi}_y(x)} \rho_{euv}(x, x^*), \end{aligned}$$

where $\rho_{euv}(x, x^*) = \mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}, x^*)\} - p(\boldsymbol{e}, \boldsymbol{v}, x)p(\boldsymbol{e}, \boldsymbol{v}, x^*) \big| \boldsymbol{e} \in S_{mm}(x, x^*)\big]$;

(*iii*) If $x^* \neq x$ then the bounds on q_{euv} coincide with those on q_{ev}.

Proof. Part (i) is trivial and part (ii) tedious yet mechanical, so we focus on part (iii). We establish the result for the case $y = y^* = (1, 0)$, where the other combinations follow analogously.

Now, note from part (ii) that $\tilde{\pi}_y(x) q_{euv}(y^* \mid x^*, x, y)$ equals

$$\begin{aligned}\pi_{yy^*}(x, x^*) &+ \pi_{ym}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x^*) \mid \boldsymbol{e} \in S_{ym}(x, x^*)\} \\ &+ \pi_{my^*}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_{my^*}(x, x^*)\} \\ &+ \pi_{mm}(x, x^*)\mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}, x^*)\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big]. \quad (53)\end{aligned}$$

Because $\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \,\big|\, \cdot\big\} \leq \min\big\{\mathbb{E}(\boldsymbol{p} \mid \cdot), \mathbb{E}(\boldsymbol{p}^* \mid \cdot)\big\}$, $\mathbb{E}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x) \,\big|\, \boldsymbol{e} \in S_m(x)\big\} = \mu_m(x)$, and $\mathbb{E}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x^*) \,\big|\, \boldsymbol{e} \in S_m(x^*)\big\} = \mu_m(x^*)$, (53) is bounded above by $\pi_{yy^*}(x, x^*) + U_{yy^*}(x, x^*)$, where U_{yy^*} is as defined in (15). For sharpness, note that (53) is bounded below by (45).

Now the lower bound. Since (53) is no less than (45), the lower bound for (45) is a lower bound here, also. It remains to be shown that it can be attained. Recall from the proof of theorem 5 that (45) attains the lower bound $\pi_{yy^*}(x, x^*) + L_{yy^*}(x, x^*)$ when $\boldsymbol{p}$ and $\boldsymbol{p}^*$ are binary, in which case $\boldsymbol{p}\boldsymbol{p}^* = \min(\boldsymbol{p}, \boldsymbol{p}^*)$. Hence, (45) equals (53). □

C. Useful formulas

In this appendix we provide the formulas for the correction terms in theorems 1 and 3 that are implied by the maximum entropy solution f^*. As we explained below theorem 7, the quantity $\rho(x, x^*)$ entering the correction terms is for the maximum entropy solution the same for q_v and q_{ev}: i.e.

$$\begin{aligned}\rho^*(x, x^*) &= \mathrm{Cov}\big\{\hbar^*(\boldsymbol{v}, x)\hbar^*(\boldsymbol{v}, x^*)\big\} \\ &= \mathbb{E}\big[\mathrm{Cov}\{p^*(\boldsymbol{e}, \boldsymbol{v}, x)p^*(\boldsymbol{e}, \boldsymbol{v}, x^*) \mid \boldsymbol{e}\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big] = \\ &\begin{cases} \displaystyle\int_0^1 \frac{\log\big(1 + v[\exp\{\lambda_m(x)\} - 1]\big)\log\big(1 + v[\exp\{\lambda_m(x^*)\} - 1]\big)}{\lambda_m(x)\lambda_m(x^*)}\,\mathrm{d}v \\ \qquad\qquad -\mu_m(x)\mu_m(x^*), & \lambda_m(x), \lambda_m(x^*) \neq 0, \\ \dfrac{1}{4\lambda_m(x)} + \dfrac{1}{2\lambda_m(x)\{\mathrm{e}^{\lambda_m(x)} - 1\}} - \dfrac{\mathrm{e}^{\lambda_m(x)}}{2\{e^{\lambda_m(x)-1}\}^2}, & \lambda_m(x) \neq \lambda_m(x^*) = 0, \\ \dfrac{1}{4\lambda_m(x^*)} + \dfrac{1}{2\lambda_m(x^*)\{\mathrm{e}^{\lambda_m(x^*)} - 1\}} - \dfrac{\mathrm{e}^{\lambda_m(x^*)}}{2\{e^{\lambda_m(x^*)-1}\}^2}, & \lambda_m(x^*) \neq \lambda_m(x) = 0, \\ \dfrac{1}{12}, & \lambda_m(x) = \lambda_m(x^*) = 0. \end{cases}\end{aligned}$$

Let a_y^* be the function a_y defined in theorem 3 corresponding to f^*. Then,

$$a_y^*(e, x) = c_y(e, x) + c_m(e, x)\mathbb{E}b_y^*(e, \boldsymbol{v}, x)$$

$$= \begin{cases} c_y(e,x) + c_m(e,x)\mu_m(x), & y = (1,0), \\ c_y(e,x) + c_m(e,x)\{1-\mu_m(x)\}, & y = (0,1), \\ c_y(e,x), & y \in \{(0,0),(1,1)\}, \end{cases} \tag{54}$$

from which the formula for $\mathrm{Cov}\{a^*_y(\boldsymbol{e},x), a^*_{y^*}(\boldsymbol{e},x^*)\}$ follows.

D. Solving the minmax problem

For some constants $c_{\sigma 1}, c_{\sigma 2}$, let $\kappa_\sigma(v) = \epsilon_\sigma^{-1}(c_{\sigma 1} v + c_{\sigma 2})$, $\kappa_{\sigma 0} = \kappa_\sigma(0)$, $\kappa_{\sigma 1} = \kappa_\sigma(1)$. By construction $c_{\sigma 2} = \epsilon_\sigma(\kappa_{\sigma 0})$ and $c_{\sigma 1} = \epsilon_\sigma(\kappa_{\sigma 1}) - \epsilon_\sigma(\kappa_{\sigma 0})$. For $j = 0, 1, 2$, define $I_{\sigma j} = \int_0^1 \kappa_\sigma^j(t)\,\mathrm{d}t$. By substitution, $I_{\sigma j} = \int_{\kappa_{\sigma 0}}^{\kappa_{\sigma 1}} t^j \exp(-\sigma t^2)\,\mathrm{d}t / c_{\sigma 1}$.

D.1 Algebraic lemmas:

Lemma 2. $\kappa'_\sigma(v) = c_{\sigma 1} \exp\{\sigma \kappa_\sigma^2(v)\}$.

Proof. Trivial. □

Lemma 3. $\forall v \in [0,1]$, $\int_v^1 \kappa_\sigma(t)\,\mathrm{d}t = \left[\exp\{-\sigma\kappa_\sigma^2(v)\} - \exp\{-\sigma\kappa_\sigma^2(1)\}\right]/(2\sigma c_{\sigma 1})$.

Proof. Substitute $s = \kappa_\sigma(t)$ and integrate. □

D.2 Proof of Theorem 8: Since e is irrelevant, we drop it from the notation. Further, use the shorthand $p = p(\cdot, x)$ and $p^* = p(\cdot, x^*)$. Similarly, we will write μ_m, μ_m for $\mu_m(x)$ and $\mu_m(x^*)$, respectively. So, we consider

$$\max_{p} \operatorname*{or}_{p^*} \min \int_0^1 p(v)p^*(v)\,\mathrm{d}v - \mu_m \mu_m^* \quad \text{subject to} \quad \begin{cases} \displaystyle\int_0^1 p(v)\,\mathrm{d}v = \mu_m, \\ \displaystyle\int_0^1 p^*(v)\,\mathrm{d}v = \mu_m^*, \\ \displaystyle\int_0^1 \log a(v)\,\mathrm{d}v + \int_0^1 \log a^*(v)\,\mathrm{d}v \geq \mathcal{E}^*, \\ p(1), p^*(1) \leq 1, \\ p(0), p^*(0) \geq 0, \end{cases} \tag{55}$$

where a and a^* denote $\partial_v p$ and $\partial_v p^*$, respectively: $a(v) > 0$ and $a^*(v) > 0$ are implicit by the entropy inequality constraint. Noting that $p(v) = \int_0^v a(t)\,\mathrm{d}t$, we have

$$\int_0^1 p(v)p^*(v)\,\mathrm{d}v = \int_0^1 \int_0^v p^*(v)a(t)\,\mathrm{d}t\,\mathrm{d}v = \int_0^1 a(v) \int_v^1 p^*(t)\,\mathrm{d}t\,\mathrm{d}v,$$

$$\int_0^1 p(v)\,\mathrm{d}v = \int_0^1 \int_0^v a(t)\,\mathrm{d}t\,\mathrm{d}v = \int_0^1 \int_t^1 a(t)\,\mathrm{d}v\,\mathrm{d}t = \int_0^1 a(v)(1-v)\,\mathrm{d}v,$$

etcetera. Using this formulation, we solve (55) with respect to a and a^*, from which we obtain the first order conditions

$$\forall v : \begin{cases} \displaystyle\int_v^1 p^*(v)\,\mathrm{d}v + \frac{\lambda}{a(v)} - \psi(1-v) - \omega = 0, \\ \displaystyle\int_v^1 p(v)\,\mathrm{d}v + \frac{\lambda}{a^*(v)} - \psi^*(1-v) - \omega^* = 0, \end{cases} \tag{56}$$

where λ, ψ, ω are Lagrange multipliers for the entropy constraint, $\int_0^1 p(v)\,\mathrm{d}v = \mu_m$, and $\int_0^1 a(v)\,\mathrm{d}v \le 1$, respectively: ψ^* and ω^* are the corresponding multipliers for a^*. The conditions in (56) are in addition to the constraints in (55). Note that in the maximization problem, $\lambda, \omega, \omega^* \ge 0$, and in the minimization problem $\lambda, \omega, \omega^* \le 0$.

Differentiating (56) yields differential equations, from which we deduce that there are solutions of the form

$$\begin{cases} p_\sigma(v) = \mu_m + b_\sigma\{\kappa_\sigma(v) - I_{\sigma 1}\}, \\ p^*_\sigma(v) = \mu^*_m + b^*_\sigma\{\kappa_\sigma(v) - I_{\sigma 1}\}. \end{cases} \tag{57}$$

Indeed, with (57), the multipliers satisfy

$$\lambda = -\frac{b_\sigma b^*_\sigma}{2\sigma}, \quad \psi = \mu^*_m - b^*_\sigma I_{\sigma 1}, \quad \omega = -\frac{b^*_\sigma \exp\bigl(-\sigma\kappa^2_{\sigma 1}\bigr)}{2\sigma c_{\sigma 1}},$$
$$\psi^* = \mu_m - b_\sigma I_{\sigma 1}, \quad \omega^* = -\frac{b_\sigma \exp\bigl(-\sigma\kappa^2_{\sigma 0}\bigr)}{2\sigma c_{\sigma 1}},$$

where we note that $\sigma = \pm 1$ determines the signs of $\lambda, \omega, \omega^*$ in accordance with the Kuhn Tucker requirements. So, all that remains to be done is to determine the constants $b_1, b_{-1}, c_{11}, c_{-11}, c_{12}, c_{-12}$. Below we discuss how to determine $c_{\sigma 1}, c_{\sigma 2}$ or equivalently, $\kappa_{\sigma 0}, \kappa_{\sigma 1}$.

Now, note that if $\lambda = 0$, then either $\forall v : a(v) = 0$ or $\forall v : a^*(v) = 0$, which would violate the entropy inequality constraint since $\mathcal{E}^*$ was assumed to be finite. Since the entropy inequality constraint must be binding, we have

$$\log(b_\sigma b^*_\sigma) + 2\log c_{\sigma 1} + 2\sigma I_{\sigma 2} = \mathcal{E}^*. \tag{58}$$

Further, $\lambda \neq 0$ implies that $\omega \neq 0$ and $\omega^* \neq 0$. Therefore, the boundary conditions at $v = 1$ must be binding, also:

$$p_\sigma(1) = \mu_m + b_\sigma(\kappa_{\sigma 1} - I_{\sigma 1}) = 1 \quad \text{and} \quad p^*_\sigma(1) = \mu^*_m + b^*_\sigma(\kappa_{\sigma 1} - I_{\sigma 1}) = 1. \tag{59}$$

By combining (58) and (59), we obtain a relationship between $\kappa_{\sigma 0}, \kappa_{\sigma 1}$: i.e.

$$\log(1-\mu_m)(1-\mu^*_m) + 2\log\bigl\{c_{\sigma 1}/(\kappa_{\sigma 1} - I_{\sigma 1})\bigr\} + 2\sigma I_{\sigma 2} = \mathcal{E}^*. \tag{60}$$

Note also that $\kappa_{\sigma 0}$ and $\kappa_{\sigma 1}$ must satisfy the boundary condition at $v = 0$. So,

$$\begin{cases} p_\sigma(0) = \mu_m - (1-\mu_m)(I_{\sigma 1} - \kappa_{\sigma 0})/(\kappa_{\sigma 1} - I_{\sigma 1}) \geq 0, \\ p^*_\sigma(0) = \mu^*_m - (1-\mu^*_m)(I_{\sigma 1} - \kappa_{\sigma 0})/(\kappa_{\sigma 1} - I_{\sigma 1}) \geq 0. \end{cases} \tag{61}$$

Finally, the objective function in (55) that corresponds to p_σ, p^*_σ can be written as

$$\int_0^1 p_\sigma(v) p^*_\sigma(v)\,\mathrm{d}v - \mu_m \mu^*_m = b_\sigma b^*_\sigma (I_{\sigma 2} - I^2_{\sigma 1}) = (I_{\sigma 2} - I^2_{\sigma 1}) \exp\big(\mathcal{E}^* - 2\sigma I_{\sigma 2}\big)/c^2_{\sigma 1}.$$

Therefore, we can determine $\kappa_{\sigma 0}, \kappa_{\sigma 1}$ by maximizing ($\sigma = -1$) or minimizing ($\sigma = 1$)

$$(I_{\sigma 2} - I^2_{\sigma 1}) \exp\big(\mathcal{E}^* - 2\sigma I_{\sigma 2}\big)/c^2_{\sigma 1}$$

subject to (60) and (61). □

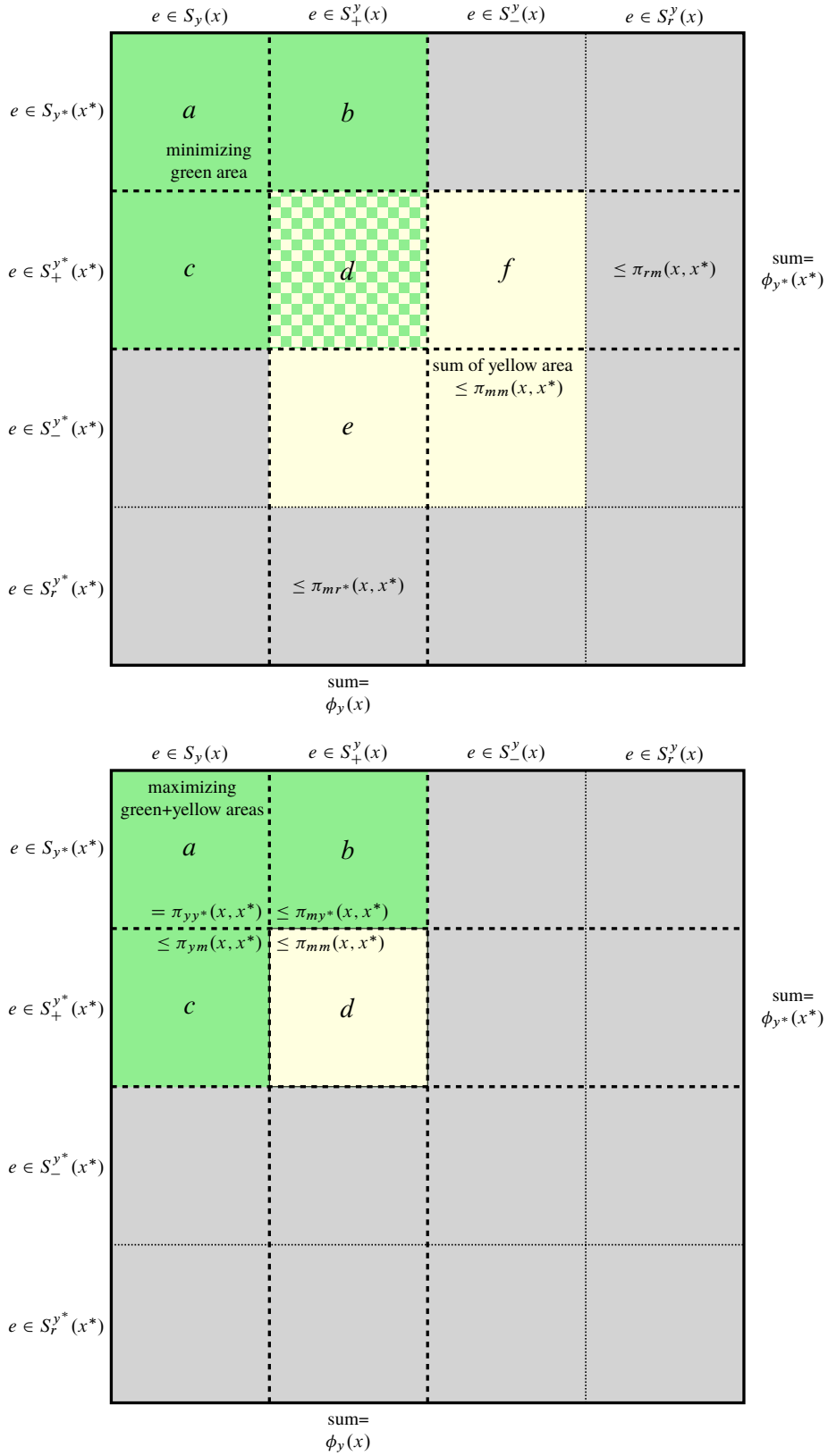

Figure 16: Finding the lower bound (top) and the upper bound (bottom)

References

Aguirregabiria, Victor and Pedro Mira (2007). "Sequential estimation of dynamic discrete games". *Econometrica* 75.1, pp. 1–53.

— (2018). "Identification of games of incomplete information with multiple equilibria and unobserved heterogeneity". *University of Toronto Department of Economics Working Paper*.

Aradillas-Lopez, Andres (2011). "Nonparametric probability bounds for Nash equilibrium actions in a simultaneous discrete game". *Quantitative Economics* 2.2, pp. 135–171.

Aradillas-López, Andres and Elie Tamer (2008). "The identification power of equilibrium in simple games". *Journal of Business & Economic Statistics* 26.3, pp. 261–283.

Athey, Susan and Guido W Imbens (2007). "Discrete choice models with multiple unobserved choice characteristics*". *International Economic Review* 48.4, pp. 1159–1192.

Aumann, Robert J (1961). "Borel structures for function spaces". *Illinois Journal of Mathematics* 5.4, pp. 614–630.

Bajari, Patrick, C Lanier Benkard, and Jonathan Levin (2007). "Estimating dynamic models of imperfect competition". *Econometrica* 75.5, pp. 1331–1370.

Bajari, Patrick, Jinyong Hahn, Han Hong, and Geert Ridder (2011). "A note on semiparametric estimation of finite mixtures of discrete choice models with application to game theoretic models". *International Economic Review* 52.3, pp. 807–824.

Bajari, Patrick, Han Hong, and Stephen P Ryan (2010). "Identification and estimation of a discrete game of complete information". *Econometrica* 78.5, pp. 1529–1568.

Baker, Alan (2013). "Simplicity". *Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2013.

Bjorn, Paul and Quang Vuong (1984). *Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation*. Tech. rep. CalTech.

Bresnahan, Timothy F and Peter C Reiss (1991a). "Empirical models of discrete games". *Journal of Econometrics* 48.1, pp. 57–81.

— (1991b). "Entry and competition in concentrated markets". *Journal of Political Economy*, pp. 977–1009.

Briesch, Richard A, Pradeep K Chintagunta, and Rosa L Matzkin (2012). "Nonparametric discrete choice models with unobserved heterogeneity". *Journal of Business & Economic Statistics*.

Bulow, Jeremy I, John D Geanakoplos, and Paul D Klemperer (1985). "Multimarket oligopoly: Strategic substitutes and complements". *Journal of Political economy* 93.3, pp. 488–511.

Cass, David and Karl Shell (1983). "Do sunspots matter?" *Journal of Political Economy* 91.2, pp. 193–227.

Chen, Xiaohong and Demian Pouzo (2012). "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals". *Econometrica* 80.1, pp. 277–321.

Ciliberto, Federico and Elie Tamer (2009). "Market structure and multiple equilibria in airline markets". *Econometrica* 77.6, pp. 1791–1828.

Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons.

Foster, Dean and Peyton Young (1990). "Stochastic evolutionary game dynamics∗". *Theoretical population biology* 38.2, pp. 219–232.

Fréchet, Maurice (1935). "Généralisations du théorème des probabilites totales". *Fundamenta Mathematica* 25, pp. 379–387.

Galichon, Alfred and Marc Henry (2011). "Set identification in models with multiple equilibria". *Review of Economic Studies*, pp. 1264–1298.

Golan, Amos (2017). *Foundations of info-metrics: modeling and inference with imperfect information*. Oxford University Press.

Golan, Amos, George G Judge, and Douglas Miller (1996). *Maximum entropy econometrics: Robust estimation with limited data*. Wiley New York.

Grieco, Paul LE (2014). "Discrete games with flexible information structures: An application to local grocery markets". *Rand Journal of Economics* 45.2, pp. 303–340.

Grünwald, Peter D and A Philip Dawid (2004). "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory". *Annals of Statistics*, pp. 1367–1433.

Haile, Philip A and Elie Tamer (2003). "Inference with an incomplete model of English auctions". *Journal of Political Economy* 111.1, pp. 1–51.

Harremoës, Peter and Flemming Topsøe (2001). "Maximum entropy fundamentals". *Entropy* 3.3, pp. 191–226.

Harsanyi, John C and Reinhard Selten (1988). "A general theory of equilibrium selection in games". *MIT Press Books* 1.

Jaynes, E.T. (1957a). "Information theory and statistical mechanics". *Physical Review* 106, pp. 620–630.

— (1957b). "Information theory and statistical mechanics II". *Physical Review* 108, pp. 171–190.

Jia, Panle (2008). "What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry". *Econometrica*, pp. 1263–1316.

Jun, Sung Jae and Joris Pinkse (2017). *Point decisions in partially identified auction models: an information–theoretic approach*. Tech. rep. The Pennsylvania State University.

Kalai, Adam and Ehud Kalai (2012). "Cooperation in strategic games revisited". *Quarterly Journal of Economics*, qjs074.

Kandori, Michihiro, George J Mailath, and Rafael Rob (1993). "Learning, mutation, and long run equilibria in games". *Econometrica*, pp. 29–56.

Kashaev, Nail (2015). *Testing for Nash behavior in entry games with complete information*. Tech. rep. Penn State.

Kashaev, Nail and Bruno Salcedo (2015). *Identification of solution concepts for discrete semiparametric games with complete information*. Tech. rep. Penn State.

Kasy, Maximilian (2011). "Identification in triangular systems using control functions". *Econometric Theory* 27.03, pp. 663–671.

Kitagawa, Toru (2012). *Estimation and inference for set–identified parameters using posterior lower probability*. Tech. rep. University College London.

Kline, Brendan (2015). "Identification of complete information games". *Journal of Econometrics* 189.1, pp. 117–131.

Kline, Brendan and Elie Tamer (2012). "Bounds for best response functions in binary games". *Journal of Econometrics* 166.1, pp. 92–105.

Kooreman, Peter (1994). "Estimation of econometric models of some discrete games". *Journal of Applied Econometrics* 9.3, pp. 255–268.

Liu, Nianqing, Quang Vuong, and Haiqing Xu (2013). *Rationalization and identification of discrete games with correlated types*. Tech. rep. University of Texas.

Magnolfi, Lorenzo and Camilla Roncoroni (2016). *Estimation of discrete games with weak assumptions on information*. Tech. rep. Yale.

Manski, Charles (2015). *Interpreting point predictions*. Tech. rep. Northwestern University.

Murray, Iain and Edward Snelson (2006). "A pragmatic Bayesian approach to predictive uncertainty". *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Springer, pp. 33–40.

Pakes, Ariel, Michael Ostrovsky, and Steven Berry (2007). "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)". *The Rand Journal of Economics* 38.2, pp. 373–399.

Porter, Robert H. and J. Douglas Zona (1999). "Ohio school milk markets: an analysis of bidding". *RAND Journal* 30.2, pp. 263–288. ISSN: 07416261. URL: http://www.jstor.org/stable/2556080.

Reguant, Mar (2016). *Bounding equilibria in counterfactual analysis*. Tech. rep. Northwestern University.

Schennach, Susanne M (2014). "Entropic latent variable integration via simulation". *Econometrica* 82.1, pp. 345–385.

Seim, Katja (2006). "An empirical model of firm entry with endogenous product–type choices". *Rand Journal of Economics* 37.3, pp. 619–640.

Shore, John and Rodney Johnson (1980). "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy". *IEEE Transactions on information theory* 26.1, pp. 26–37.

Soetevent, Adriaan R and Peter Kooreman (2007). "A discrete-choice model with social interactions: with an application to high school teen behavior". *Journal of Applied Econometrics* 22.3, pp. 599–624.

Song, Kyungchul (2014). "Point decisions for interval–identified parameters". *Econometric Theory* 30.02, pp. 334–356.

Tamer, Elie (2003). "Incomplete simultaneous discrete response model with multiple equilibria". *Review of Economic Studies* 70.1, pp. 147–165.

Teh, Yee Whye (2010). "Dirichlet process". *Encyclopedia of machine learning*. Springer, pp. 280–287.

Topsøe, Flemming (1979). "Information-theoretical optimization techniques". *Kybernetika* 15.1, pp. 8–27.

van Damme, Eric (1995). *Equilibrium selection in team games*. Tech. rep. Tilburg University.

Xu, Haiqing (2014). "Estimation of discrete games with correlated types". *The Econometrics Journal* 17.3, pp. 241–270.