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

August 9, 2016

We study the problem of counterfactual prediction in discrete decision games with complete information, pure strategies, and Nash equilibria. We show that the presence of multiple equilibria poses unique challenges for the problem of counterfactual prediction even if the payoff structure is known in its entirety. We show that multiple types of counterfactuals can be defined and that the prediction probabilities are not generally point–identified. We establish the sharp identified bounds of the prediction probabilities. We further propose, compare, and contrast various decision methods for the purpose of producing a point prediction, namely midpoint prediction, a decision–theoretic possibility using a Dirichlet–based prior, and a maximum–entropy approach. 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.

*We thank the Human Capital Foundation for their support of CAPCP. We thank Paul Grieco, Ken Hendricks, Andrés Aradillas–López, Joachim Freyberger, Ron Gallant, Bruce Hansen, Nail Kashaev, Francesca Molinari, Ulrich Müller, Jack Porter, Mar Reguant, Bruno Salcedo, Xiaoxia Shi, Guofu Tan, conference participants at the Penn State–Cornell econometrics conference at Ithaca, and seminar participants at Texas A&M, the University of Wisconsin at Madison, and Syracuse University for helpful comments and suggestions.

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

‡joris@psu.edu

1. Introduction

We study the problem of counterfactual prediction in discrete decision games with complete information, pure strategies, and Nash equilibria. 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 problems associated with identification and estimation of payoffs in such games, which are mostly a consequence of the existence of multiple equilibria, have been well–studied and it is well–known how to address them (see e.g. Tamer, 2003; Kline and Tamer, 2012; Kline, 2015). However, the presence of multiple equilibria poses unique challenges for the problem of counterfactual prediction even if the payoff structure is known in its entirety. It is the counterfactual prediction problem that is the subject of this paper. What makes the counterfactual prediction problem studied here especially interesting from an econometric perspective is the *incompleteness* of the model because it features a partially identified function p that characterizes equilibrium selection.

To facilitate both the analysis and the exposition, we focus on the simplest possible (and somewhat hackneyed) case that has all relevant features: a single–shot game with two players, binary decisions, and strategic substitutability (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. 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.

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 $\boldsymbol{y}$. Because of strategic substitutability, for other combinations of $\boldsymbol{e}, \boldsymbol{x}$ there exist multiple (in our case two) equilibria: $(1, 0)$ and $(0, 1)$. 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 $\boldsymbol{x} = x$, let $S_m(x)$ be the 'multiplicity region,' i.e. the collection of e–values for which there exist two equilibria, and let $S_y(x)$ the region of e–values for which $\boldsymbol{y} = y$ is the unique equilibrium outcome. The reasons why a particular outcome arises in the multiplicity region is unknown to us. Since identification of the payoff structure does not require knowledge of such reasons (e.g. Tamer, 2003; Kline, 2015), we will not seek to *specify* players' behavior in the multiplicity region, so we will *not* offer an equilibrium refinement. Instead, for an unknown function p, we *characterize* the behavior of players by a probability $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ for which

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

where $\boldsymbol{v}$ represents potential unobserved 'market' heterogeneity (in addition to the payoff shifters $\boldsymbol{e}, \boldsymbol{x}$) that can affect the probability of $(1, 0)$ being selected under multiplicity. The only restriction

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

that we impose on p is that it is *weakly* monotonic in v, which is without loss of generality.[2]

The characterization using p does not impose assumptions because there always exists a probability that a particular equilibrium is reached, albeit that it limits the scope of counterfactual experiments that can be conducted in a sense that is discussed in section 2. 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 more detailed discussion. We know of no other papers that allow for heterogeneity like $\boldsymbol{v}$ in the present context: its role is somewhat reminiscent of that of sunspots (Cass and Shell, 1983) in that $\boldsymbol{v}$ does not affect the payoffs but *can* affect the selection probability under multiplicity.

Now, 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 will be realized. Thus, 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 $\boldsymbol{e} \notin S_m(\boldsymbol{x})$, there is only one equilibrium and hence 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.

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.

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 $\boldsymbol{y}^*$.[3] Since $\boldsymbol{y}^*$ is categorical, we focus on the conditional probabilities of the counterfactual outcome given the observables, i.e.

$$\mathbb{P}(\boldsymbol{y}^* = y^* \mid \boldsymbol{x}^* = x^*, \boldsymbol{x} = x, \boldsymbol{y} = y), \tag{2}$$

where $y, y^* \in \mathcal{S} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. We also condition on $\boldsymbol{x}$, $\boldsymbol{y}$ in (2) 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*

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

is the same as (2), 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 (3) differs from the counterfactual prediction probability (2). In the main text, we focus on two scenarios that we believe are representative, i.e. fixing $\boldsymbol{v}$ and fixing both $\boldsymbol{e}$ and $\boldsymbol{v}$: we provide results for the leading cases in sections 2 and 3 and for other possibilities in section 3.3 and appendix B. We denote the prediction probability (2) in the leading cases by

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

[2] In contexts with more than one multiplicity region, vector–valued p may be needed, but that is not an issue here.

[3] The type of counterfactual we consider in this paper is thus more commonly found in e.g. the treatment effects literature (Heckman, 2005) than in e.g. the industrial organization literature (Reguant, 2016).

respectively and 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 q_v, q_{ev}, 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 for q_v, the counterfactual probability of having a monopoly collapses to the corresponding regression prediction and is hence point–identified, although q_v itself is not point–identified.[4] However, for q_{ev}, even the probability of monopoly is only partially identified.[5]

Paul:would like a distribution over outcomes (perhaps we can do the entropy loss thing I mentioned)

There are several solutions to our partial identification problem. One instinctive 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 *identified set* of counterfactual probability values. We determine the sharp identified sets for both $q_v(y^* \mid x^*, x, y)$ and $q_{ev}(y^* \mid x^*, x, y)$ in section 3: the bounds for other cases are provided in appendix B. 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. Second, 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. Third 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 6. 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[6] 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 6.1, there does not generally exist a single function p that can generate midpoint predictions for different counterfactuals.

Another possibility is to use a decision–theoretic approach. This entails 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). The integrated loss function, namely the risk, is then aggregated into an 'average risk' by integrating over a parameter space consisting of p–functions that are consistent with the distribution of observables. We can determine the function $\tilde{p}$ that minimizes

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

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

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

the average risk and use it to compute counterfactual probabilities.

We investigate the decision–theoretic approach in section 5. 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 5 or something else, the decision–theoretic approach does not solve the fundamental problem at hand. Indeed, 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 decision–theoretic 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.[7] 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 5, the decision–theoretic approach generates predictions that are similar to our preferred method, *maximum entropy,* which is discussed in section 4.

With the maximum entropy method of the information theory literature (Jaynes, 1957a,b; 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 $\boldsymbol{e}, \boldsymbol{x}$).[8] There are numerous philosophical justifications for the application of Occam's razor; we refer the reader to Baker (2013).

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.[9]

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

[7] This objection can be removed by focusing on worst–case risk (minmax), but minmax takes us back to the midpoint prediction, which we have already discussed.

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

[9] 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.

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 (e.g. Cover and Thomas, 2012) and 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. Finally, as noted above, for the choices made in section 5, the Dirichlet and maximum entropy approaches appear to generate similar predictions while the maximum entropy approach is considerably easier to implement.

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 if 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 5 and a new point prediction method based on the maximum entropy concept from the information theory literature in section 4.[10] Further, we demonstrate the virtues of the maximum entropy approach (and to a lesser extent the Dirichlet approach) in a number of examples in section 6. Finally, analogs of our 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

Consider a standard two player binary decision game with strategic substitutability with complete information and pure strategies, i.e.

$$\begin{cases} \boldsymbol{y}_1 = \mathbb{1}\{\varsigma_1(\boldsymbol{x}, \boldsymbol{y}_2) + \boldsymbol{e}_1 \geq 0\}, \\ \boldsymbol{y}_2 = \mathbb{1}\{\varsigma_2(\boldsymbol{x}, \boldsymbol{y}_1) + \boldsymbol{e}_2 \geq 0\}, \end{cases} \tag{4}$$

where $\boldsymbol{x}$ is a vector of exogenous covariates, $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2)$ are errors that are independent of $\boldsymbol{x}$, and $\varsigma_i(x, 1) \leq \varsigma_i(x, 0)$ for all x and $i = 1, 2$. The model in (4) 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 (see e.g. Bresnahan and Reiss, 1991a; Tamer, 2003).

The properties of the model in (4) have been studied extensively and conditions under which ς_j and the distribution of $(\boldsymbol{e}, \boldsymbol{x})$ are identified are well–understood (e.g. Tamer, 2003). From hereon, we therefore take the functions ς_1, ς_2 and the distribution of $(\boldsymbol{e}, \boldsymbol{x})$ as given.

[10] 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.

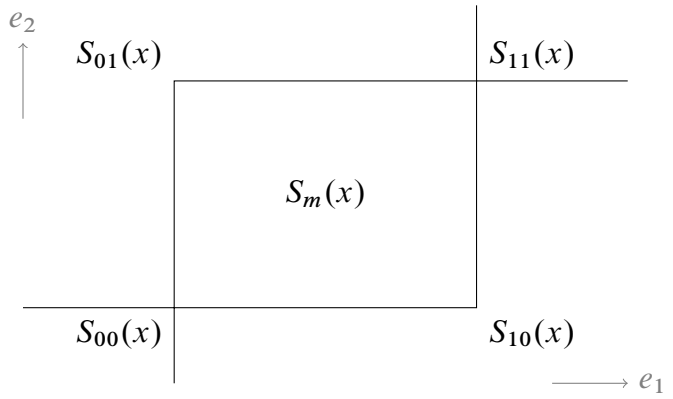

Figure 1: Regions of e–values corresponding to either a specific equilibrium and a multiplicity region.

We follow Tamer (2003) and split the (e_1, e_2)–space ($\mathbb{R}^2$) into five regions: four regions ($S_{00}(x)$, $S_{01}(x)$, $S_{10}(x)$, $S_{11}(x)$) in which a unique Nash equilibrium arises plus a *multiplicity region* $S_m(x)$, in which either $(1, 0)$ or $(0, 1)$ is the outcome. See figure 1.

Since we are taking the functions ι_j 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.

Recall from (1) that for some random variable $\boldsymbol{p} = p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x})$ and all p, x, if $e \in S_m(x)$ then

$$\mathbb{P}\left\{\boldsymbol{y} = (1,0) \,\middle|\, \boldsymbol{e} = e, \boldsymbol{p} = p, \boldsymbol{x} = x\right\} = p, \tag{5}$$

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.

The relationship in (5) is a characterization, not an assumption. In other words, there always exists such a function p and such a random variable $\boldsymbol{v}$. Equation (5) says that in a given market, under multiplicity $(1, 0)$ is selected with probability $\boldsymbol{p}$, where $\boldsymbol{p}$ can depend on payoff shifters and on (other) unobserved market heterogeneity $\boldsymbol{v}$. In view of the (weak) monotonicity of p in v, $\boldsymbol{v}$ reflects factors (unobservable to the econometrician) that affect the probability of choosing one equilibrium over another over and above what is explained by the payoff shifters. However, (5) is not necessarily a description of players' behavior in the multiplicity region. In particular, it does *not* assume that players randomize in their choice of equilibrium: it merely says that from the econometrician's perspective this choice *can* be viewed as random ($\boldsymbol{p}$ is allowed to be binary).

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) and also nests the structures in Bjorn and Vuong (1984), where $p(e, v, x) = \mu_m(x)$, and Jia (2008), where $p(e, v, x) = 1$. But, unlike Grieco (2014), we allow for an unobservable $\boldsymbol{v}$ in p and p can be flat in e, v. Our specification is a *characterization* instead of an *assumption* and it is consistent with *any* equilibrium refinement.[11] The same can be said for Grieco (2014) 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 some of the unobserved payoff–irrelevant heterogeneity is fixed and some changes. 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).

Although our characterization is consistent with *any* equilibrium refinement, our characterization does limit the nature of the counterfactual experiments discussed in section 3. Indeed, in section 3 we hold $\boldsymbol{v}$ (and possibly $\boldsymbol{e}$) fixed (in our leading cases), but it is possible that an underlying structural model of 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 (5) 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, (5) does not restrict the dependence of $\boldsymbol{e}$ and $\boldsymbol{p}$.

In view of (5) we can represent the outcome $\boldsymbol{y}$ by

$$\boldsymbol{y} = y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}) = \begin{cases} y, & \boldsymbol{e} \in S_y(\boldsymbol{x}), \\ (1,0), & \boldsymbol{e} \in S_m(\boldsymbol{x}) \text{ and } \boldsymbol{u} \leq p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}), \\ (0,1), & \boldsymbol{e} \in S_m(\boldsymbol{x}) \text{ and } \boldsymbol{u} > p(\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}), \end{cases} \tag{6}$$

where $\boldsymbol{u}$ is uniform and independent of $\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}$. So, the outcome is determined by four random elements: $\boldsymbol{e}, \boldsymbol{x}$ are unobservable and observable payoff variables, which *can* also impact $\boldsymbol{p}$ in the multiplicity region,[12] and $\boldsymbol{v}$ captures extra unobserved market heterogeneity that affects $\boldsymbol{p}$ independent of $\boldsymbol{e}, \boldsymbol{x}$. Further, $\boldsymbol{u}$ is an unobservable variable, which, together with $\boldsymbol{p}$, determines which equilibrium is played in the multiplicity region. Note that fixing the payoff variables and $\boldsymbol{v}$ does not determine the outcome because $y(e, \boldsymbol{u}, v, x)$ is still random: only the probability of which outcome is selected under multiplicity is determined by $\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}$, not (necessarily) the outcome itself.

In view of (6), (5) *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 (5) imposes is by itself of limited help for the prediction problem studied in this paper.

[11] Kalai and Kalai (2012) (and references therein) provides refinements for two by two games, albeit that in Kalai and 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.

[12] We do not impose exclusion restrictions here.

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 $(\boldsymbol{e}, \boldsymbol{p}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$.[13] 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 (6) 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 counterfactual of interest is $\boldsymbol{y}^* = y(\boldsymbol{e}, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{x}^*)$.[14] 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,[15] and provide a brief summary thereof in section 3.3.

For each of the counterfactuals, we wish to construct a prediction of $\boldsymbol{y}^*$. Since $\boldsymbol{y}^*$ is a categorical variable, we obtain the distribution of $\boldsymbol{y}^*$ given observables. 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,

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

where q represents the case with no constraints and q_v, q_{ev} represent the cases where (part of) unobserved market heterogeneity is kept unchanged. Please note that we 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

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

[13] 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.

[14] 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.

[15] 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}.

The *regression prediction* $\tilde{\pi}_{y^*}(x^*)$ is simple, but it represents the case where nothing 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. For this purpose, we define

$$\mu_m(x) = \mathbb{E}\{\boldsymbol{p} \mid \boldsymbol{e} \in S_m(x), \boldsymbol{x} = x\}. \tag{7}$$

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 $\pi_m(x) > 0$.[16] 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 $p(\cdot, \cdot, x)$ only through $\mu_m(x)$.

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

$$\delta_y = \mathbb{1}\{y = (1,0)\} - \mathbb{1}\{y = (0,1)\}. \tag{8}$$

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

$$\begin{aligned}\rho(x, x^*) = \mathrm{Cov}\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}^*, \boldsymbol{v}, x^*) \mid \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\} \\ = \mathrm{Cov}\{\hbar(\boldsymbol{v}, x), \hbar(\boldsymbol{v}, x^*)\},\end{aligned} \tag{9}$$

where

$$\hbar(v, x) = \mathbb{E}\{\boldsymbol{p} \mid \boldsymbol{e} \in S_m(x), \boldsymbol{v} = v, \boldsymbol{x} = x\} = \mathbb{E}\{p(\boldsymbol{e}, v, x) \mid \boldsymbol{e} \in S_m(x)\}. \tag{10}$$

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 it belongs to the multiplicity region or not: 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 (8). Then, $q_v(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \tilde{\rho}_{yy^*}(x, x^*)$. □

The results of theorem 1 are intuitive. 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. 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 $\pi_m(x^*)$

[16] If $\pi_m(x) = 0$ then the value of $\mu_m(x)$ is immaterial.

is large) then the correction term is large. 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 $p(\boldsymbol{e}, \boldsymbol{v}, x)$ is highly correlated with $p(\boldsymbol{e}^*, \boldsymbol{v}, x^*)$ 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

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

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. 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, both the maximum difference between the optimal and regression predictions and the maximum attainable length of the identified set of $q_v\{(1,0) \mid x^*, x, (1,0)\}$ equal

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

which 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, the maximum length of the identified set as $\pi_m(x), \pi_m(x^*) \to 1$ is $\min\{1, \mu_m(x^*) \,/\, \mu_m(x)\} - \mu_m(x^*)$. More generally, the maximum attainable length of the identified set varies with the ratio $\pi_{10}(x) \,/\, \pi_m(x)$ as depicted in figure 2. Much the same applies to other combinations of $y, y^* \in \{(1,0), (0,1)\}$.

3.2 Case 2: $\boldsymbol{e}^* = \boldsymbol{e}$ and $\boldsymbol{v}^* = \boldsymbol{v}$: We now turn to the case where both $\boldsymbol{e}$ and $\boldsymbol{v}$ are fixed, i.e. the probability $\boldsymbol{p}$ 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 $\boldsymbol{e}$ is important for prediction. For instance, if $x^* = x$ then the S–regions are unchanged and $\boldsymbol{p}^* = \boldsymbol{p}$.[17] If $x^* \neq x$ then the S–regions will be different and $\boldsymbol{p}^*$ can be different from $\boldsymbol{p}$, 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.

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

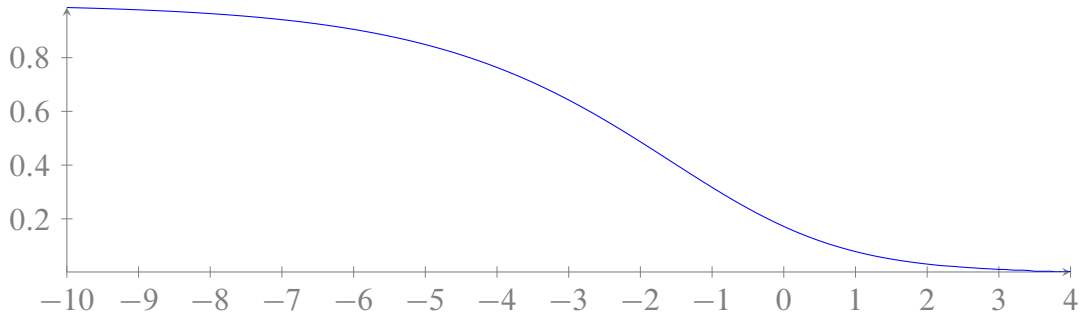

Figure 2: Maximum possible value of $\tilde{\rho}_{yy^*}(x, x^*)$ as a function of $\log\{\pi_y(x) \,/\, \pi_m(x)\}$ if $\mu_m(x^*) = \mu_m(x)$ and $y^* = y = (1, 0)$

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.[18] We further define $S_r(x) = S_r^y(x) = \left\{S_y(x) \cup S_m(x)\right\}^c$ and $S_{r^*}(x) = S_r^{y^*}(x) = \left\{S_{y^*}(x) \cup S_m(x)\right\}^c$. In order to determine the bounds on q_{ev}, it matters where the S–regions intersect. Let

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

which is similar to (9) but uses $\boldsymbol{e} = \boldsymbol{e}^*$ in both p functions, whereas in (9) $\boldsymbol{e}^*$ was an independent copy of $\boldsymbol{e}$. Indeed, $\rho(x, x^*)$ serves the same role as in theorem 1: to correct for the information contained in $\boldsymbol{x}$, $\boldsymbol{y}$ about the value of $\boldsymbol{p}^*$.

But we need another correction term to address the fact that $\boldsymbol{x}$, $\boldsymbol{y}$ contain information about the $S.(x^*)$–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 (5), $q_{ev}(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \alpha_{yy^*}(x, x^*) + \tilde{\rho}_{yy^*}(x, x^*)$. □

The intuition for the $\tilde{\rho}$–term in theorem 3 is similar to that in theorem 1. For the α term in theorem 3, consider first the case that $y = y^* = (0, 0)$. Then the α_{yy^*} term simply increases the

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

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. The intuition for the other cases are merely more complicated versions of this argument.

Before we proceed, we introduce some notation. 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

$$\begin{aligned} q_{ev}\big\{(1,0) \;\big|\; x^*, x, y\big\} + q_{ev}\big\{(0,1) \;\big|\; x^*, x, y\big\} \\ = \tilde{\pi}_{10}(x^*) + \tilde{\pi}_{01}(x^*) + \frac{\mathrm{Cov}\big\{a_y(\boldsymbol{e}, x), a_{10}(\boldsymbol{e}, x^*) + a_{01}(\boldsymbol{e}, x^*)\big\}}{\tilde{\pi}_y(x)}. \end{aligned} \tag{11}$$

The covariance on the right hand side in (11) 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: i.e. the probability of monopoly conditional on no monopoly 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 $q_{ev}\left(y^* \mid x^*, x, y\right)$ 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

$$U_{yy^*}(x, x^*) = \text{min. of} \begin{cases} \phi_y(x) + \phi_{y^*}(x^*), \\ \phi_y(x) + \pi_{ym}(x, x^*), \\ \phi_{y^*}(x^*) + \pi_{my^*}(x, x^*), \\ \pi_{mm}(x, x^*) + \pi_{ym}(x, x^*) + \pi_{my^*}(x, x^*), \end{cases} \tag{12}$$

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 3. Depending on the values of $\phi_y(x)$ and $\phi_y(x^*)$, different bounds are binding.

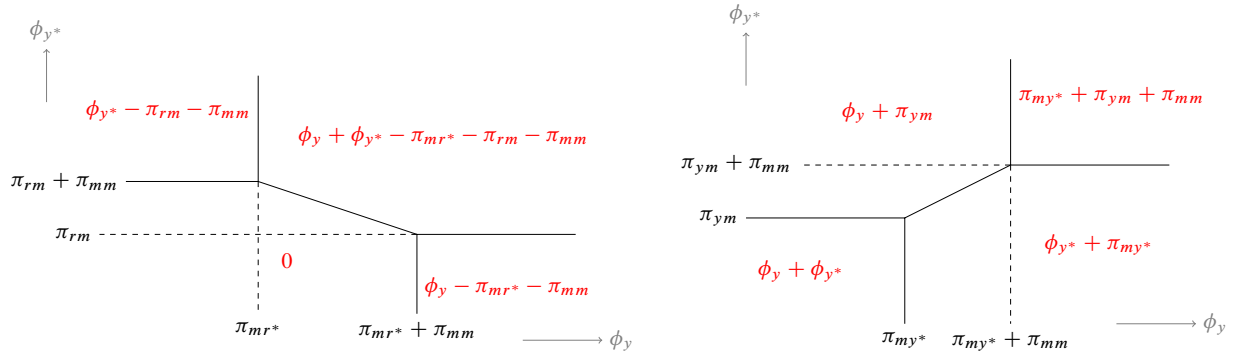

Figure 3: Regions of $\left(\phi_y(x), \phi_{y^*}(x^*)\right)$ and the corresponding values of $L_{yy^*}(x, x^*), U_{yy^*}(x, x^*)$.

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

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

The bounds are sharp in both cases. □

The fact that the formulas for the bounds in theorem 5 depend fundamentally on whether or not $x = x^*$ 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 $p(e, v, x)$ is 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 6.3 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

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

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.

$$\begin{cases} q(y^* \mid x^*, x, y) \;\leq\; q_v(y^* \mid x^*, x, y) \;\leq\; q_{uv}(y^* \mid x^*, x, y), \\ q_e(y^* \mid x^*, x, y) \;\leq\; q_{ev}(y^* \mid x^*, x, y) \;\leq\; q_{euv}(y^* \mid x^*, x, y), \end{cases} \tag{13}$$

which is intuitive. However, if $x \neq x^*$, as is shown in theorems 8 to 12 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}.[19] Since the bounds coincide, so do the midpoint predictions. Thus, despite the fact that the q–functions can be ordered as indicated in (13), the midpoint predictions are the same in each case.[20] The maximum entropy solution proposed in section 4 does not have this unfortunate feature.

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,b). The entropy measures the amount of uncertainty that a probability distribution represents.[21]

Can we create 'bounds' by showing what a loss of e.g. one unit of entropy would do to the predictions?

In our case, this entails finding the joint density of $\boldsymbol{e}, \boldsymbol{p}, \boldsymbol{x}$ which maximizes the entropy: recall from (5) 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 $\boldsymbol{x}$ as being 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 (7), is identified for all x, and hence it imposes another constraint on f. Thus, using all information available, we consider maximizing the *conditional entropy*[22]

$$
\begin{aligned}
f^* = \operatorname*{argmin}_f \iiint_0^1 f(p \mid e, x) \log f(p \mid e, x) \,\mathrm{d}p \, f_e(e) f_x(x) \,\mathrm{d}e \,\mathrm{d}x \\
\text{subject to} \begin{cases} \forall e, x : \int_0^1 f(p \mid e, x) \,\mathrm{d}p = 1, \\ \forall x : \int_{S_m(x)} \int_0^1 p f(p \mid e, x) \,\mathrm{d}p \, f_e(e) \,\mathrm{d}e = \mu_m(x) \pi_m(x). \end{cases}
\end{aligned} \tag{14}
$$

In (14) we are minimizing minus the entropy to get rid of the minus sign in the entropy definition. The optimization problem in (14) can be solved using standard constrained optimization techniques (see e.g. Cover and Thomas, 2012). The result is provided in theorem 6 below. Let

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

Theorem 6. The solution to (14) is given by

$$
f^*(p \mid e, x) = \begin{cases} \mathbb{1}(0 \leq p \leq 1), & e \notin S_m(x), \\ A\{p, \lambda_m(x)\} \mathbb{1}(0 \leq p \leq 1), & e \in S_m(x), \end{cases} \tag{15}
$$

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

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

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

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

where

$$\lambda_m(x) = \operatorname*{argmin}_{\lambda}\{\mathcal{L}(\lambda) - \mu_m(x)\lambda\}, \qquad A(p,\lambda) = \exp(p\lambda) \;/\; I(\lambda). \tag{16}$$

□

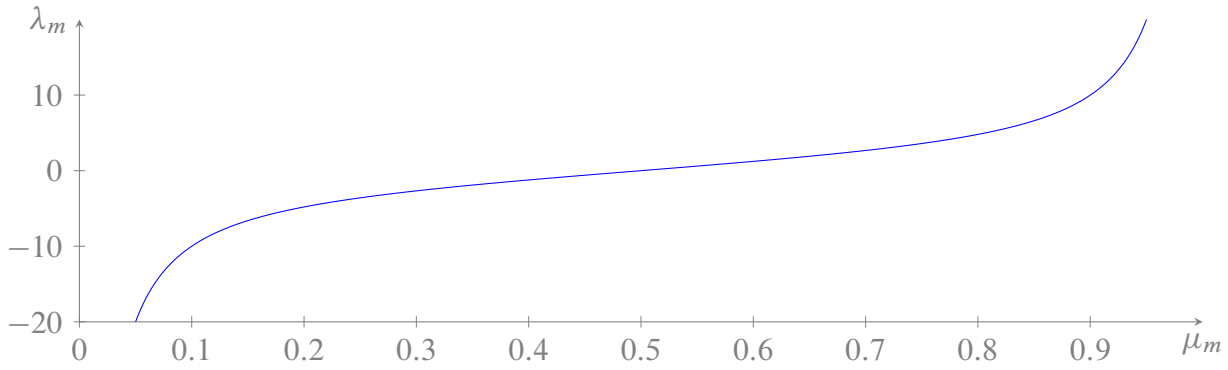

Figure 4: λ_m as a function of μ_m.

The relationship between λ_m and μ_m is depicted in figure 4. We can see from the first order condition of the minimization problem in (16) that the solution λ_m satisfies

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

where it should be noted that $\mathcal{L}'$ is well–behaved at zero.[23]

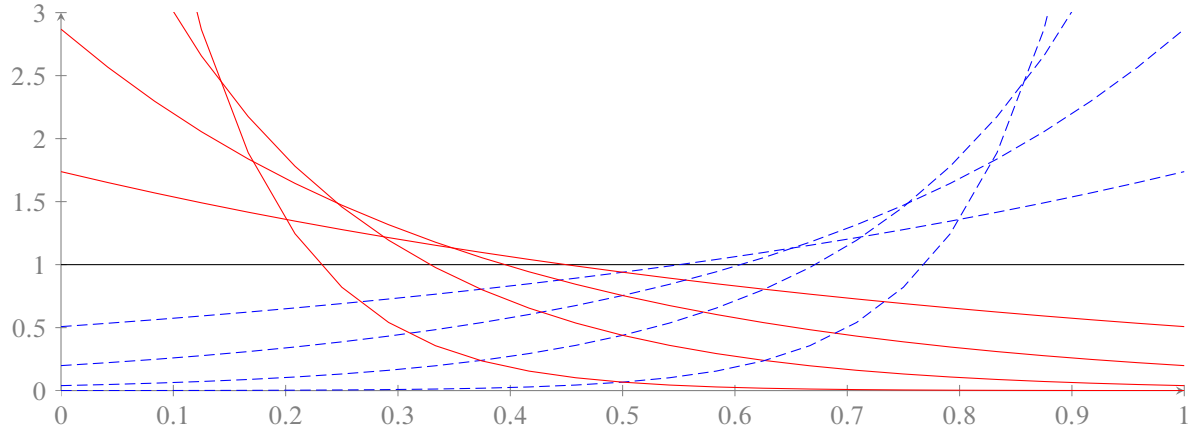

Figure 5: Conditional densities $f^*(p \mid e, x)$ for $e \in S_m(x)$ for values of $\mu_m = 0.1, \ldots, 0.9$, with blue dashed curves corresponding to $\mu_m > 0.5$ and red solid curves to $\mu_m < 0.5$.

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

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

Recall again 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 (10) 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

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

and

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

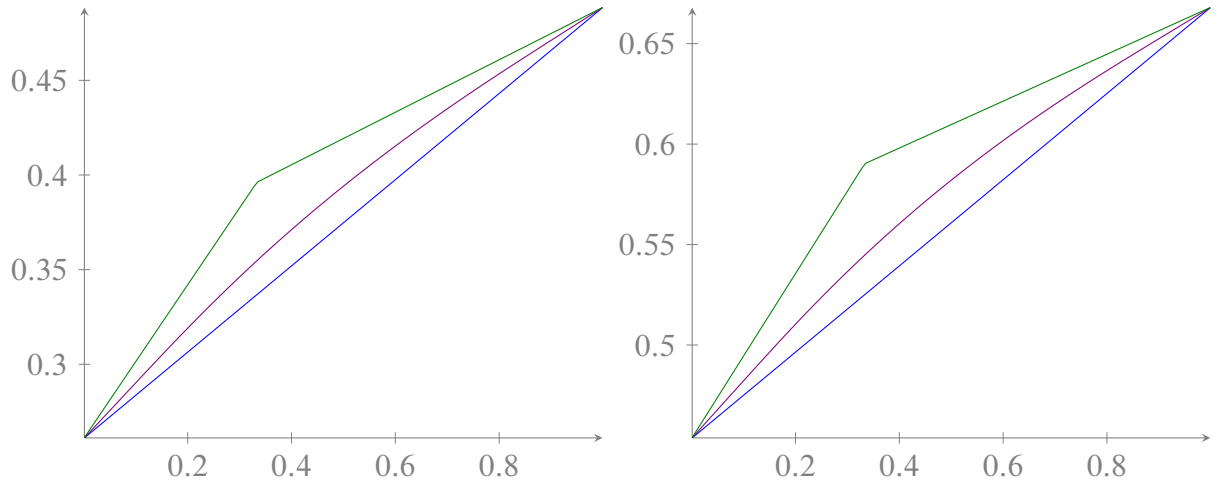

Figure 6: Maximum entropy solutions for $q_\cdot, q_v, q_{uv}$ (left graph, bottom to top) and q_e, q_{ev}, q_{euv} (right graph, bottom to top) as a function of $\mu_m(x^*)$ if $y = y^* = (1,0)$, $\mu_m(x) = 1\ /\ 3$, $\boldsymbol{e}_1, \boldsymbol{e}_2$ independent $N(0,1)$, shift in x value moves multiplicity region from $[-1,-1]^2$ to $[0,2]^2$.

For one simple design, the maximum entropy predictions are depicted in figure 6. 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 6, we will compare the maximum entropy solutions with various alternatives using a number of stylized examples.

The maximum entropy solution is our preferred method to make a point decision. Its main alternative, the Dirichlet solution, is discussed in section 5. The Dirichlet solution is of intrinsic interest on its own but since its derivation is more technical, readers who are more interested in the comparison of methods than in the details of the Dirichlet solution can skip ahead to section 6.

5. Dirichlet approach

5.1 Overview: In this section we discuss an alternative possibility, for which Dirichlet processes will be used. The approach we discuss has a decision–theoretic foundation, but its implementation is considerably more complicated than maximum entropy and it does not do much to mitigate the inherent arbitrariness of the problem at hand. We will compare the results to the maximum entropy ones in section 6.

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.[24] Suppose further that $\boldsymbol{G}$ is a draw from the prior. Then, our prediction based on the draw $\boldsymbol{G}$ will be

$$T_{\boldsymbol{G}^{-1}}(x, x^*, y, y^*), \tag{17}$$

which is a random object. The approach considered in this section is based on the probability distribution of (17). We show in section 5.2 that the mean of (17) 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 (5) that for $e \in S_m(x)$,

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

which implies

$$\frac{1}{\pi_m(x)} \int_{S_m(x)} \mathbb{P}\{\boldsymbol{y} = (1,0) \mid \boldsymbol{e} = e, \boldsymbol{x} = x; p\} f_e(e)\, \mathrm{d}e = \mu_m(x). \tag{18}$$

Equation (18) 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.

5.2 Average risk optimality: Our decision problem is to make the 'best' guess for $T_p(x, x^*, y, y^*)$ based on $\boldsymbol{x} = x, \boldsymbol{y} = y$ and $\boldsymbol{x}^* = x^*, \boldsymbol{y}^* = y^*$. Denote our decision by $d(x, x^*, y, y^*)$. For a given loss function ℓ, the risk (for any combination x, x^*) is then

$$\begin{aligned} &R(d, p, x, x^*) \\ &\quad = \sum_{y, y^*} \ell\big\{d(x, x^*, y, y^*), T_p(x, x^*, y, y^*)\big\} \mathbb{P}(\boldsymbol{y} = y \mid \boldsymbol{x} = x; p) \mathbb{P}(\boldsymbol{y}^* = y^* \mid \boldsymbol{x}^* = x^*; p) \\ &\qquad\qquad = \sum_{y, y^*} \ell\big\{d(x, x^*, y, y^*), T_p(x, x^*, y, y^*)\big\} L(y \mid x) L(y^* \mid x^*), \end{aligned} \tag{19}$$

where the likelihood function L is known and independent of p (other than via μ_m), as was established in (18). However, since the prediction of interest depends on p, the risk in (19) depends on p, also. Therefore, we consider the average risk, i.e. averaging out over p, for which we need a prior for p.

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

$$\mathcal{F}_x = \left\{p^{-1}(\cdot, \cdot, x) : \ \frac{1}{\pi_m(x)} \int_{S_m(x)} \int_0^1 p(e, v, x) f_e(e)\, \mathrm{d}e\, \mathrm{d}v = \mu_m(x)\right\}. \tag{20}$$

[24] 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 $Q^{-1}(p) = \sup\{p : \ Q(v) \le p\}$, which is a distribution function. Similarly, for a distribution function G, the generalized inverse G^{-1} is defined by $G^{-1}(v) = \inf\{x : \ G(x) \ge v\}$, which is a quantile function.

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 5.3.

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

$$\begin{aligned}\bar{R}(d) &= \iiint_{\mathcal{F}_{xx^*}} R(d, p, x, x^*)\, \mathrm{d}\omega_{xx^*}(p^{-1}) f_{\boldsymbol{x}}(x) f_{\boldsymbol{x}}(x^*)\, \mathrm{d}x\, \mathrm{d}x^* \\ &\geq \iint \sum_{y,y^*} \min_s \int_{\mathcal{F}_{xx^*}} \ell\{s, T_p(x, x^*, y, y^*)\}\, \mathrm{d}\omega_{xx^*}(p^{-1}) L(y \mid x) L(y^* \mid x^*) f_{\boldsymbol{x}}(x) f_{\boldsymbol{x}}(x^*)\, \mathrm{d}x\, \mathrm{d}x^*,\end{aligned}$$

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 $\bar{R}$ is given by

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

Thus, when the loss function is quadratic, $d^*(x, x^*, y, y^*)$ is simply a mean as we explained below (17).

5.3 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

$$\mathcal{F}_x = \left\{ p^{-1}(\cdot, x) : \int_0^1 p(v, x)\, \mathrm{d}v = \mu_m(x) \right\}, \tag{21}$$

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

Paul:why is this intro in here twice?

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

$$\forall\, j : \boldsymbol{W}_j = \boldsymbol{\beta}_j \boldsymbol{\tau}_j + (1 - \boldsymbol{\beta}_j) \boldsymbol{W}_{j+1}. \tag{22}$$

Further, the Bayes rule says that

$$f_{\boldsymbol{\tau}_j | \boldsymbol{\beta}_j, \boldsymbol{W}_j}(\tau \mid \beta, \mu) = \frac{f_{\boldsymbol{W}_j | \boldsymbol{\tau}_j, \boldsymbol{\beta}_j}(\mu \mid \tau, \beta)}{f_{\boldsymbol{W}_j | \boldsymbol{\beta}_j}(\mu \mid \beta)}. \tag{23}$$

Here, by (22) and the fact that the distribution of $\boldsymbol{W}_j$ does not depend on j, it follows that

$$\mathbb{P}(\boldsymbol{W}_j \le w \mid \boldsymbol{\tau}_j = \tau_j, \boldsymbol{\beta}_j = \beta_j) = \mathbb{P}\big\{\beta_j \tau_j + (1 - \beta_j)\tau_j \boldsymbol{W}_{j+1} \le w\big\} = F^\circ\bigg(\frac{w - \beta_j \tau_j}{1 - \beta_j}\bigg), \tag{24}$$

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

Finally, note from (23) and (24) that $F_{\boldsymbol{\tau}_j | \boldsymbol{\beta}_j, \boldsymbol{W}_j}(\tau \mid \beta, \mu) = \tilde{F}^\circ(\tau; \beta, \mu)$, where

$$\tilde{F}^\circ(t; \tilde{\beta}, \tilde{\mu}) = \frac{F^\circ\Big(\frac{\tilde{\mu}}{1 - \tilde{\beta}}\Big) - F^\circ\Big(\frac{\tilde{\mu} - \tilde{\beta} t}{1 - \tilde{\beta}}\Big)}{F^\circ\Big(\frac{\tilde{\mu}}{1 - \tilde{\beta}}\Big) - F^\circ\Big(\frac{\tilde{\mu} - \tilde{\beta}}{1 - \tilde{\beta}}\Big)}, \quad \max\Big\{0, \frac{\tilde{\mu} - (1 - \tilde{\beta})}{\tilde{\beta}}\Big\} \le t \le \min\Big(1, \frac{\tilde{\mu}}{\tilde{\beta}}\Big). \tag{25}$$

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

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

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

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

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

It now remains to be shown how to compute F°. From (22) 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.

6. Comparison of methods

In this section we compare the maximum entropy approach with other possibilities. Section 6.1 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 6.2 uses two simple designs to compare the sharp bounds on q_v to regression predictions, midpoint predictions, and maximum entropy predictions. Section 6.3 does the same for q_{ev}. Then, section 6.4 compares and contrasts all methods to the Dirichlet–process–based idea of section 5.

6.1 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

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

because the midpoint predictions of $q_v(y^* \mid \bar{x}, \bar{x}, y)$ and $q_{ev}(y^* \mid \bar{x}, \bar{x}, y)$ coincide by theorems 2 and 5. Therefore, we have

$$\mathbb{E}\mathbb{V}\{p^\circ(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{v}\} = 0,$$

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

$$q_{ev}^\circ(y^* \mid x^*, x, y) = \frac{\mathbb{E}\{p^\circ(\boldsymbol{v}, x)p^\circ(\boldsymbol{v}, x^*)\}}{\tilde{\mu}} = \frac{1}{2}, \tag{26}$$

$$q_{ev}^\circ(y^* \mid x, x, y) = \frac{\mathbb{E}p^{\circ 2}(\boldsymbol{v}, x)}{\tilde{\mu}} = \frac{3}{4}, \tag{27}$$

$$q_{ev}^\circ(y^* \mid x^*, x^*, y) = \frac{\mathbb{E}p^{\circ 2}(\boldsymbol{v}, x^*)}{\tilde{\mu}} = \frac{3}{4}. \tag{28}$$

Equation (26) 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 (27) or (28). 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 6.

6.2 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 6.3 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

$$\rho^*(x, x^*) = \mathbb{V}\{\hbar^*(\boldsymbol{v}, \boldsymbol{x}) \mid \boldsymbol{x} = x\}$$
$$= \int_0^1 p^2 A\{p, \lambda_m(x)\}\,\mathrm{d}p - \left\{\int_0^1 pA\{p, \lambda_m(x)\}\,\mathrm{d}p\right\}^2 = \mathscr{L}''\{\lambda_m(x)\}, \quad (29)$$

where A is defined in (16). Example 2 is based on (29).

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 7 depicts prediction probabilities as a function of $\tilde{\mu}$ in each case.

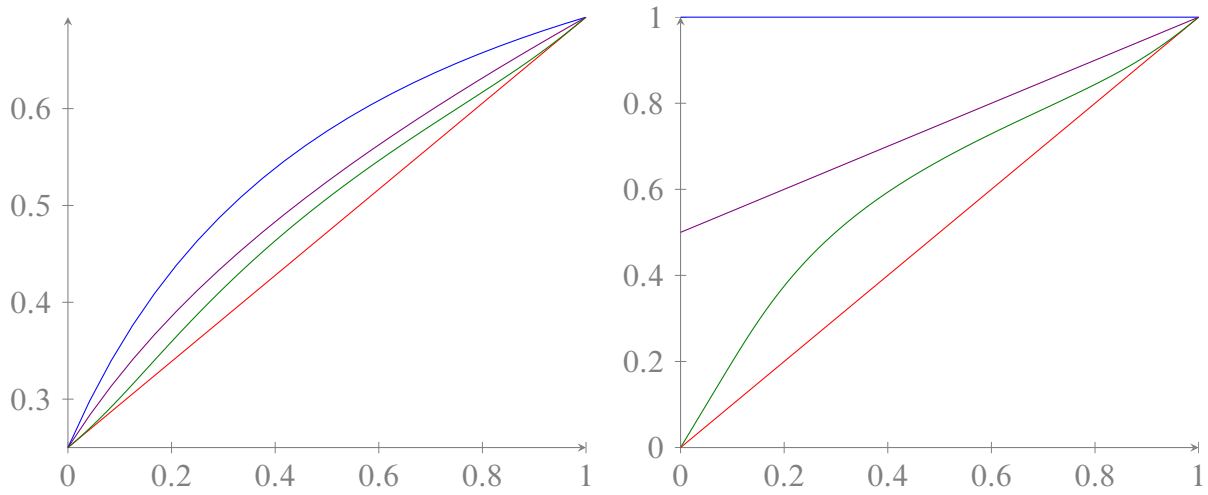

Figure 7: Infeasible predictions as a function of $\mu_m(x) = \mu_m(x^*)$ if $y = y^* = (1, 0)$ using regression (red; also smallest), largest (blue), midpoint (purple), and maximum entropy (green) methods. In both panels $\pi_m(x) = \pi_m(x^*)$ and $\pi_y(x) = \pi_y(x^*)$. In the left panel, $\pi_m(x) = 4/9$ and $\pi_y(x) = 1/4$, whereas in the right panel $\pi_m(x) = 1$ and $\pi_y(x) = 0$.

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 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

$$p(e, v, x) = \mathbb{1}(v > 1 - \tilde{\mu}), \quad e \in S_m(x), \quad (30)$$

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

$$q_v\{(1, 0) \mid x^*, x, (1, 0)\} = \frac{1}{\tilde{\mu}} \int_0^1 \hbar(v, x)\hbar(v, x^*)\,\mathrm{d}v. \quad (31)$$

For most functions p, (31) will be close to zero if $\tilde{\mu}$ is close to zero. However, for p in (30), 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 (30); see also example 5. The maximum entropy method, in contrast, does not have this problem. □

6.3 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 6.2. Consequently, the formula for $\rho^*(x, x^*)$ is again given by (29), i.e. $\rho^*(x, x^*) = \mathscr{L}''\{\lambda_m(x^*)\}$. The function a_y in theorem 3 using p^* is provided in (47) 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

$$\big[\pi_y(x)\{1 - \pi_y(x)\} + \mu_m^2(x)\pi_m(x)\{1 - \pi_m(x)\} - 2\pi_y(x)\pi_m(x)\mu_m(x)\big] \,/\, \tilde{\pi}_y(x). \tag{32}$$

Example 3 below is based on (29) and (32).

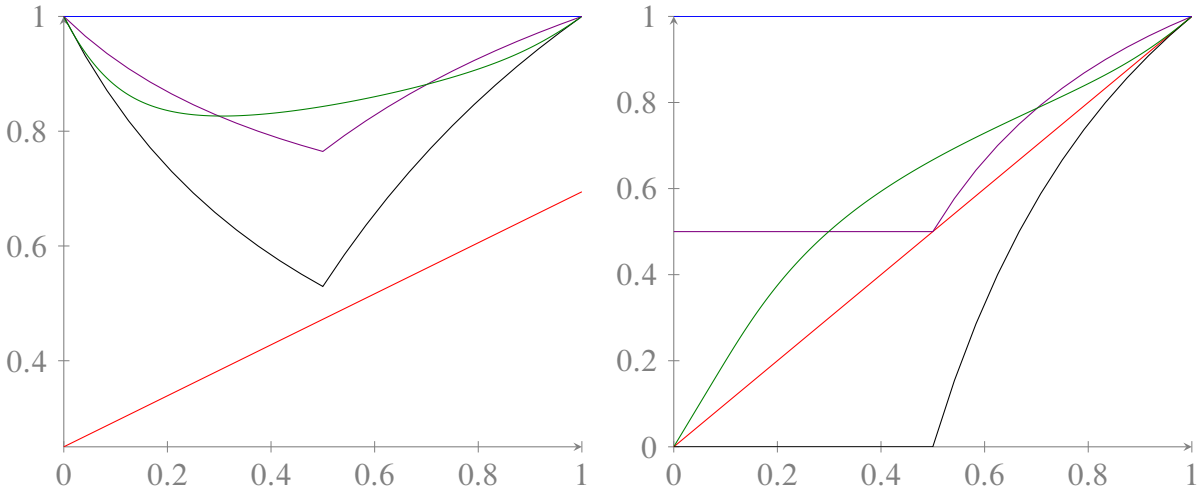

Figure 8: Infeasible predictions as a function of $\mu_m(x) = \mu_m(x^*)$ if x and x^* correspond to the same S–regions with $x \neq x^*$, and $y = y^* = (1, 0)$. The predictions are produced using the regression (red), smallest (black), midpoint (purple), largest (blue), and maximum entropy (green) 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 $\pi_y(x) = 0$.

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 8 and 9 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 8. 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

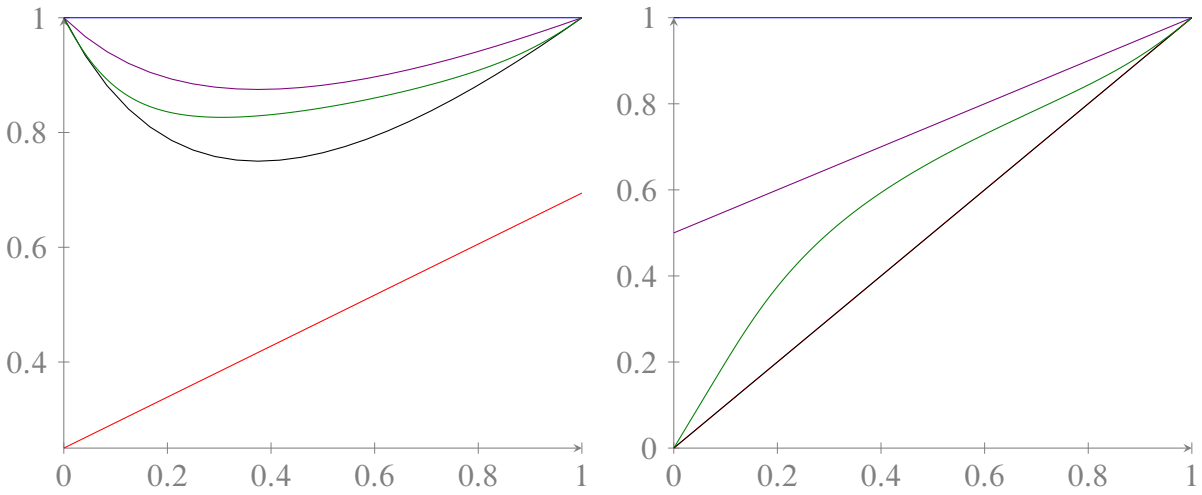

Figure 9: Infeasible predictions as a function of $\mu_m(x)$ if $x = x^*$ and $y = y^* = (1, 0)$, using regression (red), smallest (black), midpoint (purple), largest (blue), and maximum entropy (green) methods. In the left panel, $\pi_m(x) = 4 \ / \ 9$ and $\pi_y(x) = 1 \ / \ 4$, whereas in the right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

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 8 and 9 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

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

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 5, we restricted p to be flat in e. However, since the parameter space $\mathcal{F}_x$ in (20) 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.

6.4 Dirichlet–based predictions, midpoints, and maximum entropy: We now consider the average risk optimal predictions using the prior we proposed in section 5. Instead of choosing a particular loss function, we consider the distribution of the prediction when p is drawn by the algorithm

described in section 5.3. 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 7 and 8, 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

$$\frac{\mathbb{E}\{\hbar(\boldsymbol{v}, x)\hbar(\boldsymbol{v}, x^*)\}}{\mu_m(x)}. \tag{33}$$

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

$$\frac{\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x)p(\boldsymbol{e}, \boldsymbol{v}, x^*)\}}{\mu_m(x)}. \tag{34}$$

Now, consider the restricted parameter space defined in (21). Then, in our current design, (33) and (34) coincide and equal

Paul:restricted b/c of independence of e?

$$T_p(x, x^*, y, y^*) = \frac{\mathbb{E}\{p(\boldsymbol{v}, x)p(\boldsymbol{v}, x^*)\}}{\mu_m(x)}. \tag{35}$$

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 (35) is an upper bound of both (33) and (34).

Figure 10 depicts the distribution of (35) when the function p is a random draw from the Dirichlet–like–process $\mathcal{D}^*_{\mu_m}$ described in section 5.3. Recall from section 5 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.

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 (35) for a different function p.

The results of our experiments are depicted in figure 10. The left panel in figure 10 is identical to the right panel in figure 7 and the right panel in figure 10 is identical to the right panel in figure 8, 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 grey 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 (35), 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 10 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.

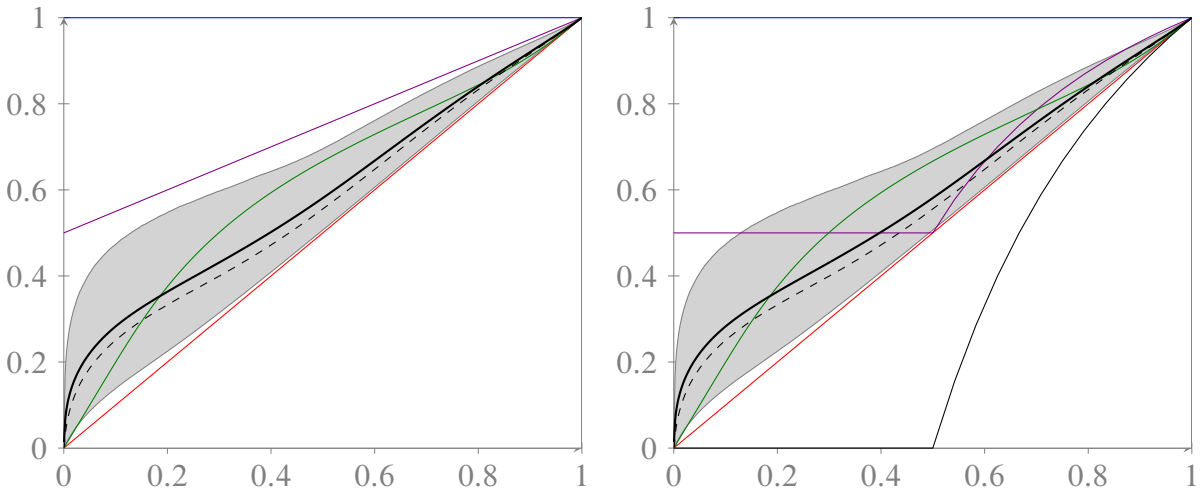

Figure 10: Infeasible predictions as a function of $\mu_m(x) = \mu_m(x^*)$ if $y = y^* = (1, 0)$ using regression (red), largest (blue), midpoint (purple), and maximum entropy (green) methods. The grey regions depict the area between 10% and 90% quantiles of the Dirichlet process–based experiment described in section 5.3 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.

Further, we chose a uniform prior and pseudocount hyperparameter equal to one for the Dirichlet process for convenience.[25] 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 (35) 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,

[25] 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.

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. □

7. 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 four 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, 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,

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

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

$$\operatorname{Cov}\big[\mathbb{1}\{y(\boldsymbol{e}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, x^*) = y^*\}, \mathbb{1}\{y(\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{v}^*, x) = y\} \,\big|\, \boldsymbol{e} \in S_m(x), \boldsymbol{e}^* \in S_m(x^*)\big]. \quad (36)$$

The covariance in (36) equals zero unless both y and y^* belong to $\{(1,0),(0,1)\}$. For $y = y^* = (1,0)$ the covariance in (36) equals (by the law of iterated expectations)

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

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

$$\begin{aligned}
&\mathbb{P}\big\{y(\boldsymbol{u}^*, \boldsymbol{e}, \boldsymbol{v}, x^*) = y^*, y(\boldsymbol{u}, \boldsymbol{e}, \boldsymbol{v}, x) = y\big\} \\
&\qquad = \mathbb{E}\big\{c_y(\boldsymbol{e}, x)c_{y^*}(\boldsymbol{e}, x^*)\big\} + \mathbb{E}\big\{c_y(\boldsymbol{e}, x)c_m(\boldsymbol{e}, x^*)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*)\big\} \\
&\quad + \mathbb{E}\big\{c_m(\boldsymbol{e}, x)c_{y^*}(\boldsymbol{e}, x^*)b_y(\boldsymbol{e}, \boldsymbol{v}, x)\big\} + \mathbb{E}\big\{c_m(\boldsymbol{e}, x)c_m(\boldsymbol{e}, x^*)b_y(\boldsymbol{e}, \boldsymbol{v}, x)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*)\big\} \\
&\qquad = \pi_{mm}\mathrm{Cov}\big\{b_y(\boldsymbol{e}, \boldsymbol{v}, x), b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*) \;\big|\; \boldsymbol{e} \in S_{mm}(x, x^*)\big\} \\
&\qquad\qquad + \mathrm{Cov}\big\{a_y(\boldsymbol{e}, x), a_{y^*}(\boldsymbol{e}, x^*)\big\} + \mathbb{E}a_y(\boldsymbol{e}, x)\mathbb{E}a_{y^*}(\boldsymbol{e}, x^*). \quad (37)
\end{aligned}$$

The stated result follows if one divides the right hand side in (37) 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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}\Big[\big\{c_y(\boldsymbol{e}, x) + c_m(\boldsymbol{e}, x)b_y(\boldsymbol{e}, \boldsymbol{v}, x)\big\}\big\{c_{y^*}(\boldsymbol{e}, x^*) + c_m(\boldsymbol{e}, x^*)b_{y^*}(\boldsymbol{e}, \boldsymbol{v}, x^*)\big\}\Big] \\
&\qquad = \pi_{yy^*}(x, x^*) + \pi_{my^*}(x, x^*)\mathbb{E}\big\{\bar{b}_y(\boldsymbol{e}, x) \;\big|\; \boldsymbol{e} \in S_{my^*}(x, x^*)\big\}
\end{aligned}$$

$$+ \pi_{ym}(x,x^*)\mathbb{E}\big\{\bar{\mathscr{b}}_{y^*}(\boldsymbol{e},x^*) \;\big|\; \boldsymbol{e} \in S_{ym}(x,x^*)\big\}$$
$$+ \pi_{mm}(x,x^*)\mathbb{E}\big\{\mathscr{b}_y(\boldsymbol{e},\boldsymbol{v},x)\mathscr{b}_{y^*}(\boldsymbol{e},\boldsymbol{v},x^*) \;\big|\; \boldsymbol{e} \in S_{mm}(x,x^*)\big\}. \quad (38)$$

Note that $\mathscr{b}_y(e,v,x)$ is unrestricted in $e \in S_m(x)$ but is restricted (i.e. monotonic) in v, albeit that we have the condition $\mathbb{E}\big\{p(\boldsymbol{e},\boldsymbol{v},x) \;\big|\; \boldsymbol{e} \in S_m(x)\big\} = \mu_m(x)$ and $\mathscr{b}_y(e,v,x)$ 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

$$\mathscr{z}_{my^*}^{xy}(x,x^*) = \mathbb{E}\big\{\mathscr{b}_y(\boldsymbol{e},x) \;\big|\; \boldsymbol{e} \in S_{my^*}(x,x^*)\big\},$$

and let $\mathscr{z}$'s with different subscript and superscript combinations be defined analogously. Then the right hand side in (38) reduces to

$$\pi_{yy^*}(x,x^*) + \pi_{my^*}(x,x^*)\mathscr{z}_{my^*}^{xy}(x,x^*) + \pi_{ym}(x,x^*)\mathscr{z}_{ym}^{x^*y^*}(x,x^*)$$
$$+ \pi_{mm}(x,x^*)\mathbb{E}\big\{\mathscr{b}_y(\boldsymbol{e},x)\mathscr{b}_{y^*}(\boldsymbol{e},x^*) \;\big|\; \boldsymbol{e} \in S_{mm}(x,x^*)\big\}. \quad (39)$$

Therefore, we seek to minimize/maximize (39) subject to the (mean) restrictions, namely (dropping the (x,x^*) arguments)

$$\begin{cases} \pi_{ym}\mathscr{z}_{ym}^{x^*y^*} + \pi_{mm}\mathscr{z}_{mm}^{x^*y^*} + \pi_{rm}\mathscr{z}_{rm}^{x^*y^*} = \phi_{y^*}(x^*), \\ \pi_{my^*}\mathscr{z}_{my^*}^{xy} + \pi_{mm}\mathscr{z}_{mm}^{xy} + \pi_{mr^*}\mathscr{z}_{mr^*}^{xy} = \phi_y(x). \end{cases} \quad (40)$$

Since $\mathscr{b}_y(e,x)$ must belong to $[0,1]$, (39) implies that for the lower bound we want to strive to make $\mathscr{b}_y(e,x)\mathscr{b}_{y^*}(e,x^*)$ 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 $\mathscr{b}_y(e,x)$ is binary for all e,x,y. So, we assume that $\mathscr{b}_y(e,x)$ is either one or zero hereafter. Define

$$S_+^y(x) = \{e \in S_m(x) :\ \mathscr{b}_y(e,x) = 1\} \quad \text{and} \quad S_-^y(x) = \{e \in S_m(x) :\ \mathscr{b}_y(e,x) = 0\},$$

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

The optimization of (39) subject to (40) is illustrated in figure 11: 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 11 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 (39), (e.g. $\pi_{my^*}\mathscr{z}_{my^*}^{xy} = b$), because by construction $\mathscr{b}_y(e,x)$ is equal to one when $e \in S_+^y(x)$ and equal to zero, otherwise. The constraints in (40) can likewise be expressed in terms of the probabilities illustrated at the top of figure 11. So, using the shorthand $\phi = \phi_y(x)$ and $\phi^* = \phi_{y^*}(x^*)$, minimizing (39) subject to (40) 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,

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

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 11. Then, maximizing (39) subject to (40) can be formulated as

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

Now,

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

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

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

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 (15).

Further, the second condition in (16) 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 (15) 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. □

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 8. $q_e(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \alpha_{yy^*}(x, x^*)$. Further, the bounds on $q_e(y^* \mid x^*, x, y)$ 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

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

which coincides with (38), except that b in the last term of (38) 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 9.

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

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 $\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \,\big|\, \cdot\big\} \leq \min\big\{\mathbb{E}(\boldsymbol{p} \,|\, \cdot), \mathbb{E}(\boldsymbol{p}^* \,|\, \cdot)\big\}$, 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 $\mathbb{E}\big\{\min(\boldsymbol{p}, \boldsymbol{p}^*) \,\big|\, \cdot\big\} \geq \mathbb{E}(\boldsymbol{p}\boldsymbol{p}^* \mid \cdot)$, where $\boldsymbol{p}, \boldsymbol{p}^*$ are independent given $\boldsymbol{e}, \boldsymbol{e}^*$. □

Theorem 10.

$$
q_{uv}(y^* \mid x^*, x, y) = \tilde{\pi}_{y^*}(x^*) + \delta_y \delta_{y^*} \frac{\pi_m(x) \pi_m(x^*)}{\tilde{\pi}_y(x)} \rho_{uv}(x, x^*), \tag{41}
$$

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

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

Note that

$$\begin{aligned} &\mathbb{P}\{y(\boldsymbol{e}^*,\boldsymbol{u},\boldsymbol{v},x^*) = y^*, y(\boldsymbol{e},\boldsymbol{u},\boldsymbol{v},x) = y\} \\ &\quad = \tilde{\pi}_y(x)\tilde{\pi}_{y^*}(x,x^*) + \\ &\qquad \pi_m(x)\pi_m(x^*)\mathrm{Cov}\{\hbar_y(\boldsymbol{e},\boldsymbol{u},\boldsymbol{v},x), \hbar_{y^*}(\boldsymbol{e}^*,\boldsymbol{u},\boldsymbol{v},x^*) \mid \boldsymbol{e} \in S_m(x),\ \boldsymbol{e}^* \in S_m(x^*)\}. \quad (42) \end{aligned}$$

Tedious but simple algebra shows that the covariance in (42) equals $\delta_y\delta_{y^*}\rho_{uv}(x,x^*)$, which yields (41).

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 $\mathbb{E}\{\min(\boldsymbol{p},\boldsymbol{p}^*) \mid \cdot\} \le \min\{\mathbb{E}(\boldsymbol{p} \mid \cdot), \mathbb{E}(\boldsymbol{p}^* \mid \cdot)\}$, with equality for $p(e,v,x) = \mathbb{1}\{v \ge 1 - \mu_m(x)\}$. For the lower bound,

$$\mathbb{E}\{\min(\boldsymbol{p},\boldsymbol{p}^*) \mid \boldsymbol{e},\boldsymbol{e}^*\} \ge \mathbb{E}(\boldsymbol{p}\boldsymbol{p}^* \mid \boldsymbol{e},\boldsymbol{e}^*) \ge \mathbb{E}(\boldsymbol{p} \mid \boldsymbol{e})\mathbb{E}(\boldsymbol{p}^* \mid \boldsymbol{e}^*) \quad (43)$$

where the inequalities hold with equality for $p(e,v,x) = \mathbb{1}\{e \in S_m(x)\}$. Take the expectations in (43) 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 11.

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

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

$$\begin{aligned} \pi_{yy^*}(x,x^*) &+ \pi_{ym}(x,x^*)\mathbb{E}\{p(\boldsymbol{e},\boldsymbol{v}^*,x^*) \mid \boldsymbol{e} \in S_{ym}(x,x^*)\} \\ &+ \pi_{my^*}(x,x^*)\mathbb{E}\{p(\boldsymbol{e},\boldsymbol{v},x) \mid \boldsymbol{e} \in S_{my^*}(x,x^*)\} \\ &+ \pi_{mm}(x,x^*)\mathbb{E}\big[\min\{p(\boldsymbol{e},\boldsymbol{v},x), p(\boldsymbol{e},\boldsymbol{v}^*,x^*)\} \mid \boldsymbol{e} \in S_{mm}(x,x^*)\big]. \quad (44) \end{aligned}$$

If $x = x^*$, the second and third terms in (44) are equal to zero, so we only need to consider the last term. The upper bound follows from the fact that

$$\mathbb{E}\{\min\{\boldsymbol{p},\boldsymbol{p}^*\} \mid \cdot\} \le \min\{\mathbb{E}(\boldsymbol{p} \mid \cdot), \mathbb{E}(\boldsymbol{p}^* \mid \cdot)\}. \quad (45)$$

For the lower bound, note that by the Jensen inequality

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

If $x \neq x^*$ then (45) 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 (44) is bounded above by $\pi_{yy^*}(x, x^*) + U_{yy^*}(x, x^*)$, where U_{yy^*} is as defined in (12). Sharpness follows from the fact that (44) is bounded below by (38). For the lower bound, note that $\mathbb{E}\big\{\min\{\boldsymbol{p}, \boldsymbol{p}^*\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big\} \geq \mathbb{E}\big\{\bar{\ell}_y(\boldsymbol{e}, x)\bar{\ell}_{y^*}(\boldsymbol{e}, x^*) \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big\}$, which is identical to the second factor of the last term in (38), except that ℓ_y is replaced with $\bar{\ell}_y$. □

Theorem 12.

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

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

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

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

$$\pi_{yy^*}(x, x^*) + \pi_{ym}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x^*) \mid \boldsymbol{e} \in S_{ym}(x, x^*)\}$$
$$+ \pi_{my^*}(x, x^*)\mathbb{E}\{p(\boldsymbol{e}, \boldsymbol{v}, x) \mid \boldsymbol{e} \in S_{my^*}(x, x^*)\}$$
$$+ \pi_{mm}(x, x^*)\mathbb{E}\big[\min\{p(\boldsymbol{e}, \boldsymbol{v}, x), p(\boldsymbol{e}, \boldsymbol{v}, x^*)\} \,\big|\, \boldsymbol{e} \in S_{mm}(x, x^*)\big]. \quad (46)$$

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 $\mathbb{E}\big\{p(\boldsymbol{e}, \boldsymbol{v}, x^*) \,\big|\, \boldsymbol{e} \in S_m(x^*)\big\} = \mu_m(x^*)$, (46) is bounded above by $\pi_{yy^*}(x, x^*) + U_{yy^*}(x, x^*)$, where U_{yy^*} is as defined in (12). For sharpness, note that (46) is bounded below by (38).

Now the lower bound. Since (46) is no less than (38), the lower bound for (38) is a lower bound here, also. It remains to be shown that it can be attained. Recall from the proof of theorem 5 that (38) 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, (38) equals (46). □

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.

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

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

$$\begin{aligned}a_y^*(e,x) &= c_y(e,x)+c_m(e,x)\mathbb{E}b_y^*(e,\boldsymbol{v},x)\\ &= \begin{cases}c_y(e,x)+c_m(e,x)\mu_m(x), & y=(1,0),\\ c_y(e,x)+c_m(e,x)\{1-\mu_m(x)\}, & y=(0,1),\\ c_y(e,x), & y\in\{(0,0),(1,1)\},\end{cases}\end{aligned}\tag{47}$$

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

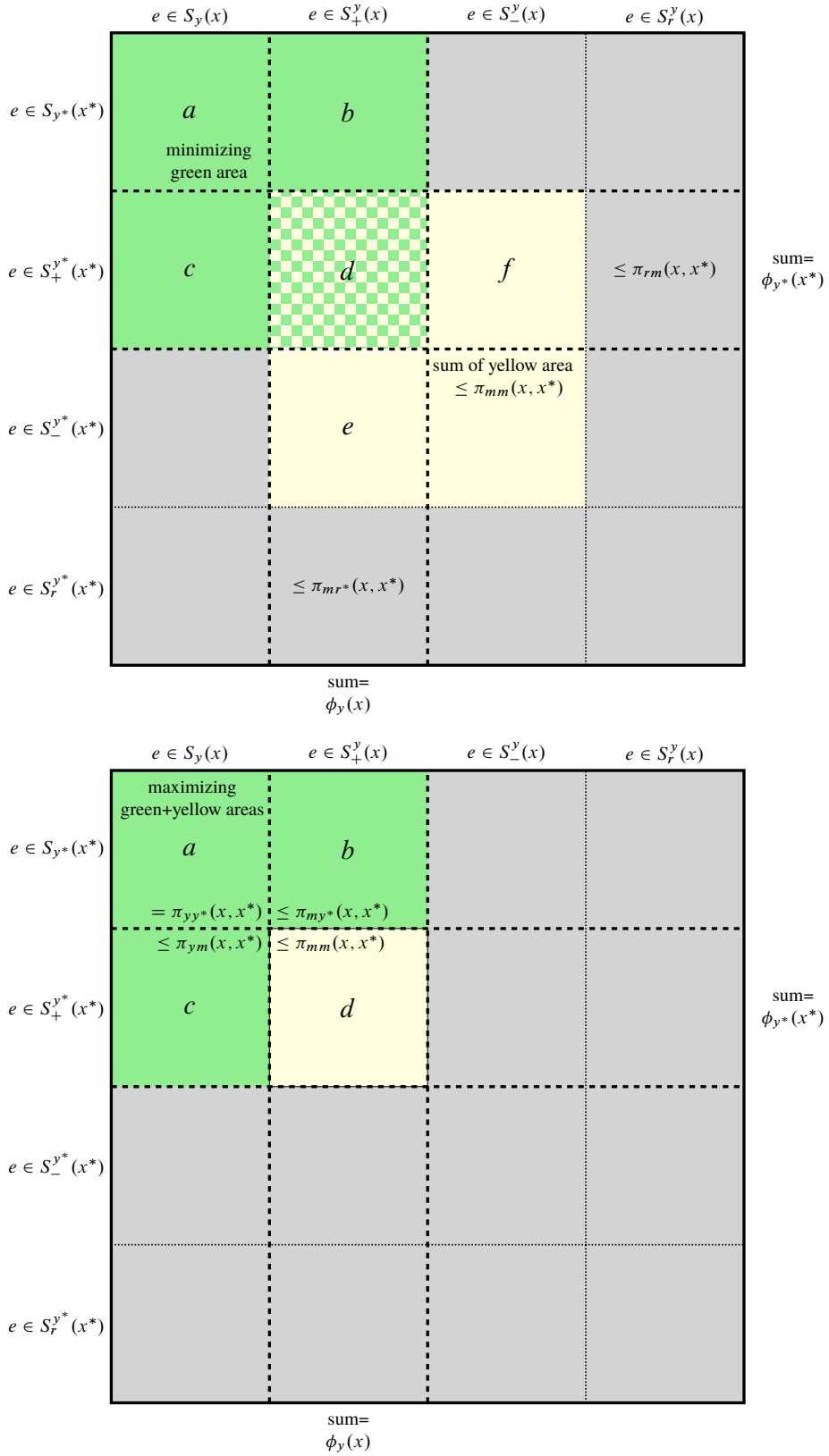

Figure 11: Finding the lower bound (top) and the upper bound (bottom)

References

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

Aradillas-Lopez, A. (2011): "Nonparametric probability bounds for Nash equilibrium actions in a simultaneous discrete game," *Quantitative Economics*, 2(2), 135–171.

Aradillas-López, A., and E. Tamer (2008): "The identification power of equilibrium in simple games," *Journal of Business & Economic Statistics*, 26(3), 261–283.

Athey, S., and G. W. Imbens (2007): "Discrete choice models with multiple unobserved choice characteristics*," *International Economic Review*, 48(4), 1159–1192.

Aumann, R. J. (1961): "Borel structures for function spaces," *Illinois Journal of Mathematics*, 5(4), 614–630.

Bajari, P., C. L. Benkard, and J. Levin (2007): "Estimating dynamic models of imperfect competition," *Econometrica*, 75(5), 1331–1370.

Bajari, P., H. Hong, and S. P. Ryan (2010): "Identification and estimation of a discrete game of complete information," *Econometrica*, 78(5), 1529–1568.

Baker, A. (2013): "Simplicity," in *Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta. Fall 2013 edn.

Bjorn, P. A., and Q. H. Vuong (1984): "Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation," Discussion paper, CalTech.

Bresnahan, T. F., and P. C. Reiss (1991a): "Empirical models of discrete games," *Journal of Econometrics*, 48(1), 57–81.

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

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

Bulow, J. I., J. D. Geanakoplos, and P. D. Klemperer (1985): "Multimarket oligopoly: Strategic substitutes and complements," *Journal of Political economy*, 93(3), 488–511.

Cass, D., and K. Shell (1983): "Do sunspots matter?," *Journal of Political Economy*, 91(2), 193–227.

Chen, X., and D. Pouzo (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80(1), 277–321.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 1791–1828.

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

Galichon, A., and M. Henry (2011): "Set identification in models with multiple equilibria," *The Review of Economic Studies*, p. rdr008.

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

Grieco, P. L. (2014): "Discrete games with flexible information structures: An application to local grocery markets," *Rand Journal of Economics*, 45(2), 303–340.

Heckman, J. J. (2005): "The scientific model of causality," *Sociological methodology*, 35(1), 1–97.

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

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

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

Kalai, A., and E. Kalai (2012): "Cooperation in strategic games revisited," *Quarterly Journal of Economics*, p. qjs074.

Kashaev, N. (2015): "Testing for Nash behavior in entry games with complete information," Discussion paper, Penn State.

Kashaev, N., and B. Salcedo (2015): "Identification of solution concepts for discrete semi-parametric games with complete information," Discussion paper, Penn State.

Kasy, M. (2011): "Identification in triangular systems using control functions," *Econometric Theory*, 27(03), 663–671.

Kline, B. (2015): "Identification of complete information games," *Journal of Econometrics*, 189(1), 117–131.

Kline, B., and E. Tamer (2012): "Bounds for best response functions in binary games," *Journal of Econometrics*, 166(1), 92–105.

Kooreman, P. (1994): "Estimation of econometric models of some discrete games," *Journal of Applied Econometrics*, 9(3), 255–268.

Liu, N., Q. Vuong, and H. Xu (2013): "Rationalization and identification of discrete games with correlated types," Discussion paper, University of Texas.

Magnolfi, L., and C. Roncoroni (2016): "Estimation of discrete games with weak assumptions on information," Discussion paper, Yale.

Manski, C. (2015): "Interpreting point predictions," Discussion paper, Northwestern University.

Pakes, A., M. Ostrovsky, and S. Berry (2007): "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)," *The Rand Journal of Economics*, 38(2), 373–399.

Reguant, M. (2016): "Bounding equilibria in counterfactual analysis," Discussion paper, Northwestern University.

Schennach, S. M. (2014): "Entropic latent variable integration via simulation," *Econometrica*, 82(1), 345–385.

Seim, K. (2006): "An empirical model of firm entry with endogenous product–type choices," *Rand Journal of Economics*, 37(3), 619–640.

Soetevent, A. R., and P. Kooreman (2007): "A discrete-choice model with social interactions: with an application to high school teen behavior," *Journal of Applied Econometrics*, 22(3), 599–624.

Song, K. (2014): "Point decisions for interval–identified parameters," *Econometric Theory*, 30(02), 334–356.

Tamer, E. (2003): "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 70(1), 147–165.

Teh, Y. W. (2010): "Dirichlet process," in *Encyclopedia of machine learning*, pp. 280–287. Springer.

Xu, H. (2014): "Estimation of discrete games with correlated types," *The Econometrics Journal*, 17(3), 241–270.