COUNTERFACTUALS WITH LATENT INFORMATION

By

Dirk Bergemann, Benjamin Brooks, and Stephen Morris

January 2019

Revised March 2021

COWLES FOUNDATION DISCUSSION PAPER NO. 2162R2

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Counterfactuals with Latent Information*

Dirk Bergemann[†] Benjamin Brooks[‡] Stephen Morris[§]

March 31, 2021

We describe a methodology for making counterfactual predictions when the information held by strategic agents is a latent parameter. The analyst observes behavior which is rationalized by a Bayesian model, in which agents maximize expected utility, given partial and differential information about payoff-relevant states of the world, represented as an information structure. A counterfactual prediction is desired about behavior in another strategic setting, under the hypothesis that the distribution of the state and agents' information about the state are held fixed. When the data and the desired counterfactual prediction pertain to environments with finitely many states, players, and actions, there is a finite dimensional description of the sharp counterfactual prediction, even though the latent parameter, the information structure, is infinite dimensional.

Keywords: Counterfactuals, Bayes correlated equilibrium, information structure, linear program.

JEL Classification: C72, D44, D82, D83.

[*] We acknowledge financial support through NSF Grants SES 1459899, SES 1824137 and SES 2049744. We are grateful for valuable discussions with Stéphane Bonhomme, Jeff Ely, Hanming Fang, Phil Haile, Tibor Heumann, Peter Hull, Lorenzo Magnolfi, Eduardo Souza-Rodrigues, Philipp Strack, Vasilis Syrgkanis, and Elie Tamer; and excellent research assistance from Aroon Narayanan and Roi Orzach.

[†] Department of Economics, Yale University, dirk.bergemann@yale.edu

[‡] Department of Economics, University of Chicago, babrooks@uchicago.edu

[§] Department of Economics, Massachusetts Institute of Technology, semorris@mit.edu

1 Introduction

When making counterfactual predictions in the presence of incomplete information, it is standard for the analyst to assume that the information structure is known. When there is strategic interaction, the relevant information structure must specify what economic agents, or players, know about payoff-relevant variables, but also what they know about other players' information. This may be an unrealistic assumption that the analyst would like to relax. We provide a characterization of counterfactual predictions that an analyst could make when they do not know the information structure but want to hold the (unknown) information structure fixed.

We implement a completely non-parametric approach to the identifying counterfactuals by treating the information structure as a nuisance parameter—that is, a parameter which is not of intrinsic interest but needs to be accounted for in making counterfactual predictions. We thus avoid the complexity that would be involved in trying to identify the information structure. In particular, when there is more than one player, the set of belief hierarchies that arise in some information structure is an infinite dimensional vector space that is known as the universal type space. By focusing on a given counterfactual of interest and finite dimensional restrictions on information implicit in observed behavior, we are able to derive counterfactual predictions that are finite dimensional and characterized by a finite number of linear inequalities, as long as the underlying action and state spaces are also finite.

Let us give a semi-formal description of our approach for the special case of a single-player games, i.e., decision problems. Suppose that an individual takes an action a from a finite set of possible actions. The action results in a payoff $u(a, \theta)$, where θ is a possibly uncertain state of the world that has finitely many possible values. As θ is uncertain, so too may be the individual's information about θ. But we maintain the standard assumption that the action maximizes expected utility, given whatever interim beliefs about θ are held at the time the decision was made.

Over a long period of time or across a large population of representative individuals, and with suitable ergodicity assumptions, what may be observable is the *distribution* of outcomes, e.g., the actions that were taken. In fact, for the purposes of the current discussion, let us further suppose that a and θ are both observable ex post, so that the joint distribution thereof, denoted $\phi(a, \theta)$, can be estimated. (We allow very general forms for the data in our main theorem.) What is not observable is what was *known* about θ at the time the action was taken.

This information may be canonically expressed as an *experiment* in the sense of Blackwell (1951): the individual observes a signal s, and the distribution of s conditional on θ is known to follow a conditional probability law $\pi(s|\theta)$. By combining this experiment with the observed prior over θ and applying Bayes' rule, we would obtain the distribution of the decision maker's interim beliefs. An experiment rationalizes the observed data if an expected utility maximizer who observed the signal would optimally behave in a way that results in the observed joint distribution ϕ.

Now suppose we wish to predict how the same representative individual will behave in a new decision problem, where an action $\widehat{a}$ leads to a payoff $\widehat{u}(\widehat{a}, \theta)$. Importantly, we shall assume that while the decision problem changes, the distribution of θ and the information (i.e., the Blackwell experiment) remain the same.[1] The distribution of θ can be computed directly from the joint distribution ϕ, but the experiment, i.e., the set of signals and the conditional distribution π, is a latent parameter. The question is: which joint distributions $\widehat{\phi}(\widehat{a}, \theta)$ could be induced by optimal behavior for some experiment which also rationalizes the observed data ϕ? This counterfactual prediction can then be used to test a model of preferences, do welfare analysis, etc.

A direct approach would be to first compute the set of experiments which can rationalize ϕ, and then for each such experiment, compute the optimal strategies and the resulting $\widehat{\phi}$ in the counterfactual. But the signals in a Blackwell experiment are an abstract set, and we

[1] The assumption of a fixed prior distribution over θ is essentially without loss of generality; as discussed in Section 8.5 .

have not even assumed a particular space in which these signals should live. And as noted above, the set of information structures with multiple players is vastly more complex.

Instead, we proceed directly to counterfactual predictions. This can be done as follows. Imagine that rather than performing an abstract thought experiment, the individual actually did choose $\widehat{a}$ at the same time as a was chosen, and we simply did not observe it. The payoff was simply the sum of the payoffs across the two decision problems, so that there was no interaction between the two choices except through the common information. Moreover, since both actions were taken based on the same information about the same state, there will be correlation between θ, a, and $\widehat{a}$, and we can write $\overline{\phi}\,(a, \widehat{a}, \theta)$ for the joint distribution of these objects. We could even conceptualize there being a single *linked decision problem*, in which the action is an ordered pair $(a, \widehat{a})$. If $\overline{\phi}$ is to be consistent with our data, the marginal of $\overline{\phi}$ on (a, θ) must be ϕ. The counterfactual prediction $\widehat{\phi}$ is simply the marginal of $\overline{\phi}$ on $(\widehat{a}, \theta)$.

Thus, the problem of computing counterfactual predictions can be reduced to computing those outcomes $\overline{\phi}$ for the linked game which are consistent with Bayesian rationality. But this problem has already been solved. Bergemann and Morris (2013, 2016) have shown that the solution concept of Bayes correlated equilibrium (BCE) (specialized to single-player games) describes precisely those joint distributions of actions and states which are consistent with optimal behavior with respect to some information, and corresponds to a convex set of $\overline{\phi}$ that satisfy a finite collection of “obedience constraints,” which represent Bayesian optimality.

When we add in the constraint that the marginal on (a, θ) is the observed ϕ, we obtain a convex polytope of joint distributions $\overline{\phi}$ on $(a, \widehat{a}, \theta)$ which are consistent with rationality in the linked decision problem and are consistent with the data. The set of possible counterfactual outcomes can then be obtained as the marginals on $(\widehat{a}, \theta)$. The net result is that the counterfactual prediction is a convex polytope, which can be described by a finite collection of linear constraints, as long as the underlying action and state spaces are finite and the data is characterized by linear restrictions. These constraints can then be used, for example, to compute counterfactual welfare outcomes by simply solving linear programs.

The discussion thus far considers the special case of a one-player finite-action game (i.e., a decision problem) where the distribution over fundamentals is observed. Our subsequent analysis shows that this logic goes through in general many-player finite-action games, where arbitrary data about behavior and fundamentals is revealed in the form of linear constraints on ϕ. For example, it might be that the fundamental θ is not observed but the distribution of actions is observed. Or it might be that only some statistics of players' actions is observed, such as the winning bid in an auction. The BCE characterization of outcomes that could arise with any information structure holds for arbitrary many-player games. The argument that the set of feasible counterfactuals is characterized by a set of linear inequalities is completely general, and follows the same steps as our informal discussion above and is stated in our main Theorem 2.[2]

The key assumption in our analysis is that information is fixed: the same information structure that generated the data is assumed to be the information structure in the counterfactual. An alternative approach would be to analyze *variable information counterfactuals*, where the analyst does not know the information structure that generated the data and believes that a completely different information structure may be relevant for the counterfactual. This alternative variable information approach can also utilize the BCE characterization of Bergemann and Morris (2013, 2016), and was discussed in Bergemann and Morris (2013) and Bergemann, Brooks, and Morris (2017). Naturally, the latter approach will provide more robust but weaker predictions. The fixed information approach is perhaps more consistent with the meaning of counterfactual (what happens if I change one things leaving all others the same)[3], but in a specific applications there will be context specific arguments

[2] The proof of Theorem 2 appeals to a more general result on joint predictions with fixed information (Theorem 1) that is of independent interest. We discuss this result at the end of the introduction and in the concluding Section 7.

[3] Lewis (1973) argues that the natural meaning of the counterfactual "if X, then Y" is that Y is true in the counterfactual world that is as close as possible to the true world but where X was true.

for one approach or the other, as we briefly discuss in Section 7.3. For comparison, we formally define variable information counterfactuals in Section 3.3 and discuss the difference in examples.

A growing number of empirical papers use the BCE characterization in developing informational robust identification and counterfactuals. The early works of Syrgkanis, Tamer, and Ziani (2018) and Magnolfi and Roncoroni (2020) considered variable information counterfactuals in the context of auctions and entry games respectively. Following earlier versions of this paper, Magnolfi and Roncoroni (2020) have added a comparison with fixed information counterfactuals and Gualdani and Sinha (2020) maintain the fixed counterfactuals assumption in their application to a single agent discrete choice model of voting. Canen and Song (2020) introduce a decomposition approach used frequently in labor economics to offer counterfactual predictions in strategic settings using BCE as solution concept. This scope of this paper is limited to the question of partial identification of counterfactuals, when the "data" is in the form of an exact moment (or moments) of the population outcome. We do not address the fundamental issue of how to conduct inference on the counterfactual prediction when one only has access to a noisy estimate derived from a finite data set, which of course plays a central role in the empirical papers described above.

Following our main result, we report some extensions and examples to illustrate the adaptability of the approach and provide intuition. Our main result is presented in a stark theoretical benchmark where the analyst knows nothing about players' information ex ante. There are many natural settings where the analyst knows more than in our benchmark result but less than in standard parameterized models of the information structure. In Section 4, we show how partial knowledge of the information structure will enable the analyst to narrow the set of possible counterfactual predictions. As examples, the analyst might be confident that players knows those features of the state that affect their own preferences, which we refer to as the case of private values; or the analyst may observe payoff shifters for each player, observed by that player and the analyst, but known not to be observed by other players and independent across players. We believe that empirical researchers will want to impose restrictions of this kind. Indeed Syrgkanis, Tamer, and Ziani (2018) present results in a private value environment and Magnolfi and Roncoroni (2020) exploit payoff shifters.

We report two examples to illustrate the logic of our approach and visualize the approach in simple models in the empirical literature. First, we discuss how Roy's (1951) model of self selection can be mapped into our framework. Although there are some special features of the one-player case (which we discuss), most features of the many-player case are clearly illustrated when in the single player case. In particular, we illustrate how fixed-information counterfactuals refine variable-information counterfactuals, how more limited data leads to a larger counterfactual prediction, and how payoff shifters can shrink the counterfactual prediction. We also analyze a two player two action game with strategic substitutes. This example maps into the model of market entry studied in the empirical literature and, in particular, in the empirical analysis of Magnolfi and Roncoroni (2020). As well as illustrating in a strategic problem the power of the fixed-information approach in refining counterfactuals, we also illustrate the role of private values in refining counterfactuals, and show how with many players, counterfactuals may not be tight even in the original game. Because the examples are somewhat involved, we postpone their development to the last section of the paper, but we will preview in our theoretical development the features that will be illustrated by the examples.

Before proving our main result, we first solve a foundational result concerning joint predictions when the players engage in K games simultaneously with a fixed information structure. We formalize this idea of a large "linked game" and the notion of a joint prediction in Theorem 1. Our main counterfactual result then follows when we consider just two games, one of which corresponds to the observed game and a counterfactual game. We also appeal the joint predictions result in subsequent results tightening counterfactuals in Section 4. We believe it can be used for other purposes, and in Section 7.2, we describe how it can be used to derive "informationally robust" comparative statics in a fixed game. Heumann (2019) has analyzed this problem in a class of symmetric games with normal uncertainty and linear best responses. The methods and results are complementary. By solving for general games and non-local counterfactuals, we describe an approach that can be most easily adapted for empirical work incorporating uncertainty about fundamentals, all while maintaining the structure of the linear program. It will, however, be hard to prove general analytic results. The extra structure of normal distributed uncertainty and linear-best-response games means that it is possible to derive interpretable analytic results regarding local behavior. However, the structural assumptions are heavily exploited and the extension to more general settings may be more difficult.

The rest of this paper proceeds as follows. Section 2 establishes the basic notation. Section 3 presents our notion of a counterfactual prediction and our main theorem characterizing the set of counterfactuals consistent with data. In Section 4, we discuss how one can add more structure to information and preferences to tighten the counterfactual prediction. Section 5 illustrates our results with the classic Roy model as an example of a single player game. Section 6 considers the entry game as widely analyzed two player binary action game. Section 7 concludes the paper with a discussion of our assumptions and directions for future work. In a Supplemental Appendix, we report further analysis of the two examples in the paper, and also two more applications (two-player zero-sum games and private-value auctions).

2 Preliminaries

The game consists of N players, indexed by $i = 1, \ldots, N$. Players' preferences depend on a state of the world $\theta \in \Theta$, where Θ is a finite set. The players and state space will be held fixed throughout our analysis.

The players interact through a *base game*, denoted $\mathcal{G}$, which consists of the following objects. Each player has finite set of actions A_i for each player, with $A = \times_{i=1}^{N} A_i$ denoting the set of action profiles. Players are expected utility maximizers, and preferences are represented by utility indices $u_i : A \times \Theta \to \mathbb{R}$. Thus $\mathcal{G} = (A_i, u_i)_{i=1}^N$.

Players' information about the state is represented with a common-prior information structure, denoted by $\mathcal{I}$, which consists of the following objects. Each player has a measurable set of signal*s* S_i, with $S = \times_{i=1}^N S_i$ denoting the set of signal profiles, and there is a conditional probability measure $\pi : \Theta \to \Delta(S)$ over signal profiles as a function of the state. Thus $\mathcal{I} = \left((S_i)_{i=1}^N, \pi \right)$.[4]

The *prior distribution* over states is denoted $\mu \in \Delta(\Theta)$.

A *Bayesian game* is a tuple $(\mu, \mathcal{G}, \mathcal{I})$. A *strategy* for player i in the Bayesian game is a measurable mapping $\sigma_i : S_i \to \Delta(A_i)$. We write $\sigma_i(a_i|s_i)$ for the probability of an action a_i given the signal s_i. A strategy profile is denoted $\sigma = (\sigma_1, \ldots, \sigma_N)$ and is associated with the product mapping $\sigma : S \to \Delta(A)$, where $\sigma(a|s) = \times_{i=1}^N \sigma_i(a_i|s_i)$. Player i's expected utility under the strategy profile σ is

$$U_i(\sigma) = \sum_{\theta \in \Theta} \int_{s \in S} \sum_{a \in A} u_i(a, \theta)\, \sigma(a|s)\, \pi(ds|\theta)\, \mu(\theta).$$

A strategy profile σ is a *Nash equilibrium* if $U_i(\sigma) \geq U_i(\sigma_i', \sigma_{-i})$ for all i and for all strategies σ_i'.

An *outcome* of a Bayesian game is a distribution $\phi \in \Delta(A \times \Theta)$. Note that the outcome contains all the information required in order to compute players' payoffs or any Bayesian welfare criterion that only depends on realized actions and states. The outcome ϕ is *induced* by a strategy profile σ in Bayesian game $(\mu, \mathcal{G}, \mathcal{I})$ if

$$\phi(a, \theta) = \int_{s \in S} \sigma(a|s)\, \pi(ds|\theta)\, \mu(\theta).$$

[4] Note that we allow the information structure to be infinite, while the other objects in the model are finite. This richness is necessary to accommodate the full range of possible higher order beliefs and correspondingly the full range of equilibrium behavior across all counterfactuals.

An outcome ϕ is a *Bayes correlated equilibrium (BCE)* of the *base game* $\mathcal{G}$ if the following *obedience constraints* are satisfied: for all i, a_i, and a_i',

$$\sum_{\theta \in \Theta} \sum_{a_{-i} \in A_{-i}} \left(u_i\left(a_i, a_{-i}, \theta\right) - u_i\left(a_i', a_{-i}, \theta\right)\right) \phi\left(a_i, a_{-i}, \theta\right) \geq 0. \tag{1}$$

It is a theorem of Bergemann and Morris (2013, 2016) that ϕ is a BCE of $\mathcal{G}$ if and only if there exists a prior μ, an information structure $\mathcal{I}$, and a Nash equilibrium σ of $(\mu, \mathcal{G}, \mathcal{I})$ such that ϕ is induced by σ. We will write BCE$(\mu, \mathcal{G})$ for the set of BCE of $(\mu, \mathcal{G})$.

3 Joint Predictions and Counterfactuals

Before developing our results on counterfactuals, we address the following question: Suppose that the players were to simultaneously play K games, $\mathcal{G}^1, ..., \mathcal{G}^K$. What joint predictions can we make about the outcomes that would arise in equilibria of the respective games, assuming that the players have the same information structure in every game? We have already observed that the set of outcomes that could arise in game $\mathcal{G}^k$ is BCE$\big(\mu, \mathcal{G}^k\big)$. But because players have the same information in each of the K games, there will be additional consistency requirements across the games arising from fixed information. The purpose of this section is to characterize those extra restrictions. Our main result regarding counterfactual behavior and other later results will be established using this characterization when we impose particular structure or interpretations on the games $\mathcal{G}^k$.

3.1 Joint Predictions with Fixed Information

We say that an outcome profile

$$\left(\phi^1, ..., \phi^K\right) \in \Delta\left(A^1 \times \Theta\right) \times \cdots \times \Delta\left(A^K \times \Theta\right)$$

is a *joint prediction* if there exists an information structure $\mathcal{I}$ such that, for each $k = 1, \dots, K$, there is a Nash equilibrium σ^k of $\left(\mu, \mathcal{G}^k, \mathcal{I}\right)$ that induces ϕ^k.

We denote by $\overline{\mathcal{G}}$ the following *linked game*, where player i's actions are $\overline{A}_i = A_i^1 \times \cdots \times A_i^K$, and payoffs are given by

$$\overline{u}_i\left(\overline{a}, \theta\right) = \sum_{k=1,\dots,K} u_i^k\left(a^k, \theta\right),$$

where $\overline{a}_i = \left(a_i^1, ..., a_i^K\right)$ for all i. We refer to $\mathcal{G}^k$ as the *component games* of $\overline{\mathcal{G}}$. A Bayes correlated equilibrium $\overline{\phi}$ of $\left(\mu, \overline{\mathcal{G}}, \mathcal{I}\right)$ can be identified with a joint distribution in $\Delta\left(A^1 \times .... \times A^K \times \Theta\right)$.

Theorem 1 (Joint Predictions)**.**
An outcome profile $\left(\phi^1, ..., \phi^K\right)$ is a joint prediction for the component games $\mathcal{G}^1, ..., \mathcal{G}^K$ if and only if there exists a Bayes correlated equilibrium $\overline{\phi}$ of the linked game $\overline{\mathcal{G}}$ for which the marginal of $\overline{\phi}$ on $A^k \times \Theta$ is ϕ^k for each k.

Proof of Theorem 1. Fix an information structure $\mathcal{I}$. Any strategy profile $\overline{\sigma}$ in the linked game can be identified with a profile of strategy profiles $\sigma^k = \left(\sigma_1^k, ..., \sigma_N^k\right)$ in the component games, where $\sigma_i^k\left(\cdot|s_i\right)$ is the marginal of $\overline{\sigma}\left(\cdot, \cdot|s_i\right)$ on A_i^k.

Claim: $\bar{\sigma}$ is a Nash equilibrium of $\left(\mu, \overline{\mathcal{G}}, \mathcal{I}\right)$ if and only if each σ^k is a Nash equilibrium of $\left(\mu, \mathcal{G}^k, \mathcal{I}\right)$. This follows from the identity:

$$\begin{aligned}
\overline{U}_i\left(\overline{\sigma}\right) &= \sum_{\theta \in \Theta} \int_{s \in S} \sum_{\overline{a} \in \overline{A}} \bar{u}_i\left(\overline{a}, \theta\right) \overline{\sigma}\left(\overline{a}|s\right) \pi\left(ds|\theta\right) \mu\left(\theta\right) \\
&= \sum_{\theta \in \Theta} \int_{s \in S} \sum_{k=1,\dots,K} \left[\sum_{a \in A} u_i^k\left(a, \theta\right) \sigma^k\left(a|s\right)\right] \pi\left(ds|\theta\right) \mu\left(\theta\right) \\
&= \sum_{k=1,\dots,K} U_i^k\left(\sigma^k\right).
\end{aligned}$$

Thus, if $\bar{\sigma}$ is not a Nash equilibrium, then there exists i and a linked game strategy $\overline{\tau}_i$ such that

$$\begin{aligned}\sum_{k=1,\dots,K} U_i^k\left(\sigma^k\right) &= \overline{U}_i\left(\overline{\sigma}\right) \\ &< \overline{U}_i\left(\bar{\tau}_i, \overline{\sigma}_{-i}\right) \\ &= \sum_{k=1,\dots,K} U_i^k\left(\tau_i^k, \sigma_{-i}\right),\end{aligned}$$

where τ_i^k is the marginal of $\overline{\tau}_i$ on A_i^k. Thus, for at least one k, τ_i^k is a profitable deviation in $\left(\mu, \mathcal{G}^k, \mathcal{I}\right)$. Similarly, if there is a profitable deviation in one of the component games, say to τ_i^k for player i in $\mathcal{G}^k$, then the product strategy defined by $\overline{\tau}_i\left(a_i|s_i\right) = \tau_i\left(a_i^k|s_i\right)\overline{\sigma}_i\left(a_i^{-k}|s_i\right)$ is a profitable deviation in the linked game. □

Remark 1. There is an important theoretical subtlety about our notion of fixed-information joint predictions and Theorem 1. An information structure can always be decomposed into "higher-order belief" signals, representing differences in beliefs and higher-order beliefs about Θ; and correlation devices, signals that allow perhaps quite complicated correlation of the actions among the players but without changing their higher-order beliefs (see Liu (2015) for a description of this decomposition). It is always possible to construct equilibria where players' behavior does not depend on correlating devices. This means that any outcome in games $1, 2, ..., K-1$ are consistent with the existence of a correlation device in $\mathcal{I}$ that is only used in the play of game $\mathcal{G}^K$. The set of outcomes that can arise from Bayes Nash equilibrium if there is a correlating device is the set of belief-invariant Bayes correlated equilibria (Liu, 2015; Bergemann and Morris, 2017). An implication is that the set of joint predictions also characterizes the set of profiles of belief-invariant BCE outcomes that could arise for a fixed information structure.

3.2 Counterfactuals when Information is Latent and Fixed

We will use the preceding framework to study counterfactual predictions. An analyst has partial information about play in an *observed game* $\mathcal{G}$. The sets of possible states, actions, and players' payoff functions are known, but the prior μ, the information structure $\mathcal{I}$, and the players' strategies are unknown.[5] Given the partial information revealed about μ and the information structure $\mathcal{I}$, the analyst would like to make counterfactual predictions for what might have happened if the *unobserved game* $\widehat{\mathcal{G}}$ were played. But in making the counterfactual prediction, the analyst wants to assume that the unobserved μ and $\mathcal{I}$ remain the same. We extend the convention that objects without accents correspond to the (partially) observed game and objects accented with a circumflex correspond to the unobserved game. For example, we denote outcomes for the two games by ϕ and $\widehat{\phi}$, respectively.

There is data on behavior in the observed game. We want to predict behavior in the unobserved game. We suppose that the only data that is available, and the only prediction that is desired, pertains to the outcomes. All we know is that the outcome ϕ (i) lies in a set of moment conditions $M \subseteq \Delta(A \times \Theta)$, (ii) it was generated under some prior μ and information structure $\mathcal{I}$, and (iii) it was induced by a Nash equilibrium of $(\mu, \mathcal{G}, \mathcal{I})$. We ask which outcomes $\widehat{\phi}$ could be induced by some equilibrium of $\left(\mu, \widehat{\mathcal{G}}, \mathcal{I}\right)$?

Formally, an outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta\right)$ is a *counterfactual prediction* if there exist μ, $\mathcal{I}$, and Nash equilibria σ and $\widehat{\sigma}$ of $(\mu, \mathcal{G}, \mathcal{I})$ and $\left(\mu, \widehat{\mathcal{G}}, \mathcal{I}\right)$, respectively, such that the outcome ϕ induced by σ is in M and such that $\widehat{\phi}$ is induced by $\widehat{\sigma}$. In other words, $\widehat{\phi}$ can be rationalized as an equilibrium outcome for some $(\mu, \mathcal{I})$, such that $(\mu, \mathcal{I})$ can also rationalize an equilibrium outcome in M. The set of counterfactual predictions is denoted $\widehat{\Phi}(M)$, where we emphasize the dependence on the conditions M.

There are various possible specifications for M, which represent different kinds of observed data. For example:

[5] The assumption that the sets of possible states and actions, and players' payoff functions are known is without loss of generality: if they were was uncertainty, we could construct a larger model that represented that uncertainty. This is discussed in Section 8.5.

1. $M = \{\phi\}$ for some particular ϕ. This corresponds to the case described in the introduction, where the joint distribution of states and actions is observed. It is only the information structure $\mathcal{I}$ that is a latent parameter that we wish to identify from the data.

2. $M = \{\phi \in \Delta(A \times \Theta) \,|\text{marg}_A \phi = \psi\}$ for some $\psi \in \Delta(A)$. In this case, the joint distribution of actions is known, but both information and the distribution of θ are latent variables.

3. $M = \{\phi \in \Delta(A \times \Theta) \,|\text{marg}_A \phi \in \Psi\}$ for some $\Psi \subseteq \Delta(A)$. In this case, we do not even observe the entire distribution of the players' actions. For example, it could be that only some statistic, such as the average action or the highest action is observed.

In our examples in Sections 5 and 6, we mostly focus on the case 1 where the whole distribution on actions and states is observed in the data; this corresponds to the "purest" instance of our problem, in which players' information is the only latent variable that needs to be identified. We note that 1–3 are only examples of data restrictions, and much richer applications are possible. In the context of the Roy model, we consider a case that is particularly relevant in applications involving self-selection , which is that the data is censored based on the endogenous choices of the players. For example, we may wish to estimate potential wages of workers and workers' beliefs about potential wages before they decide whether to pursue formal employment, but wages are observed only for those workers who choose to be employed. .

Our second main result is the following exact characterization of $\widehat{\Phi}(M)$, the set of counterfactual predictions that are consistent with M. We denote by $\bar{\mathcal{G}}$ the linked game with component games $\mathcal{G}$ and $\widehat{\mathcal{G}}$.

Theorem 2 (Counterfactual Predictions)**.**
An outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta\right)$ *is in* $\widehat{\Phi}(M)$ *if and only if there is a BCE* $\overline{\phi}$ *of the linked game* $\overline{\mathcal{G}}$ *such that (i) the marginal of* $\overline{\phi}$ *on* $A \times \Theta$ *is in* M *and (ii)* $\widehat{\phi}$ *is the marginal of* $\overline{\phi}$ *on* $\widehat{A} \times \Theta$*.*

Proof of Theorem 2. It is immediate from the definition of a counterfactual prediction that $\widehat{\phi} \in \widehat{\phi}(M)$ if and only if there is a $\phi \in M$ such that $\left(\phi, \widehat{\phi}\right)$ is a joint prediction for the linked game. The conclusion of the theorem is then an immediate consequence of Theorem 1. □

Remark 2. Our model includes a number of assumptions that, at first glance, seem to restrict the applicability of Theorems 1 and 2. In particular, we have assumed that information is described by a single information structure, preferences are known, the prior over the state is held fixed between observed and counterfactual games, and there is a single equilibrium that is played in the observed game and a single equilibrium in the counterfactual. In fact, all of these assumptions are merely normalizations. For example, variation in the information structure across players or across instances of the observed game can be captured as additional dimensions of players' signals, and uncertainty about players' preferences can be modeled by a expanding the state space. Section 8.5 in the Supplemental Appendix gives further details why all of these features of the model are actually without loss of generality.

Remark 3. The assumption that there is a single observed game and a single counterfactual game is only to simplify exposition. This result easily generalizes to a case where there is more than one observed game and more than one counterfactual game. This is discussed further in Section 7.

Remark 4. A leading case is when M is a polytope, i.e., the set of outcomes that satisfy a finite number of linear inequalities. For example, this is the case when $M = \{\phi\}$, so that states and actions are observed, or when $M = \{\psi\} \times \Delta(\Theta)$, so that actions are observed and states are not. When M is a polytope, $\widehat{\Phi}(M)$ is also a polytope, being the projection onto $\Delta(A \times \Theta)$ of the set of BCE of the linked game which satisfy the finitely many obedience constraints (1), one for each $(a_i, \widehat{a}_i)$, and the constraints corresponding to M. This is still a finite dimensional set, although the dimension grows exponentially in the number of players. If we fix a Bayesian welfare criterion $w(\widehat{a}, \theta)$ over ex post counterfactual outcomes, then the range of expected values of w across all counterfactuals can be obtained by solving a pair of finite-dimensional linear programs. We will use this fact in the examples in Section 5.

3.3 Counterfactuals when Information is Latent and Variable

Our focus in this paper is mainly on identifying counterfactual predictions when the game form changes but the distribution of fundamentals and players' information are held fixed. A distinct counterfactual exercise that has recently been considered in the literature is that the game form changes, the distribution of fundamentals is held fixed, but information can vary in an arbitrary way between the observed and counterfactual outcomes (Syrgkanis et al., 2018; Magnolfi and Roncoroni, 2020). Obviously, such counterfactuals with variable information are more permissive than the fixed information counterfactuals characterized by Theorem 2. In order to obtain a sharp comparison with the literature, we provide an analogous statement of Theorem 2 for counterfactuals with variable information.

Formally, an outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta\right)$ is a *variable information counterfactual prediction* if there exist μ, $\mathcal{I}$, $\widehat{\mathcal{I}}$, and Nash equilibria σ and $\widehat{\sigma}$ of $(\mu, \mathcal{G}, \mathcal{I})$ and $\left(\mu, \widehat{\mathcal{G}}, \widehat{\mathcal{I}}\right)$, respectively, such that the outcome ϕ induced by σ is in M and the outcome $\widehat{\phi}$ is induced by $\widehat{\sigma}$. In other words, $\widehat{\phi}$ can be rationalized as an equilibrium outcome for some $\left(\mu, \widehat{\mathcal{I}}\right)$, such that $(\mu, \mathcal{I})$ can also rationalize an equilibrium outcome in M. The set of variable information counterfactual predictions is denoted $\widehat{\Phi}_V(M)$, which will necessarily be a superset of $\widehat{\Phi}(M)$ (since counterfactual predictions further impose that $\mathcal{I} = \widehat{\mathcal{I}}$). We can now state an extension of our previous result for (fixed information) counterfactuals to variable information.

Proposition 1 (Counterfactual Predictions with Variable Information).
An outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta\right)$ is in $\widehat{\Phi}_V(M)$ if and only if there is an outcome $\overline{\phi}$ of the linked game $\overline{\mathcal{G}}$ such that (i) the marginal of $\overline{\phi}$ on $A \times \Theta$ is in M and (ii) $\widehat{\phi}$ is the marginal of $\overline{\phi}$ on $\widehat{A} \times \Theta$, and (iii) the marginals of $\overline{\phi}$ on $A \times \Theta$ and on $\widehat{A} \times \Theta$ are BCE of the component games $\mathcal{G}$ and $\widehat{\mathcal{G}}$, respectively.

4 Tightening Counterfactuals

Theorem 2 gives a precise characterization of what can be learned about counterfactual behavior when we hold fixed the players' information, and in the context of the benchmark model of Section 2. In the next section, we will describe contexts where the set of counterfactual predictions is small, and others where it is large. Four features of our benchmark model push in the direction of large sets of counterfactual predictions. First, the data may be limited. Second, there may be many information structures that rationalize the data. Third, even for a fixed information structure, there may be multiple Bayes Nash equilibria. Finally, as noted previously, our notion of an information structure also gives the players access to correlation devices that are orthogonal to the state, so that in fact the counterfactual prediction also includes those outcomes corresponding to *correlated* equilibria, and in general there can be many more correlated equilibria than Bayes Nash equilibria.

As an illustration of the potential for multiplicity, consider the case in which the counterfactual game is the same as the original game. We call this the *local counterfactual prediction.* One might hope that the local counterfactual would necessarily coincide with the observed outcome. In fact, this is not generally the case, and there will be more than one local counterfactual prediction whenever the observed outcome can be rationalized by an information structure for which the observed game has multiple equilibria. We note that there are special cases where the local counterfactual prediction is unique, which we discuss further in Section 5 and 6.

Nonetheless, this discussion points to a critical issue: the counterfactual prediction in the benchmark model can be quite permissive. In this section, we describe two ways to enrich the benchmark model that shrink the set of counterfactual predictions. All of these extensions can be understood as restrictions on the information structures that are used to rationalize the data. In Section 4.1, we describe how one can put lower bounds on players' information by assuming that they have some minimal information about the state. A special case is when there is some publicly observed covariate of θ that all players may condition on. We also discuss a methodology for imposing an upper bound on information. In Section 4.2 we consider the addition of exogenous payoff shifters, which can be viewed as a component of the state that is only observed by and only payoff relevant to a particular player.

4.1 Lower and Upper Bounds on Information

In our methodology for counterfactual predictions, we implicitly restrict attention to information structures that can rationalize the data, which takes the form of a restriction on the outcome in the observed game. This constraint can both viewed as both a lower bound and an upper bound on the information structures that can rationalize the data. It is a lower bound, in that whatever information structure rationalizes the data, players must at least know which action they played in the observed game, so any admissible information structure must be weakly more informative than the outcome of the observed game. At the same time, the data also places an upper bound on information; given that a player is taking some action in the observed game, there is a set of beliefs that the player might have that rationalize the action they took as a best response. If a player were too informed (both about the state and others' actions) then we would violate the equilibrium conditions. While these restrictions on the information structure are derived from the data, one could use a similar methodology to impose tighter bounds on the information structure, as we now explain.

Let us first illustrate how this can be done for lower bounds on information. We begin with a simple case, in which there are covariates of θ that are publicly observed by the players. Suppose that there is a quantity x which is jointly distributed with θ. Further suppose that x is publicly observed by all the players before they choose their actions. This variable may represent a public regressor that is observable to both the players and the econometrician. An outcome now describes the joint distribution of (a, θ, x), and with slight abuse of notation, we write $\phi(a, \theta, x)$, $\widehat{\phi}(\widehat{a}, \theta, x)$, and $\overline{\phi}(\overline{a}, \theta, x)$ for the outcomes in the observed, counterfactual, and linked games, respectively. The moment condition $M \subseteq \Delta(A \times \Theta \times X)$ may now incorporate restrictions on x, such as a known marginal on x or a known joint distribution of (θ, x). Importantly, we still assume that u depends on only a and θ, so that the content of x is purely informational.[6]

As x is common knowledge among the players, they can condition on it when evaluating the payoff from a deviation. This ability to condition on x is reflected in the definition of a BCE and, in particular, the obedience constraints. Recall that the obedience condition (1) requires that for each player and conditional on each recommended action, a player cannot profit by deviating to a different action. When x is publicly observed, this condition must be strengthened so that conditional on each recommended action and conditional on x, there is no profitable deviation. Applied to the linked game, the obedience constraint is now that for all i, $\overline{a}_i$, $\overline{a}'_i$, and x,

$$\sum_{\theta \in \Theta} \sum_{\overline{a}_{-i} \in A_{-i}} \left(u_i\left(\overline{a}_i, \overline{a}_{-i}, \theta\right) - u_i\left(\overline{a}'_i, \overline{a}_{-i}, \theta\right)\right) \overline{\phi}\left(\overline{a}_i, \overline{a}_{-i}, \theta, x\right) \geq 0. \tag{2}$$

The analogue of Theorem 2, under the additional hypothesis that the players publicly observe x, is as follows: An outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta \times X\right)$ is a counterfactual prediction if and only if there exists a BCE $\overline{\phi} \in \Delta\left(\overline{A} \times \Theta \times X\right)$, satisfying the generalized obedience conditions (2), such that $\widehat{\phi}$ is the marginal of $\overline{\phi}$ on $\widehat{A} \times \Theta \times X$, and the marginal of $\overline{\phi}$ on $A \times \Theta \times X$ is in M.

Thus, we see that public regressors can easily be incorporated into our framework. The public signal x represents a lower bound on information, in the sense that the players cannot have less information about the state than that contained in x, although they may have more. An extreme case of this is when $x = \theta$, so that the players all observe the state perfectly. In this case, the players' signals contain no additional information about the state

[6] A conceptually distinct way to model the public regressor is by assuming that there is a separate instance of the observed and counterfactual games for each value of x, denoted by $\mathcal{G}_x$ and $\widehat{\mathcal{G}}_x$, where players' utilities can now depend directly on x, and by forming a larger linked game consisting of component games $\left(\mathcal{G}_x, \widehat{\mathcal{G}}_x\right)_{x \in X}$. We could then view the data for each x as a separate restriction on the marginal on actions for the observed game $\mathcal{G}_x$. With this formulation, we would implicitly be assuming that the players' signals are conditionally independent of x. In contrast, the approach developed in the main text allows for correlation between signals and public regressors.

beyond that contained in the public regressor, and signals serve only as a correlation device. As a result, the counterfactual prediction reduces to those outcomes such that $\widehat{\phi}(\cdot,\theta)/\mu(\theta)$ is a correlated equilibrium of the complete information game that is $\widehat{\mathcal{G}}$ when the state is θ.

We may also consider other lower bounds on information. For example, suppose that the games of interest are auctions in which θ represents the bidders' value for the good being sold, so that Θ is a product set $\times_{i=1}^{N}\Theta_i$, with $\Theta_i \subset \mathbb{R}$. The widely used *private value* hypothesis says that each bidder knows their own value for the good. In this case, each player's utility only depends on their own component, θ_i. Moreover, players can condition on θ_i when deviating, and hence the obedience constraints for the linked game are now written that, for all i, $\overline{a}_i$, $\overline{a}_i'$, and θ_i,

$$\sum_{\theta_{-i}\in\Theta_{-i}}\sum_{\overline{a}_{-i}\in A_{-i}}\left(u_i\left(\overline{a}_i,\overline{a}_{-i},\theta_i\right)-u_i\left(\overline{a}_i',\overline{a}_{-i},\theta_i\right)\right)\overline{\phi}\left(\overline{a}_i,\overline{a}_{-i},\theta_i,\theta_{-i},x\right)\geq 0. \tag{3}$$

Thus, we see that private values represents both a restriction on the structure of payoffs as well as a restriction on information. And again, there is an analogue of Theorem 2 for private values, which simply uses the modified obedience constraint (3) instead of (1). We give two example using this version of Theorem 2: In Section 6, we study an entry game where firms have private entry costs; and in the Supplemental Appendix, we consider first-price auctions with private values. Private costs and private values play important roles in the variable-information analysis of Syrgkanis et al. (2018) and the variable-information and fixed-information analysis of Magnolfi and Roncoroni (2020).

Both of these examples can be viewed as special cases of the following more general framework for lower bounds on information. Fix an information structure $\widetilde{\mathcal{I}}=\left(\left\{\widetilde{S}_i\right\}_{i=1}^{N},\widetilde{\pi}\right)$. We say that the players' information structure $\mathcal{I}$ is more informative than $\widetilde{\mathcal{I}}$ in the sense of *individual sufficiency*, as defined by Bergemann and Morris (2016), which says that the signal $\widetilde{s}_i$ can be written as a garbled observation of s_i. An equivalent definition is that $\mathcal{I}$ represents the same distribution over higher-order beliefs as an information structure where the sets of signals are of the form $S_i = \widetilde{S}_i \times \widehat{S}_i$, and such that the marginal of π on $\widetilde{S} \times \Theta$ is $\widetilde{\pi}$. The interpretation is that each player i observes at least their "base" signal $\widetilde{s}_i$ in $\widetilde{\mathcal{I}}$, but may observe additional information, in the form of the "auxiliary" signal $\widehat{s}_i \in \widehat{S}_i$. Outcomes now describe the joint distribution of actions and states and the minimal signals $\widetilde{s} \in \widetilde{S}$, and we write $\phi(a, \theta, \widetilde{s})$, etc. The moment restriction M can now incorporate assumptions about how the base signals $\widetilde{s}$ are distributed and are correlated with the state. Specifically, $M \subseteq \Delta\left(A \times \Theta \times \widetilde{S}\right)$. Since players observe their base signals, they can condition on them when they deviate. The obedience constraint for the linked game is therefore that for every i, $\overline{a}_i$, $\overline{a}'_i$, and $\widetilde{s}_i$,

$$\sum_{\theta\in\Theta} \sum_{\widetilde{s}_{-i}\in\widetilde{S}_{-i}} \sum_{\overline{a}_{-i}\in A_{-i}} \left(u_i\left(\overline{a}_i, \overline{a}_{-i}, \theta\right) - u_i\left(\overline{a}'_i, \overline{a}_{-i}, \theta\right)\right) \overline{\phi}\left(\overline{a}_i, \overline{a}_{-i}, \theta, \widetilde{s}_i, \widetilde{s}_{-i}\right) \geq 0. \tag{4}$$

With this formalism in place, we can now give the most general version of Theorem 2 with lower bounds on information:

Corollary 1 (Joint Prediction with Informational Constraints).
An outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta \times \widetilde{S}\right)$ is in $\widehat{\phi}(M)$ if and only if there exists a BCE $\overline{\phi} \in \Delta\left(\overline{A} \times \Theta \times \widetilde{S}\right)$, satisfying the generalized obedience conditions (4), such that (i) the marginal of $\overline{\phi}$ on $A \times \Theta \times \widetilde{S}$ is in M and (ii) $\widehat{\phi}$ is the marginal of $\overline{\phi}$ on $\widehat{A} \times \Theta \times \widetilde{S}$.

This framework nests the public-regressor and private-value examples. In particular, the public regressor can be modeled as $\widetilde{S}_i = X$ for all i. The private values example corresponds to the case where $\widetilde{S}_i = \Theta_i$ and the moment restriction M incorporates the constraint that $\phi(a, \theta, \widetilde{s}) = 0$ whenever $\theta \neq \widetilde{s}$. However, it allows us to incorporate richer examples as well, such as where players are known to observe a noisy signal of the state $\widetilde{s}_i$, and there is a known correlation structure between the base signals and the state, which is expressed as the linear restriction

$$\phi(a, \theta, \widetilde{s}) = \widetilde{\pi}(\widetilde{s}|\theta) \sum_{\widetilde{s}'\in\widetilde{S}} \sum_{a\in A} \phi(a, \theta, \widetilde{s}').$$

By varying the informativeness of this signal, one can generate counterfactual predictions covering the entire range of lower bounds on information, from no information (when $\widetilde{s}$ is independent of θ) to full information (when $\widetilde{s}_i$ perfectly identifies θ).

The preceding discussion concerned lower bounds on information. We now discuss how one could impose upper bounds on information. As indicated above, the restriction on the observed outcome is itself an upper bound on information: Players cannot have so much information in the counterfactual game that they would have preferred to deviate from their observed actions. In many cases, this upper bound on information is abstract and difficult to interpret. But we can contemplate games for which there is a natural interpretation of the implied upper bounds on information. Consider a game in which the players do not interact at all; they simply try to guess the state, and are incentivized to be as accurate as possible. If we were to observe players not always guessing the state correctly, then we would know that their information about the state must be bounded away from full information.

While we pursue this in our examples, one could take this idea even further and model constraints on players' information through *conjectures* about how the players would behave in a hypothetical strategic environment. In particular, let $\widetilde{\mathcal{G}}$ be a game form and $\widetilde{\mathcal{M}}$ be a moment restriction on the outcome ϕ in $\widetilde{\mathcal{G}}$ which jointly represent an upper bound on information. For example, $\widetilde{\mathcal{G}}$ could be a game in which players guess the state and guess one anothers' beliefs about the state, and $\widetilde{\mathcal{M}}$ could denote the set of outcomes where players' payoffs are bounded away from that of full information. We can consider counterfactual predictions for a game $\widehat{\mathcal{G}}$ that are induced by an equilibrium under an information structure that also rationalizes an observed outcome in a game $\mathcal{G}$ and is consistent with the conjectured behavior in the hypothetical game $\widetilde{\mathcal{G}}$. By Theorem 1, such counterfactuals can be characterized as marginals of BCE of the linked game with component games $\mathcal{G}$, $\widehat{\mathcal{G}}$, and $\widetilde{\mathcal{G}}$.[7]

[7] An alternative approach is to fix an information structure $\widetilde{\mathcal{I}}$ and restrict attention to information structures that are weakly less informative than $\widetilde{\mathcal{I}}$, in the sense of individual sufficiency (Bergemann and Morris, 2016). This approach is conceptually appealing, but unfortunately it breaks the linear structure of the problem. To see why, recall that $\widetilde{\mathcal{I}}$ is individually sufficient for the revelation information structure that induces an outcome of the linked game only if we can write $\overline{a}_i$ as a noisy signal of $\widetilde{s}_i$ that is conditionally independent

4.2 Known Exogenous Payoff Shifters

The next approach to tightening the counterfactual involves enriching the observed game so that the data reveals more about players' beliefs. A common device in applied microeconomics is to assume that there are exogenous payoff "shifters" that affect a player's utility from a particular action, and that the econometrician can observe the mappings from payoff shifters to actions (Tamer, 2003; Jia, 2008; Bajari et al., 2010; Somaini, 2020). This shifter is essentially an instrument for a player's payoffs, with the further restriction that the instrument for player i's payoffs is observed by player i, but other players do not get any direct signals. This mapping provides much richer information about players' latent preferences (or beliefs) than just the action itself. They are especially powerful for purposes of identification when their distribution is observed and they are assumed to be orthogonal to other uncertainty in the economy. Such payoff shifters play an important role in the work of Magnolfi and Roncoroni (2020).

We now show how such payoff shifters can be incorporated into our framework. We discuss this in the context of the observed game only. For each i, let ω_i be a payoff shifter for player i, which takes values in a finite set Ω_i. Player $i's$ payoff is now of the form $u_i(a, \theta, \omega_i)$. We assume that ω_i is directly observable to player i, but not to players $-i$. Moreover, we let $\eta(\omega|\theta)$ denote the conditional distribution of the entire profile of payoff shifters, which is assumed to be known. (This may be because this object can be directly observed, or because the shifters themselves can be observed and there are further orthogonality assumptions that allow η to be identified.) In addition, we assume that players do not learn anything about ω from their signals in the information structure. In this manner, the payoff shifters are distinct from the state θ, about which the players may obtain more information.

of $\widetilde{s}_{-i}$ and θ, i.e., there exist kernels $K_i : \tilde{S}_i \to \Delta(A_i)$ such that

$$\sum_{\overline{a}_{-i}\in\overline{A}_{-i}} \overline{\phi}\left(\overline{a}_i, \overline{a}_{-i}, \widetilde{s}_i, \widetilde{s}_{-i}, \theta\right) = K_i\left(\overline{a}_i|\widetilde{s}_i\right) \sum_{\overline{a}'\in\overline{A}} \overline{\phi}\left(\overline{a}', \tilde{s}_i, \tilde{s}_{-i}, \theta\right). \tag{5}$$

The condition (5) is non-linear in K_i and $\overline{\phi}$, which makes it much less computationally tractable.

How can this assumption be incorporated into our framework? Since each player can observe their component of ω, they may condition on it when choosing their action. We can therefore reduce this game to a normal form in which player i's action space is $\Xi_i = A_i^{\Omega_i}$, i.e., the set of pure mappings $\xi_i : \Omega_i \to A_i$. It is without loss to restrict player i to pure mappings, since we will allow them to mix over pure strategies, as in the baseline model. Conditional on a profile of such mappings, player i's payoff is

$$u\left(\xi, \theta\right) = \sum_{\omega \in \Omega} \eta\left(\omega | \theta\right) u_i\left(\xi\left(\omega\right), \theta, \omega\right).$$

Now, fix an information structure $\mathcal{I}$. Having rewritten the game with payoff shifters in normal form, we can now define a strategy for player i to be a mapping $\sigma_i : S_i \to \Delta\left(\Xi_i\right)$. Theorem 2 then goes through immediately.

The value of the payoff shifters is that the implied richness of the observed game may reveal more information about players' information and the state. We illustrate this with a single-player example in the next section.

5 One-Player Games and the Roy Model

We will present two applications to illustrate the content of our theorem. In this section, we consider the case with one player, thus a decision probelm, and relate it to Roy's model of self-selection. In the next section, we consider a two-player two-action entry game.

5.1 The Roy Model

We now specialize to the case where $N = 1$, in which case the "game" is really a decision problem being solved by a single decision maker. To highlight the connection to our general analysis, we will maintain the terminology of player and game in this special case of a one-player game. We further specialize to the case where there are two states $\Theta = \{-1, 1\}$ and two actions $A = \{0, 1\}$.[8] In the observed game, the player's payoff is the product:

$$u(a, \theta) = a\theta.$$

Thus, the player receives a payoff of zero if the action is $a = 0$, and otherwise the payoff equals the state. Note that these payoffs are without loss of generality up to normalization. In the counterfactual game, the payoff changes to

$$\widehat{u}(a, \theta) = a(\theta + z)$$

for some $z \in \mathbb{R}$, so that the payoff from the action $a = 1$ is translated by a constant z independent of the state. The observed game corresponds to the special case with $z = 0$.

This problem can be viewed as a discretized version of the canonical model of self-selection, first formulated by Roy (1951). In particular, the action can be interpreted as a decision by the player whether to opt in to some activity. Opting out (action 0) results in a certain payoff (normalized to 0), but opting in (action 1) yields an uncertain payoff. The state is normalized to be equal to this payoff.

This model has been described by Heckman (2010) as "the prototype for many models of self-selection in economics" (p. 264). For example, opting in could represent a decision to choose one occupation over another or to participate in a randomized control trial. In the employment application, the payoff represents the player's potential long-run earnings, and "opting in" represents the choice of one occupation over another. In the experimental application, the opt in payoff is the average potential treatment effect, which may be uncertain due to latent characteristics of the player that affect the treatment's efficacy. The player may have knowledge of the state that informs their decision of whether to opt in, so that the average payoff is different among the players who opt in versus those who opt out.

[8] We will note in the text how some results generalize to the many action, many state case, and report these results as propositions in the Supplementary Appendix.

a/θ	-1	1
0	α	$\frac{1}{2}-\alpha$
1	$\frac{1}{2}-\alpha$	α

Table 1: Observed distribution of actions and states.

As we vary z, we vary the average payoff from opting in, which corresponds to a shift in the distribution of potential wages or the distribution of potential treatment effects.

We will first consider the case where the analyst observes the entire distribution of actions and states. This corresponds to the case originally considered by Roy (1951), in which it is assumed that wages are observed for whatever occupation is chosen. This will make it easy to visualize the insight that fixing information gives substantially tighter counterfactuals than variable information. We then consider the case where outcomes are only observed if the player chooses action 1, i.e., we only observe potential wages for workers who are employed, and we only observe average treatment effects for agents who opt into the randomized control trial. This illustrates the logic of our approach with restricted data and is obviously the empirically relevant case for the Roy model. Finally, we will illustrate fixed-information counterfactual comparisons and the role of payoff shifters in shrinking counterfactuals.

5.2 Counterfactual Welfare with Observable Outcomes

We first construct the set of counterfactual predictions for $\widehat{u}$ corresponding to a particular outcome in the observed game u. We suppose the each state θ is equally likely and that the observed outcome (a, θ) is given by the probabilities in Table 1 for some $\alpha \in [1/4, 1/2]$.[9] Thus, both states are equally likely and the probability that the player chooses the optimal action for the true state is the same in both states.

Counterfactual welfare for this case is plotted in the left-hand panel of Figure 1. We will consider what happens for three values of α: 0.5, 0.375, and 0.25. The corresponding outcomes are rationalized by distinct information structures. If $\alpha = 0.5$, then the player

[9] The constraint $\alpha \leq 1/2$ is necessary for the probability of every outcome to be non-negative. If $\alpha < 1/4$, the player's payoff would be negative, which is inconsistent with utility maximization, since the player can always opt out and obtain a payoff of zero.

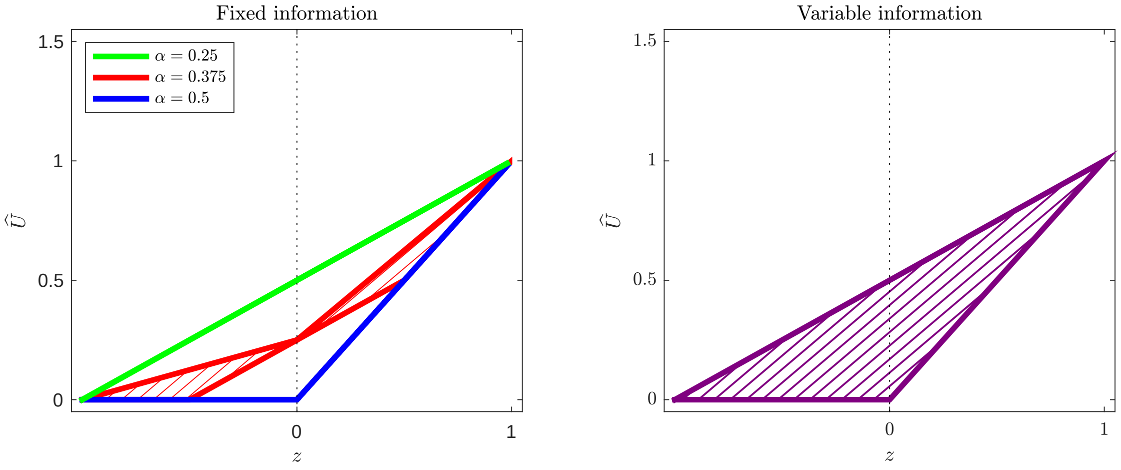

Figure 1: Counterfactual welfare in the Roy model.

opts in if and only if the state is 1. This is only possible if the player has full information about the state. If we change z, it will still be the case that the player has full information, so that counterfactual payoff must be on the green line. If $\alpha = 0.25$, then the player enters with the same probability independent of the state. This is possible only if the player has no information: Otherwise there would be some correlation between the action and the state. Again, in the counterfactual, the player must still have no information, so that the counterfactual payoff must be on the blue line. Thus, for both $\alpha = 0.5$ and $\alpha = 0.25$, we get a point prediction for the counterfactual payoff under fixed information.

If $\alpha = 0.375$, then the player's action is ex post optimal less often than can be explained by full information and more often than can be explained by no information. In fact, there is more than one information structure that could give rise to the observed outcome, Here are two examples:

(i) Half of the time, the player's signal is fully informative, so that the player opts in if and only if the state is high. The other half of the time, the signal is uninformative, and the player randomizes between opting in and opting out with equal probabilities.

(ii) The player's signal is a noisy observation of the state, that matches the state with probability $3/4$ and mismatches with probability $1/4$. The player opts in if the signal is 1 and opts out when the signal is -1.

Suppose that the player's information is given by (i), which we hold fixed as we vary z. Then the counterfactual payoff is simply the average of the payoffs from full information and no information. This is the top red line in the left panel of Figure 1, which is halfway between the blue and green lines. In contrast, if information is given by (ii), then the player's strategy when $z = 0$ remains optimal as long as $z \in [-1/2, 1/2]$. When $z < 1/2$, the noisy signal is not strong enough to induce the player to enter, and it is optimal to always opt out. Similarly, when $z > 1/2$, it is optimal to always opt in. The resulting welfare is given by the lower red line in the left panel of Figure 1. Thus, while the counterfactual does not permit a point prediction anymore, the set of counterfactual prediction remains small and degrades only slowly as we move away from the observed game.

In fact, the information structures (i) and (ii) achieve maximum and minimum counterfactual welfare. Thus, all we can say about the player's counterfactual welfare is that it lies in the hatched area between the red lines. To see why the information structure (ii) attains minimum counterfactual welfare, observe that whatever information the player has, it must be Blackwell more informative than the signal generated by the action the player chose, which corresponds precisely to the information structure (ii). Thus, welfare under this information structure must be weakly lower than welfare under any other information structure that rationalizes the player's behavior.

As for why (i) attains the maximum counterfactual payoff, we can explain this in terms of distributions of the player's belief about the state that are consistent with the observed behavior, which we parametrize by the probability that the state is high. As we discussed above, this is an equivalent way to describe information for single-player games. Note that the obedience constraint implies that when the player opts out, their belief is less than $1/2$, and when they opt in, their belief is greater than $1/2$. In addition, the *average* belief is $1/4$

conditional on opting out and 3/4 conditional on opting in. Among distributions of beliefs satisfying this property, there is a maximal distribution with respect to the mean-preserving spread order, which is the distribution that assigns probability 1/4 to a belief of 0, 1/4 to a belief of 1, and 1/2 to a belief of 1/2. This is precisely the belief distribution induced by the information structure (i). Since the Blackwell order on information structures is equivalent to the mean-preserving spread order on the associated belief distributions, we conclude that the player achieves a higher payoff under (i) than under any other information structure that rationalizes the data.

A final property that we would like to highlight from this example is that if $z = 0$, then for any value of α, there is a unique local counterfactual payoff prediction. The argument is straightforward. First, the player can always simply repeat their observed action in the counterfactual and achieve as high a payoff as in the observed outcome, so the counterfactual payoff must be weakly greater than observed payoff. By a similar argument, in the observed game, the player could have instead mimicked their action in the counterfactual, and hence the observed payoff must be greater than counterfactual welfare, and thus they are equal.

To summarize, we have identified three key properties of fixed information welfare counterfactuals in the Roy problem:

1. Minimum welfare in the counterfactual is equal to player welfare if they had only the information revealed by behavior in the observed game.

2. Maximum welfare is equal to player welfare if they had the Blackwell most informative information structure consistent with behavior in the observed game.

3. The local counterfactual is unique.

In the Supplemental Appendix, we prove that all these properties hold in all single-player counterfactual predictions, with the proviso that (2) generalizes only in the two state case (with more than two states, there may not be a Blackwell most informative information structure consistent with observed game behavior).

All three results rely on the fact that there is a single player and we are looking at counterfactuals about that player's payoff. In the Supplemental Appendix, we give an example of a single-player game where there is a not a unique counterfactual when we look at behavior instead of payoffs. We will see the failure of the uniqueness of the local counterfactual in the entry game in the next section. It is important to note, however, that there are nontrivial games for which the local counterfactual is unique. In particular, the Supplemental Appendix shows that the local counterfactual for welfare is unique in two-player zero-sum games.

5.3 Comparison with Variable Information Counterfactuals

To illustrate the power of fixing information, we now compare the preceding fixed-information counterfactuals with those obtained under variable information. When information is variable, the only restriction from the data is that both states are equally likely. Blackwell's theorem (Blackwell, 1953) implies that for every decision problem, the player's welfare is minimized when the player has no information about the state, and their welfare is maximized when they have full information about the state. These outcomes are depicted in the right panel of Figure 1. Note that the upper and lower bounds correspond to the blue and green curves in the left panel. Under no information, it is optimal for the player to opt out when $z \leq 0$, and it is optimal to opt in when $z \geq 0$. Under full information, the player learns the state, and it is optimal to opt in when the state is 1, and it is optimal to opt out when the state is -1. Any other information structure must result in welfare between these two extremes. If the information structure is variable, there is no more we can say.

5.4 Partially Observed Outcomes

We have analyzed the benchmark case in which we observe the entire joint distribution of actions and states. A natural assumption in the randomized control trial applications is that the data is censored, and we only observe average treatment effects for those who opt into

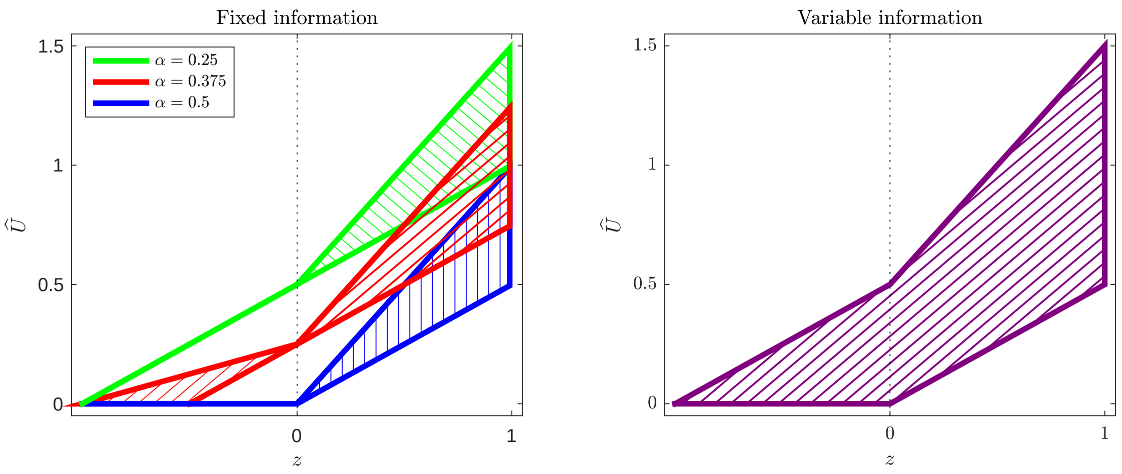

Figure 2: Variable information counterfactual welfare in the Roy model.

the trial. Even in the employment application, if "opting in" represents a decision to join the labor force, then it is natural to suppose that we do not observe potential income for the unemployed. Equivalently, we observe the distribution of actions, but we only observe the state when the player opts in. In the left panel of Figure 2, we have plotted the analogous welfare bounds for the same simulated data as discussed previously, except that we do not observe the state distribution when $a = 0$. For $z < 0$, the welfare bounds are exactly the same as when we observe the entire action/state distribution. The reason is that for this range of counterfactuals, opting in has become less attractive, so whenever the player opted out in the observed data, they will continue to do so in the counterfactual, in which case the payoff is independent of the state.

When $z > 0$, however, the player may opt in when we observed them opt out. Exactly how often this happens and what welfare results depends on the state distribution that rationalizes the data. Consider first when $\alpha = 0.5$, i.e., we observe the player opt in half the time, but when the player opts in, the state is always high. We do not know the distribution of the state when the player opts out. In the case considered previously, the state was always low when the player opted out. This is the most pessimistic case, and corresponds to the lower green curve in Figure 2. However, the obedience constraint only requires that the state be low at least half the time when the player opts out; in particular, the data can also be rationalized by the state being equally likely to be high or low when the player opted out, so that the player is indifferent between actions but breaks the tie in favor of opting out. In this case, the counterfactual payoff would be strictly higher, since when $z > 0$, the player would strictly prefer to opt in, thereby achieving the welfare on the higher green line. Any payoff between the green lines can be rationalized by some information structure and some distribution of the state when the player opts out.

A similar analysis explains the upper and lower bounds for $\alpha = 0.375$ and $\alpha = 0.25$. When $\alpha = 0.25$, the upper blue line corresponds to the case considered previously, when the state was equally likely to be high or low when the player opted out. If instead the state was always low when the player opted out, then counterfactual welfare would lie on the lower blue line. For $\alpha = 0.375$, we had previously assumed that the state was high with probability 1/4 when the player opted out. If instead the state was high with probability 1/2, we would obtain the higher red dashed line in the left panel of Figure 2. Finally, the lower red dashed line is obtained when the state is always low when the player opted out. Initially this line coincides with its counterpart in the left-hand panel, since in that case the player would also have continued opting out when we observed them opting out, as long as z was less than 1/2.

There are two broad takeaways from the partially observed case. First, the additional uncertainty about the state distribution does expand our counterfactual prediction for welfare when opting in becomes more attractive, although the direction in which our bounds expand depends on the particular observed outcome: When the player's observed welfare was high, it is the upper bound on welfare that is relaxed, whereas when observed welfare was low, it is the lower bound that is relaxed. In the intermediate case, both upper and lower bounds on counterfactual welfare are relaxed. Second, the additional uncertainty does not affect our counterfactual prediction when opting in becomes less attractive. The reason is that for such counterfactuals, the player would continue to opt out whenever they opted out in the observed decision problem. Since the payoff from opting out is independent of the state, the fact that we do not know the state distribution conditional on opting out does not affect our prediction for counterfactual welfare.

For completeness, we have also depicted the variable information prediction when the state is unobservable when the agent opts out, in the right panel of Figure 2. Once again, we see that with variable information, we cannot rule out any counterfactual welfare between the complete and no information counterfactual predictions.

5.5 Point Identification with Payoff Shifters

We conclude the discussion of the single-player case with an example showing how payoff shifters, as described in Section 4.2, can significantly shrink the set of counterfactual predictions. Suppose we first make the counterfactual problem harder, by having the analyst not observe anything about the state in the original game. But suppose instead that the player observes a payoff shifter ω, which is distributed uniformly on $[-1, 1]$ and independent of the player's signal and the state.[10] The payoff from opting out is still normalized to zero, but the payoff to a player who opts in is now $\theta + 1 - 2\omega$.

Now the analyst can observe the joint distribution of (a, ω). Let p denote the (possibly random) interim probability that the player assigns to the event $\theta = 1$. Clearly, it is optimal for the player to choose $a = 0$ only if

$$p - (1 - p) + 1 - 2\omega \leq 0 \iff p \leq \omega.$$

Thus, the probability that the player chooses the action 0 when the shifter is ω is precisely the probability that p is less than ω. Indeed, if we write $P(\omega)$ for the probability of choosing action 0 given ω, then $P(\omega)$ is the cumulative distribution of the player's interim belief. As we discussed above, the distribution of the interim belief is a canonical representation of

[10] We let the payoff shifter be a continuous variable. It would be straightforward but messy to make the same points with a finite set of realizations (consistent with the rest of our model), in which case we would be able to identify the cumulative distribution of the player's belief at finitely many points.

information for single-player games, so that the function $P(\omega)$ describes all features of the player's information that are relevant for counterfactual predictions.

6 Two Player Binary Action Games and Entry

Our second application is a simple entry game with private entry costs, in the spirit of that used in the empirical work of Ciliberto and Tamer (2009) and Magnolfi and Roncoroni (2020). The entry game is a parameterized two-player binary-action game with symmetric payoffs and strategic substitutes, where each player has a binary signal affecting only their own payoffs. We will focus on the entry game interpretation for concreteness and to highlight the relevance for the empirical literature.

6.1 The Entry Game

Two firms choose whether to enter (E) or not enter (N) a market. The payoff from not entering is zero. Each firm $i = 1, 2$ has a cost to enter the market c_i, which is either 0 or $C > 0$. If a single firm enters, that firm earns a monopoly revenue X. If both firms enter, they each earn revenue $X - \Delta$, where $\Delta > 0$ reflects the revenue decrease due to a loss of the monopoly position. These payoffs are summarized in the following matrix:

$$\begin{array}{ccc}
a_1/a_2 & N & E \\
\text{N} & (0,0) & (0, X - c_2) \\
\text{E} & (X - c_1, 0) & (X - \Delta - c_1, X - \Delta - c_2)
\end{array}$$

To map this game into our framework, we identify the state with the ordered pair of firms' entry costs, that is, $\theta = (c_1, c_2)$. In principle, firms could be uncertain about their own costs. In the version of this problem studied in the applied literature, the firms are generally assumed to know their own entry costs, but not have any information about other's costs, beyond the prior distribution thereof. We will maintain the assumption that players know their own costs. This case can be formally included in our model by following the extension for minimal information in Section 4.1, where $\widetilde{S}_i = \{0, C\}$ and $\widetilde{\pi}\left((c_1, c_2) \,|\, (c_1, c_2)\right) = 1$, i.e., each firm receives a signal equal to their entry cost with probability one. Thus, the latent parameter is the information each firm has about the other firm's cost.

6.2 Fixed Information Counterfactual with Known Private Values

In the following numerical example, we take $\Delta = C = 2$. We further assume that the observed data comes from a version of this game in which $X = 3$ and all entry-cost profiles are equally likely. The counterfactual of interest is how aggregate profit, denoted $\widehat{U}$, varies with X, which uniformly scales the payoff from entering. We can equivalently interpret an increase in X as a decrease in firms' entry costs.

Now, if each firm knows nothing about the other firm's cost, there is a unique symmetric Bayes Nash equilibrium which involves firms entering when their cost is low and not entering when their cost is high. (Note that when the cost is low, entering is a strictly dominant strategy, and if a firm enters with positive probability when their cost is high, then the other firm has a strict incentive to not enter when their cost is high.) We assume that the analyst observes the outcome induced by this equilibrium, i.e., ϕ places probability one on each firm entering if and only if their entry cost is low.

In Figure 3, we have plotted the counterfactual predictions for aggregate profit for $X \in [0, 5]$, and holding fixed all of the other parameters of the model. These predictions were computed numerically by solving the linear program for maximum and minimum counterfactual payoffs, for a fine grid of values of X. The blue graph with horizontal hatching is the counterfactual prediction with fixed information. The observed outcome corresponds to the point where $X = 3$ and $\widehat{U} = 2$. Let us describe how this graph is generated by considering the relevant cases.

We first consider which information structures are consistent with the observed outcome. Recall that we assumed that each firm knows their own cost. In fact, this assumption is redundant with the restrictions placed by the observed outcome on the information structure:

Since each firm's action is perfectly correlated with their cost, in any information structure that rationalizes the data, firms must learn their own costs. Next, when $X = 3$, it is a strictly dominant strategy to enter when the cost is low. Thus, the data does not place any additional restrictions on what firms know when their cost is low; entering is optimal regardless of the low-cost firms' higher-order beliefs. However, the data does restrict the beliefs of high-cost firms. In particular, the high-cost firms' obedience constraint is just binding, meaning that if they had no additional information beyond their cost, they are indifferent between entering and not entering. If a high-cost firm's signal contained non-trivial information about the other firm's cost, then they would sometimes believe that the other firm is more likely to have a high cost than under the prior. Under such a belief, it would be strictly optimal to enter, thus violating obedience. As a result, high-cost firms' signals cannot contain any non-trivial information about the other firm's entry cost.

Thus, we see that the observed outcome implicitly places both lower and upper bounds on the players' information. On the one hand, the firms must know their own entry costs. On the other hand, when a firm has a high cost, they cannot have learned anything non-trivial about the other firm's cost. This feature of the example aligns with our earlier discussion of bounds on information in Section 4.1..

With this characterization of which information structures are consistent with the observed data, it is possible to explicitly describe the information structures and equilibria that attain the maximum and minimum counterfactual predictions, for each value of X. As indicated in Figure 3, there are several cases, depending on the value of X. We will here explain how the counterfactual prediction is generated for $X > 2$. Explicit calculations of extreme counterfactual predictions for all values of X are given in Section 8.2 in the Supplemental Appendix.

First, if $X > 4$, then it is a strictly dominant strategy to enter regardless of the entry cost, so that there is a point prediction for welfare. There is also a unique counterfactual

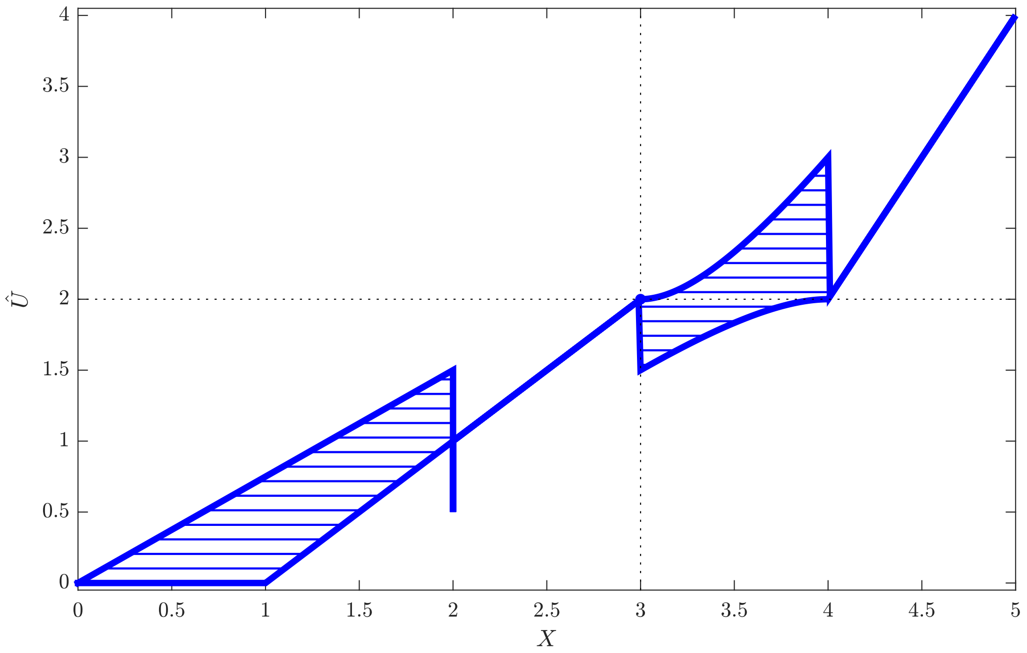

Figure 3: Counterfactual welfare in the entry game.

prediction for aggregate payoffs for $X \in (2,3)$, which is attained in the equilibrium in which firms enter if and only if their costs are low.

We now explain how the counterfacutal prediction is generated for $X \in [3,4]$. For this range of parameters, there is a mixed-strategy symmetric equilibrium in which high-cost firms are indifferent between entering and not entering. Since the high-cost firm receives a payoff of zero, all surplus is generated by the low-cost signals. Moreover, the low-cost firm's payoff from entering is just the high-cost firm's payoff from entering plus 2 (the difference in entry cost), so that aggregate payoff is 2 in this equilibrium. However, we can see in Figure 3 that there are counterfactual predictions for which the aggregate payoff is higher or lower than 2. This is possible because even though high-cost firms receive no information about the other firm's cost, they may have access to correlation devices, which can be used to achieve a higher or lower probability that both high-cost firms enter simultaneously.

In particular, suppose that in addition to learning their entry costs, firms observe the outcome of a randomization device which recommends either enter ($s_i' = e$) or not enter ($s_i' = n$) with symmetric probabilities:

s_1'/s_2'	n	e
n	$1 - \beta - 2\gamma$	γ
e	γ	β

where β and γ are non-negative scalars and $\beta + 2\gamma \leq 1$. Because the distribution of (t_1', t_2') is independent of (c_1, c_2), it does not affect the firms' higher-order beliefs. We will construct equilibria in which low-cost firms ignore the correlation device and always enter (since it is strictly dominant to do so) and high-cost firms enter if and only if $s_i' = e$.

In order to maximize firm profits, it is intuitive that the optimal value of β is zero, so as to minimize the loss from two firms entering, and that the binding obedience constraint will be the one for not entering (since we are trying to discourage firms from entering too often). This indeed turns out to be the case. The obedience constraint reduces to $\gamma \leq (X - 3) / (X - 2)$. Setting γ equal to this upper bound, the obedience constraint for not entering holds as an equality. Then γ ranges from 0 to 1/2 as X goes from 3 to 4. The aggregate payoff reduces to $2X - 4 - 2(X - 3) / (X - 2)$, which exactly coincides with the highest counterfactual aggregate payoff. When $X = 4$, this BCE involves exactly one high-cost firm always entering.

To minimize firm profits, it is intuitive that it is optimal to maximize the probability of both firms entering by setting $\beta = 1 - 2\gamma$ and that the obedience constraint for entering will bind. This reduces to $\gamma \leq (4 - X) / (5 - X)$. We again take γ equal to its upper bound, which ranges from 1/2 to 0 as X ranges from 3 to 4, whereas β ranges from 0 to 1. The aggregate payoff then reduces to $X - 1 - 1/(5 - X)$, which coincides with the lowest counterfactual aggregate payoff. When $X = 3$, this BCE involves exactly one high-cost firm entering. This lowers the aggregate payoff below the equilibrium in which high-cost firms do not enter, because it lowers the payoffs of low-cost firms.

This analysis illustrates the multiplicity of local counterfactual predictions when there is more than one player. We see in Figure 3 that not only are there multiple counterfactual predictions when the counterfactual game is different from the observed game, but this is also true when the counterfactual game is the same as the observed game. In this case, it is because there could be correlation devices that players ignore in the observed outcome but could use in a counterfactual equilibrium. We also saw that there were many information structures consistent with the observed data. In this case, when high-cost firms had no information about the other firm's cost, there was an equilibrium where high-cost firms did not enter. But there if firms observed a signal correlating their entry decisions when cost were high, there is also an equilibrium where exactly one firm always enters when costs are high.

6.3 Variable Information Counterfactuals and Lower Bounds on Information

We next contrast the fixed-information counterfactual prediction with that obtained with variable information. We consider two versions of variable information: First, we can allow any information structure, including those in which the firms do not even know their own entry costs. Second, we can consider all information structures in which firms know their own entry costs, which we refer to as *private cost* information structures. In Figure 4, the variable private-cost information counterfactual prediction is red with vertical hatching, and the unrestricted variable information counterfactual is green with diagonal hatching. For reference, we have also plotted in blue the previously described counterfactual with fixed information.

As expected, the unrestricted variable information graph contains the variable private-cost information graph. The latter in turn contains the fixed information graph. This is because the observed outcome can only be rationalized by private-cost information structures, as discussed above. In fact, the variable private-cost and fixed information graphs

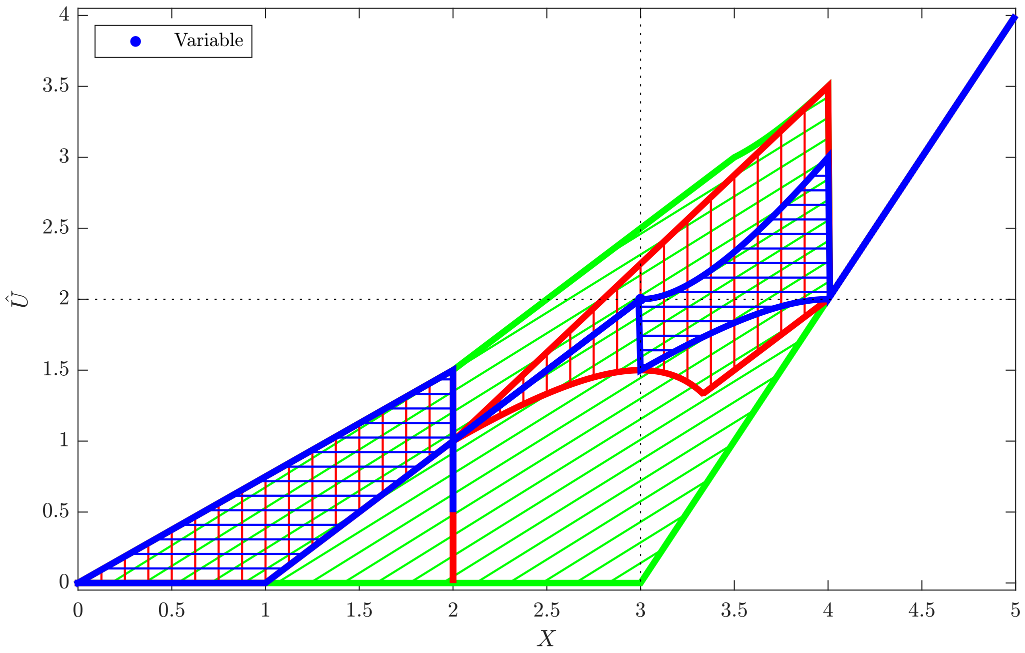

Figure 4: Counterfactual welfare in the entry game with variable information.

coincide for $X < 2$. For this range of parameters, the high-cost firms have a strictly dominant strategy to not enter, and the observed outcome places no restrictions on the low-cost firms' information, except that they know their costs are low. However, for $X \geq 2$, there are outcomes which can be obtained with some private-cost information structure but are inconsistent with fixed information. This must be because these information structures involve the high-cost firms receiving non-trivial information about the cost of the other firm. For example, when $X = 2$, there is a private-cost information structure and equilibrium in which the aggregate payoff is zero. This occurs when both firms enter when the cost is low and exactly one high-cost firm enters if any firm has high cost. This is only possible if high-cost firms can distinguish when they are facing a low-cost firm (in which case they must enter with probability one) or when they are facing another high-cost firm (in which case one of the high-cost firms must enter with probability strictly less than one). The difference between fixed and variable information is perhaps most dramatic for $X \in (2, 3)$, where there is a unique counterfactual prediction under fixed information but a fat set for variable information.

7 Discussion and Conclusion

The purpose of this paper has been to describe exactly the implications of Bayesian rationality and common priors for counterfactual predictions, under the hypothesis that information is fixed. We have shown that there is a sharp description of the set of counterfactual outcomes that are consistent with observed data. We have demonstrated through examples that the predictive power of fixed information can be significant, especially compared to what can be predicted if we do not fix information between observation and counterfactual. There are other cases where even fixing information still leads to a permissive counterfactual. We have discussed a variety of methodologies for tightening the counterfactual.

7.1 Identifying the Information Structure

We have treated the information structure as a nuisance parameter. We want to control for the information structure, but do not want to learn anything about it. One might be interested in identifying the information structure for its own sake. Or one might think that it would be a helpful step, instrumentally, as a step in performing counterfactuals. Either way, we want to reiterate our point in the introduction that information structures are complicated objects and thus hard to identify.

In single-player games, the problem is hard (although as we will see much easier than the many player case). In this case, all that matters for behavior is the player's interim beliefs about θ. In that case, we could without loss of generality restrict attention to information structures in which $S = \Delta(\Theta)$ and the signals are normalized to be equal to the interim belief. Even so, the player's interim belief may have full support on $\Delta(\Theta)$, which is an infinite dimensional vector space. No finite approximation can capture all of the relevant behavior, in that for any two distinct distributions of beliefs, we can always find a counterfactual game for which counterfactual predictions would differ across the two distributions. However, if Θ is separable, then $\Delta(\Theta)$ is a separable metric space in the weak-* topology, and behavior in single-player games with continuous utilities can be well approximated in this topology. We could in principle compute the set of rationalizing information structures whose support lives in a finite grid in $\Delta(\Theta)$ and have a fair approximation of behavior in continuous counterfactual games.

In the multi-player case, the canonical space for information structures is the *universal type space* (Brandenburger and Dekel, 1993; Mertens and Zamir, 1985). This is also an infinite-dimensional vector space, and much of our comments about the single-player case apply here as well. A qualitative difference is that types in the universal type space are much harder to approximate. For example, the set of "finite" types, which correspond to belief hierarchies that arise in finite information structures, is dense in the universal type space in the product topology. However, finite types do not provide a good approximation of behavior, in the sense that rationalizable behavior along a sequence of finite types may not be close to rationalizable behavior for the limit type (Rubinstein, 1989; Dekel, Fudenberg, and Morris, 2006). We see no simple way to even approximately compute the set of information structures that rationalize multi-player outcomes. Moreover, the universal type space only encodes players' higher order beliefs about the state, and it abstracts away from correlation devices which may be relevant for strategic interaction (cf. Liu, 2015). This makes our "implicit identification" methodology all the more appealing.

7.2 Further Applications of the Joint Predictions Theorem

Our main characterization of counterfactual results were proved via an appeal to a theory of joint predictions in multiple games described in Theorem 1. We believe that the perspective on joint predictions embodied in Theorem 1 will be useful for other problems. We now discuss some potential further applications.

First, as we discussed in Remark 3, Theorem 1 can be used to extend Theorem 2 to the case where there is more than one observed game and more than one counterfactual game. Specifically, suppose that there is data on observed games $\left(\mathcal{G}^1, \ldots, \mathcal{G}^K\right)$ which we use to generate predictions for the counterfactual games $\left(\widehat{\mathcal{G}}^1, \ldots, \widehat{\mathcal{G}}^K\right)$. Theorem 1 immediately gives us a characterization of joint outcomes of these games as being marginals of BCE of a linked game $\overline{\mathcal{G}}$, whose component games are precisely the $\mathcal{G}^k$ and $\widehat{\mathcal{G}}^k$. Now let $\left(M^1, \ldots, M^K\right)$ denote the data restrictions on the outcomes of the observed games, then we immediately have a characterization of counterfactual outcomes $\left(\widehat{\phi}^1, \ldots, \widehat{\phi}^K\right)$ in the counterfactual games that are consistent with the data, in the sense that there is an information structure $\mathcal{I}$ that can rationalize the observed data and the counterfactual outcomes as being induced by Bayes Nash equilibria under $\mathcal{I}$. In particular, $\left(\widehat{\phi}^k\right)$ are consistent in this sense if and only if there exists a BCE $\overline{\phi}$ of the linked game $\overline{\mathcal{G}}$ such that $\widehat{\phi}^k$ is the marginal of $\overline{\phi}$ on $\widehat{\mathcal{G}}^k$ for each $k = 1, \ldots, \hat{K}$, and such that the marginal of $\overline{\phi}$ on $\mathcal{G}^k$ is in M^k for each $k = 1, \ldots, K$.

One interpretation of the observed games is that they represent different instances of a public regressor that affects the players' payoffs and is common knowledge among the players. What distinguishes this formulation from the public regressors model of Section 4 is that when we model the different realizations of the public regressor as distinct games, we are implicitly assuming that players' information is uncorrelated with the public regressor. In particular, the joint distribution of signals and states is the same for every realization.

Second, note that Theorem 1 characterizes more than just which outcomes could arise from the same information. It also tells us the possible correlation structures between actions *across* games. This can be used to generate informationally-robust comparative statics[11], as we now explain.

Suppose that there are two games of interest $\mathcal{G}^k$ for $k = 1, 2$. An analyst is interested in comparing behavior in these two games. In particular, there is a value $v^k\left(\phi^k\right) \in \mathbb{R}$ for the outcome ϕ^k of $\mathcal{G}^k$, and the analyst uses these values to rank outcomes both within and across

[11] As we discussed in the introduction, this question has been pursued for linear best response games in Heumann (2019).

games. For example, $\mathcal{G}^1$ and $\mathcal{G}^2$ may be entry games with different barriers to entry, and the analyst may be interested in comparing social surplus across these games. Or they may represent different auction formats, and the analyst wishes to compare expected revenue. The analyst's question is, which game is associated with a higher value?

Without knowing the information structure, one can only characterize a set of possible values for v^k, which we denote by V^k. Clearly, if $\max V^2 \leq \min V^1$, then $\mathcal{G}^1$ dominates $\mathcal{G}^2$ in that no matter what are the information structures for the two games, the outcome of $\mathcal{G}^1$ has higher value than the outcome of $\mathcal{G}^2$. In this case, we may say that $\mathcal{G}^1$ *dominates* $\mathcal{G}^2$ *with variable information*, in that we insist on the outcome of $\mathcal{G}^1$ being superior to the outcome of $\mathcal{G}^2$ even if there are different information structures for the two games. The analyst may also be interested in a weaker notion of dominance that holds pointwise across information structures, meaning that for every information structure $\mathcal{I}$, and equilibria of $(\mathcal{I}, \mathcal{G}^1)$ and $(\mathcal{I}, \mathcal{G}^2)$, the induced outcome for $\mathcal{G}^1$ is superior to the induced outcome of $\mathcal{G}^2$. This is obviously equivalent to the following definition: $\mathcal{G}^1$ *dominates* $\mathcal{G}^2$ *with fixed information* if for every joint prediction (ϕ^1, ϕ^2) of $\mathcal{G}^1$ and $\mathcal{G}^2$, we have $v^1(\phi^1) \geq v^2(\phi^2)$.[12] It is straightforward to test dominance with fixed information using Theorem 1: $\mathcal{G}^1$ dominates $\mathcal{G}^2$ with fixed information if and only if for every BCE $\overline{\phi}$ of the linked game with components $\mathcal{G}^1$ and $\mathcal{G}^2$, we have $v^1(\phi^1) \geq v^2(\phi^2)$, where ϕ^k is the marginal of $\overline{\phi}$ on the kth game. When the games are finite, determining dominance with fixed information reduces to checking the feasibility of a finite system of linear inequalities.

One can even contemplate richer applications of Theorem 1, for example, where we combine the informationally-robust comparative static and counterfactual predictions, whereby we ask when one counterfactual game dominates another with fixed information, when we further impose data based restrictions on an observed game, and maintain fixed information throughout. In the Supplementary Appendix, we report examples of such comparative stat-

[12] We thank Jeff Ely for suggesting this notion of dominance.

ics on counterfactuals for both the Roy model and the entry game. Further developing these and other applications of Theorem 1 is an important direction for future work.

7.3 The Fixed Information Assumption

So aside from the usual assumptions of game theory (expected utility, common priors, etc), our most substantive assumption is that information is held fixed in the counterfactual. It is easy to think of reasons why information should *not* be held fixed. Why is it possible to alter the players' actions and payoffs without in any way altering the players' information? If observed and counterfactual games do not occur simultaneously, then why should not the information structure be evolving over time? Economic agents can often influence the kind of information they receive, and why should information gathering behavior remain the same when other aspects of the world have changed?

Even so, and in spite of these legitimate complaints, fixed information is still an important benchmark, and it is implicitly adopted in much of the extant literature on counterfactuals in industrial organization (e.g., Guerre, Perrigne, and Vuong, 2000; Ciliberto and Tamer, 2009). The main virtue of our methodology is that it adopts weak assumptions about the form of information and equilibrium selection. The predictions of our model are therefore quite safe, although the range of counterfactual outcomes may be larger than what would be obtained with a more structural model. The suitability of our approach to any particular application therefore depends both on the analyst's uncertainty about the form of players' information and preferences with regard to the misspecification thereof.

References

Anscombe, F. and R. Aumann (1963): "A Definition of Subjective Probability," *Annals of Mathematical Statistics*, 34, 199–205.

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

Bergemann, D., B. Brooks, and S. Morris (2017): "First Price Auctions with General Information Structures: Implications for Bidding and Revenue," *Econometrica*, 85, 107–143.

Bergemann, D. and S. Morris (2013): "Robust Predictions in Games with Incomplete Information," *Econometrica*, 81, 1251–1308.

——— (2016): "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 11, 487–522.

——— (2017): "Belief-Free Rationalizability and Informational Robustness," *Games and Economic Behavior*, 104, 744–759.

Blackwell, D. (1951): "Comparison of Experiments," in *Proc. Second Berkeley Symp. Math. Statist. Probab.*, Berkeley: University of California Press, 93–102.

——— (1953): "Equivalent Comparison of Experiments," *Annals of Mathematics and Statistics*, 24, 265–272.

Brandenburger, A. and E. Dekel (1993): "Hierarchies of Belief and Common Knowledge," *Journal of Economic Theory*, 59, 189–198.

Canen, N. and K. Song (2020): "A Decomposition Approach to Counterfactual Analysis in Game-Theoretic Models," Tech. rep.

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

Dekel, E., D. Fudenberg, and S. Morris (2006): "Topologies on Types," *Theoretical Economics*, 1, 275–309.

Gualdani, C. and S. Sinha (2020): "Identification and Inference in Discrete Choice Models with Imperfect Information," Tech. rep.

Guerre, E., I. Perrigne, and Q. Vuong (2000): "Optimal Nonparametric Estimation of First-price Auctions," *Econometrica*, 68, 525–574.

Heckman, J. J. (2010): "Selection Bias and Self-Selection," in *Microeconometrics*, ed. by S. N. Durlauf and L. E. Blume, Springer, 242–266.

Heumann, T. (2019): "Informationally Robust Comparative Statics in Incomplete Information Games," Tech. rep.

Jia, P. (2008): "What Happens When Wal-Mart Comes to Town: An Empirical Analysis of the Discount Retailing Industry," *Econometrica*, 76, 1263–1316.

Lewis, D. (1973): *Counterfactuals*, Cambridge: Harvard University Press.

Liu, Q. (2015): "Correlation and Common Priors in Games with Incomplete Information," *Journal of Economic Theory*, 157, 49–75.

Magnolfi, L. and C. Roncoroni (2020): "Estimation of Discrete Games with Weak Assumptions on Information," Tech. rep.

Mertens, J.-F. and S. Zamir (1985): "Formalization of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14, 1–29.

Peski, M. (2008): "Comparison of Information Structures in Zero-Sum Games," *Games and Economic Behavior*, 62, 732–735.

Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

Rubinstein, A. (1989): "The Electronic Mail Game: Strategic Behavior under 'Almost Common Knowledge'," *American Economic Review*, 79, 385–391.

Savage, L. (1954): *The Foundations of Statistics*, New York: Wiley, 1st ed.

Somaini, P. (2020): "Identification in Auction Models with Interdependent Costs," *Journal of Political Economy*, 128, 3820–3871.

Syrgkanis, V., E. Tamer, and J. Ziani (2018): "Inference on Auctions Under Weak Assumptions on Information," Tech. rep., Harvard University.

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

8 Supplemental Appendix

8.1 Single Player Example: Further Analysis

8.1.1 General Analysis

In Section 5, we explained how in single-player games, minimum counterfactual welfare is obtained with the minimally informative information structure, in which a player's signal is their observed action. We now give a general statement of this result:

Proposition 2 (Minimum single-player counterfactual welfare)**.** *Suppose $N = 1$, and fix an observed decision problem $\mathcal{G} = (A, u)$ and moment restriction $M = \{\phi\}$. Define an information structure $\mathcal{I} = (S, \pi)$ by $T = A$ and such that $\pi(a|\theta)\,\mu(\theta) = \phi(a, \theta)$ for all a and θ. Then $\mathcal{I}$ rationalizes the observed outcome. Moreover, for every counterfactual decision problem $\widehat{\mathcal{G}} = \left(\widehat{A}, \widehat{u}\right)$, the minimum expected counterfactual welfare across all counterfactual predictions is attained when the information structure is $\mathcal{I}$, and minimum counterfactual welfare is*

$$\sum_{a \in A} \max_{\widehat{A} \in \widehat{A}} \sum_{\theta \in \Theta} \phi(a, \theta)\, \widehat{u}\left(\widehat{A}, \theta\right).$$

The proof is elementary, and follows the argument given in the text.

We next give a general statement of the result that with binary states, there is a minimally informative information structure which attains minimum counterfactual welfare. When $\Theta = \{\theta_1, \theta_2\}$, we can represent the player's belief conditional on their signal as the probability that the state is θ_1. For each observed action $a \in A$, there is an interval of beliefs for which that action is optimal, which we can denote by $\left[x^L(a), x^H(a)\right]$. Conditional on taking the action a, every realized belief must be in this interval. The Blackwell-most informative belief distribution consistent with the data must have all of the mass concentrated on the end points of this interval. Any information structure that generates this distribution of beliefs will maximize the player's welfare in all counterfactual decision problems. One such information structure is $\mathcal{I} = (S, \pi)$ where $S = A \times \{L, H\}$ and for all a such that $x^L(a) < x^H(a)$, π

satisfies

$$\pi\left(a, H|\theta_2\right) = \frac{1}{\mu\left(\theta_2\right)} \frac{\phi\left(a, \theta_1\right) - \frac{x^L(a)}{1-x^L(a)} \phi\left(a, \theta_2\right)}{\frac{x^H(a)}{1-x^H(a)} - \frac{x^L(a)}{1-x^L(a)}};$$

$$\pi\left(a, H|\theta_1\right) = \frac{1}{\mu\left(\theta_1\right)} \frac{x^H\left(a\right)}{1-x^H\left(a\right)} \frac{\phi\left(a, \theta_1\right) - \frac{x^L(a)}{1-x^L(a)} \phi\left(a, \theta_2\right)}{\frac{x^H(a)}{1-x^H(a)} - \frac{x^L(a)}{1-x^L(a)}},$$

where $\pi\left(a, H|\theta_1\right) = \phi\left(a, \theta_1\right)/\mu\left(\theta_1\right)$ and $\pi\left(a, H|\theta_2\right) = 0$ if $x^H\left(a\right) = 1$. Otherwise, if $x^L\left(a\right) = x^H\left(a\right)$, then we can take $\pi\left(a, H|\theta_1\right) = \pi\left(a, H|\theta_2\right) = 1$. With this information structure, the player has an optimal strategy to choose a after the signals (a, H) and (a, L). Moreover, $\mathcal{I}$ is Blackwell-more informative than any other information structure that rationalizes the data. We have proven the following proposition:

Proposition 3. *Suppose that $N = 1$ and $\Theta = \{\theta_1, \theta_2\}$, and fix an observed game $\mathcal{G} = (A, u)$ and moment restriction $M = \{\phi\}$. Then the information structure $\mathcal{I}$ described in the preceding paragraph rationalizes the observed outcome. Moreover, for every counterfactual decision problem $\widehat{\mathcal{G}} = \left(\widehat{A}, \widehat{u}\right)$, the maximum expected counterfactual welfare across all counterfactual predictions is attained when the information structure is $\mathcal{I}$, and maximum counterfactual welfare is*

$$\sum_{(a,k)\in A\times\{L,H\}} \max_{\widehat{a}\in\widehat{A}} \sum_{\theta\in\Theta} \widehat{u}\left(\widehat{a}, \theta\right) \pi\left(a, k|\theta\right) \mu\left(\theta\right).$$

At a high level, this result depends on the fact that the set of distributions over beliefs partially ordered by mean-preserving spreads is a lattice when $|\Theta| = 2$. When $|\Theta| > 2$, this partially ordered set is no longer a lattice, and in particular, there need not be a most informative distribution of beliefs that rationalizes the data.

Finally, we argue that there is always a unique local counterfactual in single-player games:

Proposition 4. *Suppose that $N = 1$ and $M = \{\phi\}$. If the counterfactual game $\widehat{\mathcal{G}}$ is equal to $\mathcal{G}$, then there is a unique counterfactual welfare in all counterfactual predictions, which is* *welfare under ϕ:*

$$\sum_{a\in A}\sum_{\theta\in\Theta} u\left(a,\theta\right)\phi\left(a,\theta\right).$$

The argument is that given in the text: Fix an information structure $\mathcal{I}$ and observed and counterfactual equilibrium strategies σ and $\hat{\sigma}$ (that is, σ and $\hat{\sigma}$ are optimal decision rules). Since the two games are the same, the payoffs in the games are respectively

$$U=\sum_{\theta\in\Theta}\sum_{s\in S}\sum_{a\in A} u\left(a,\theta\right)\sigma\left(a|s\right)\pi\left(s|\theta\right)\mu\left(\theta\right) \text{ and } \widehat{U}=\sum_{\theta\in\Theta}\sum_{s\in S}\sum_{a\in A} u\left(a,\theta\right)\hat{\sigma}\left(a|s\right)\pi\left(s|\theta\right)\mu\left(\theta\right).$$

But $\hat{\sigma}$ is a feasible strategy in the observed game, so the fact that σ is an equilibrium must be $U\geq\widehat{U}$. By an analogous argument, $\widehat{U}\geq U$, so in fact they are equal. Finally, by the definition of a counterfactual prediction, we must have that σ and $\mathcal{I}$ induce ϕ, so that U is equal to welfare under ϕ.

8.1.2 Comparing two counterfactuals

We will next use the Roy model to illustrate the methodology for fixed-information dominance discussed in Section 7. Specifically, we ask for which pairs of counterfactual parameters z^1 and z^2 does the agent always attain higher welfare under z^1 than under z^2, when we restrict attention to those information structures which are consistent with the observed outcome. Figure 5 plots the set of pairs $\left(\widehat{U}^1,\widehat{U}^2\right)$ of agent welfare obtained for the pairs $(z^1,z^2)=(0.5,0.75)$ when the observed outcome corresponds to $\alpha=0.375$, both under the assumption of fully-observable outcomes (the blue set) and partially-observable outcomes (the red set).

The picture clearly shows that agent is unambiguously better off under z^2. This can be seen from the fact that all of the sets lie above the 45 degree line. Indeed, this conclusion is theoretically trivial: the counterfactual with z^2 has payoffs that are pointwise higher, so that the agent could achieve a higher payoff with z^2 than with z^1 simply by using whatever strategy was optimal for z^1. Note that while this conclusion is theoretically obvious, it is

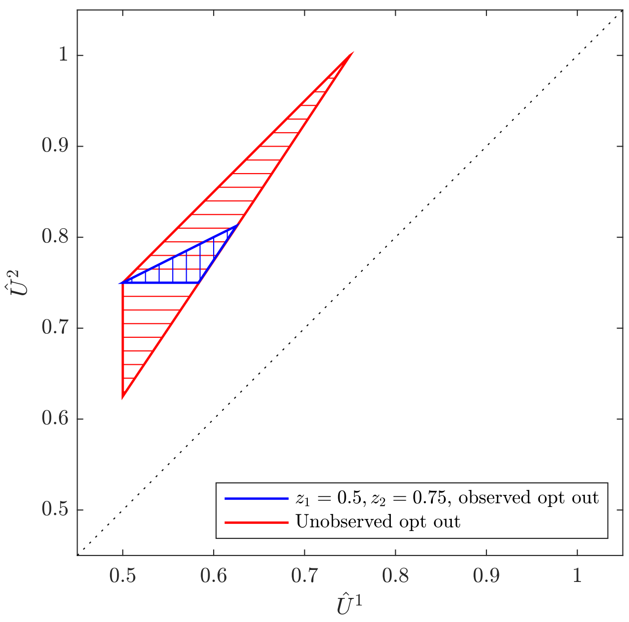

Figure 5: Ranking counterfactuals in the Roy model.

not apparent in Figure 1: For many pairs $z^2 > z^1$, the set of possible welfare outcomes for the agent overlap. It is only by plotting agent welfare resulting from joint counterfactual predictions that we can see that higher values of z dominate.

Nonetheless, this example illustrates the power of fixing information when computing informationally-robust comparative statics: Without holding information fixed, there would be no dominance ranking between z^1 and z^2, whenever the two are sufficiently close together.

8.1.3 Welfare versus behavior

In Section 5, we primarily focused on the player's welfare. This is not the only counterfactual outcome of interest. More broadly, we may ask how the player's *behavior* could change in the counterfactual, i.e., the probability of opting in for each state. While we do not analyze this question in detail, we can say that there are generally much weaker restrictions on behavior than on welfare. This is illustrated in Figure 6, which depicts the total probability that the player opts in as we vary z, for the cases considered above.

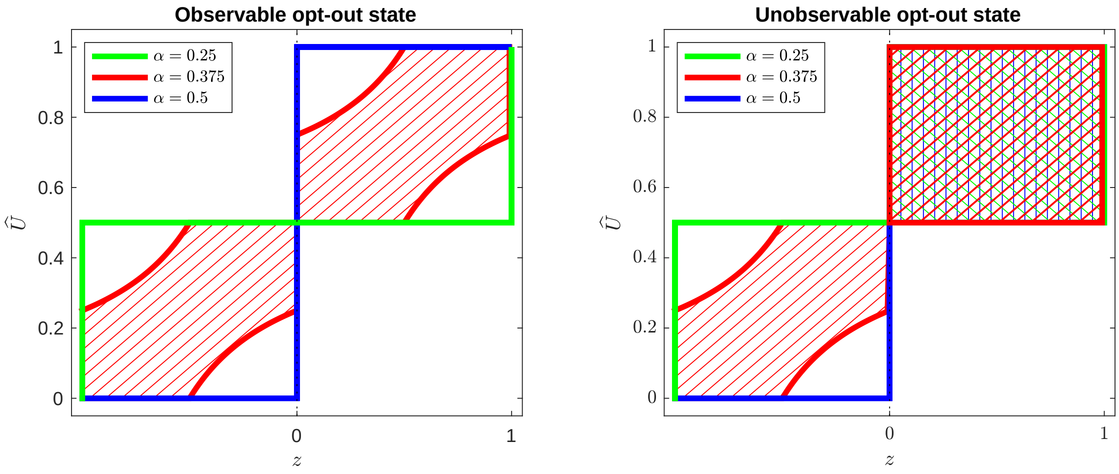

Figure 6: Counterfactual behavior in the Roy model.

The left panel describes the counterfactual probability of opting in when we observe the entire outcome, including the state distribution when the player opts out. When the observed outcome is consistent with either either no information (the green curve) or full information (the blue curve), there is generically a point prediction for counterfactual behavior. However, for no information and $z = 0$, there are counterfactual predictions consistent with any opt-in probability between zero and one. This is true even though there is a point prediction for counterfactual welfare, simply because when $z = 0$, the player is indifferent between actions. For the intermediate case of partial information, there is always a fat set of counterfactual opt-in probabilities. Again, this is true even when $z = 0$, when there is a point prediction for welfare.

The counterfactual prediction for behavior when we do not observe the state after opting out is depicted in the right panel of Figure 6. The prediction is even more permissive in this case: For every $z > 0$, any opt-in probability between 1/2 and 1 is consistent with all three cases considered. For, in each of these examples, the player must always opt in when the state is good, and there is a state distribution that rationalizes the player's observed decision to opt out when $z = 0$ but such that they would strictly prefer to enter if $z > 0$.

8.2 Entry Game: Further Analysis

8.2.1 Detailed calculations for entry counterfactuals

In this appendix, we analytically construct the equilibria that attain the boundaries of the numerically computed counterfactual prediction in Figure 3. We do not give a proof that these bounds are optimal.

Both firms always entering is an equilibrium if $X \geq \Delta + C$, and the resulting payoff is $2(X - \Delta) - C = 2X - 6$. This is the unique counterfactual prediction when $X > \Delta + C = 4$.

For $X < \Delta + C$, always entering is not an equilibrium. As long as $X - \Delta/2 - C \geq 0$, there is a mixed strategy equilibrium in which low-cost firms always enter and a firm with high cost enters with probability α, to make the other firm indifferent between entering and not entering:

$$X - (1 + \alpha)\Delta/2 - C = 0 \iff \alpha = 2\frac{X - \Delta/2 - C}{\Delta}.$$

Thus, these strategies are an equilibrium for $X \in [\Delta/2 + C, \Delta + C]$. Since this equilibrium makes high-cost firms indifferent between entering and not entering, the payoff of the high-cost firm is zero, and the payoff when the cost is low is just $C = 2$, so that the overall payoff in this equilibrium is 2.

We now construct equilibria for $X \in [\Delta/2 + C, \Delta + C]$ that attain the upper and lower bounds of the counterfactual welfare. Firms observe the outcome of a correlation device that produces signals (s_1, s_2) that are independent of the firms' costs and has the following probabilities:

s_1/s_2	0	1
0	$1 - \beta - 2\gamma$	γ
1	γ	β

where $\gamma \in [0, 1/2]$ and $\beta \in [0, (1 - \gamma)/2]$. In the equilibria we now construct, low-cost firms ignore this signal and always enter, but a high-cost firm i enters if and only if $s_i = 1$.

The obedience constraints are as follows: Conditional on $s_i = 1$, the likelihood of the other firm entering is $(\gamma + 2\beta)/(2\gamma + 2\beta)$. The reason is that the other firm will enter regardless of their signal if their cost is low, but will only enter if they get the high signal when their cost is high. Conditional on this signal, the payoff from entering must be non-negative:

$$X - \Delta \frac{\gamma + 2\beta}{2(\gamma + \beta)} - C \geq 0.$$

Similarly, conditional on being told to not enter and having a high cost, the payoff from entering must be non-positive:

$$X - \Delta \frac{1 - \beta - 2\gamma + 2\gamma}{2(1 - \beta - \gamma)} - C \leq 0$$

The equilibrium payoffs are

$$\frac{1}{2}\left(X - \Delta(1 + \gamma + \beta)/2\right) + \frac{1}{2}\left[(\gamma + \beta)(X - C) - \Delta(\gamma + \beta + \beta)/2\right].$$

To obtain minimum counterfactual welfare, we set $\beta = 1 - 2\gamma$ and make the obedience constraint for entering hold as an equality. Intuitively, we are pushing down welfare by having firms enter with high probability. Plugging in $\Delta = C = 2$, and solving for β, we obtain

$$\beta = 1 - 2\gamma = \frac{X - 3}{5 - X}.$$

It is straightforward to verify that the obedience constraint for entering is always satisfied with these values for β and γ and $X \in [3, 4]$. The resulting aggregate payoff is

$$X - 1 - \frac{1}{5 - X}.$$

which coincides with the simulated minimum counterfactual welfare.

For maximum counterfactual welfare, we set $\beta = 0$ and make the obedience constraint for not entering hold as an equality. Intuitively, we increase welfare by having firms enter less often, so as to avoid the low-payoff from duopoly. Solving for γ, we obtain

$$\gamma = 1 - \frac{1}{X-2} = \frac{X-3}{X-2}.$$

So γ goes from 0 to 1/2 as X goes from 3 to 4. Note that when $\beta = 0$, the obedience constraint for entering is unambiguously satisfied, since the left-hand side reduces to 1/2, and the right hand side is always at least 1/2. The resulting payoff is

$$2X - 4 - 2\frac{X-3}{X-2},$$

which coincides with the simulation.

We next consider the equilibrium to enter if and only if $c_i = 0$. The payoff from entering with a low cost is clearly positive. The payoff from entering with the high cost is

$$X - \Delta/2 - C,$$

and the payoff from entering with a low cost is $X - \Delta/2$, so this is an equilibrium if $X \in [\Delta/2, \Delta/2 + C] = [1, 3]$. The resulting ex ante sum of payoffs is

$$\frac{X}{2} + \frac{1}{2}\left(X - \Delta\right),$$

which reduces to $X-1$ in the numerical example. This is the unique counterfactual prediction when $X \in (2, 3)$, and it is the lower boundary of the counterfactual prediction when $X \in [1, 2]$.

If $X \in [0, \Delta/2]$, there is an equilibrium in which firms mix over whether they enter, which results in a payoff of zero. This attains the lower boundary of the counterfactual prediction for $X \in [0, 1]$.

Next we construct the profit maximizing BCE when $X \in [0, \Delta]$. Using a correlation device as we did above for $X \in [3, 4]$, we can coordinate the low firms' behavior so that firms enter only if they have low cost, a firm enters with probability one if they are the only low-cost firm, and when both firms have low-cost, and exactly one firm enters when both firms have low cost. This is obviously an equilibrium: Entering is strictly dominated for high signals, and if a firm with low cost does not enter in equilibrium, then the other low-cost firm must be entering, so the payoff from deviating would be $X - \Delta \leq 0$. The resulting aggregate payoff would be $3X/4$ (that is, $3/4$ of the time exactly one firm enters, and it is a firm with low cost). This coincides with the upper boundary of the simulation.

Finally, we construct an equilibrium that attains the low payoff at $X = \Delta = C$. First, there is a correlation device as above when $\gamma = 1/2$. In addition, we assume that low-cost firms can observe the cost of the other firm. Consider the following strategies: A high-cost firm enters if and only if $s_i = 1$. A low cost firm enters with probability 1 if the other firm's cost is low or if the other firm's cost is high and $s_i = 1$. Otherwise, when the other firm's cost is high and $s_i = 0$, the low-cost firm does not enter. The high-cost firm gets zero surplus from entering. Relative to the equilibrium where firms enter if and only if the cost is low, profit has dropped by $C/4 = 1/2$, since $1/4$ of the time it is a high cost firm entering as a monopolist rather than a low-cost firm. This equilibrium is knife edge: First, it depends on $\Delta = C$, so that the low-cost firm is indifferent to entering as a duopolist, and the high-cost firm is indifferent to entering as a monopolist. Second, if X is a little bigger than $\Delta = C$, low-cost firms would strictly prefer to enter when the high-cost firm enters, and if X is a little smaller than $\Delta = C$, the high-cost firm would be unwilling to enter.

8.2.2 Informationally-Robust Comparative Statics in the Entry Game

In this appendix, we conduct version of the joint counterfactual prediction analysis described in Section 7. In this case, we ask whether higher X are necessarily associated with higher payoffs for the firms. We conducted six versions of this counterfactual, which are depicted in Figure 7.

We computed joint predictions for counterfactual aggregate profit for the firms for two different counterfactual games: $X = 2.4$ and $X = 2.6$. Three versions of this computation under different informational assumptions are depicted in Figure 7.

First, we computed a counterfactual prediction when we restrict attention to information structures that can rationalize the data used in Section 6, namely that when $X = 3$, firms enter if and only if their cost is low. Thus, in this example, we are actually computing joint predictions for three games, where $X \in \{2.4, 2.6, 3\}$, and we impose a data restriction on the $X = 3$ game and plot the set of pairs of counterfactual aggregate profit for the games with $X = 2.4$ and $X = 2.6$. In fact, for this case, the set of joint counterfactual predictions can immediately be read from Figure 3: For $X \in (2, 3)$, there is a unique counterfactual prediction for aggregate payoffs under fixed information, and this prediction is increasing in X. Indeed, we see in Figure 7 that the joint counterfactual prediction when we have the data restriction is the single blue point. This point is above the 45 degree, meaning that the firms are unambiguously better off in the aggregate when $X = 2.6$ than they are when $X = 2.4$.

Second, for these same parameters, we computed a joint prediction when we impose that the same information structure is used for both values of X, but we allow all private-cost information structures (depicted in red). In this case, the joint prediction spans both sides of the 45 degree line, so that $X = 2.6$ does not dominated $X = 2.4$ with fixed information, when we do not have a restriction from the data. A fortiori, $X = 2.6$ does not dominated $X = 2.4$ under variable information, even when we restrict to private-cost information structures.

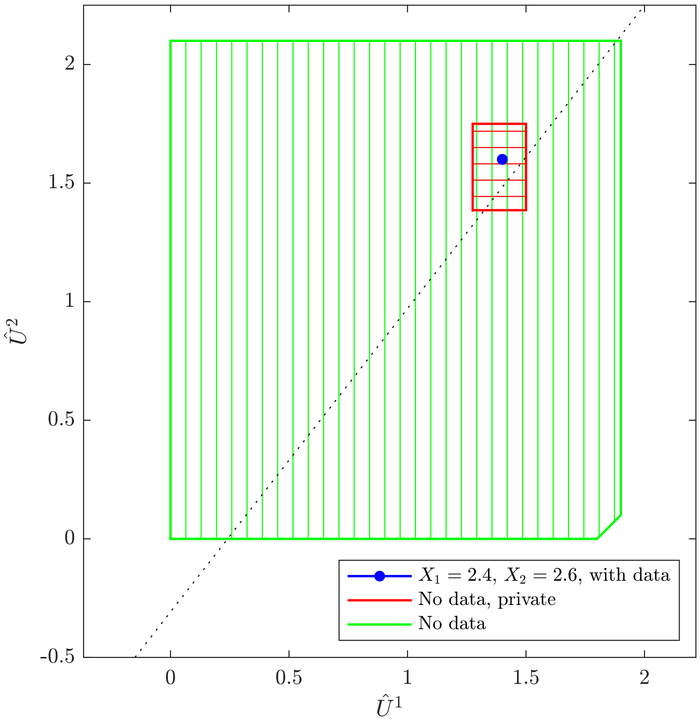

Figure 7: Joint counterfactuals in the entry game.

Third, we computed the joint prediction when we allow all information structures. This most permissive joint prediction for aggregate profit is in green. Again, it is clear that neither game dominates the other.

This example illustrates the potential benefit of combining methodologies: When we use only joint predictions for informationally-robust comparative statics, without a data-based restriction, it is not possible to rank $X = 2.4$ and $X = 2.6$. But when we use data to further refine the joint counterfactual prediction, we do obtain an unambiguous ranking.

8.3 Two-Player Zero-Sum Game

We now consider a setting with two players, binary actions, and binary states. The observed game is the following:

$\theta = 0$			$\theta = 1$		
a_1/a_2	0	1	a_1/a_2	0	1
0	$(2, -2)$	$(-1, 1)$	0	$(0, 0)$	$(-1, 1)$
1	$(-1, 1)$	$(0, 0)$	1	$(-1, 1)$	$(2, -2)$

In each state, the game has the form of an asymmetric matching pennies. Both states are equally likely, so that in expectation the game is symmetric. Thus, if the players have no information about the state, there is a unique equilibrium in which they both randomize with equal probabilities, and both players' payoffs are zero. If they have full information about the state, then there is again a unique (and symmetric) equilibrium in which they play $a = 0$ with probability $1/4$ in state $\theta = 0$, and they play $a = 0$ with probability $3/4$ in state $\theta = 1$. In both states, player 1's payoff is $-1/4$.

We assume that we have observed ϕ exactly, and $\phi(a, \theta) = 1/8$ for all (a, θ). This is the joint distribution of states and actions that arises under no information. In the counterfactual, we multiply all of the payoffs by a factor $2 - z$ in state 0 and by z in state 1, for some $z \in [0, 2]$. This comparative static is equivalent to varying the relative likelihoods of the two states. The observed game corresponds to $z = 1$. The counterfactual outcome of interest is player 1's payoff.

We numerically computed maximum and minimum payoffs for player 1 for a fine grid of z values. The range of counterfactual outcomes under variable and fixed information are depicted in Figure 8 as a function of z. When information is variable, then again, the only thing we learn from the data is that both states are equally likely. The gray lines represent upper and lower bounds on welfare. The range of possible outcomes is largest at $z = 1$, when the counterfactual game is a copy of the observed game. In this case, any payoff in $[-1/2, 1/2]$ can be attained with some information structure. The highest payoff of $1/2$ can be achieved by letting player 1 observe the state and player 2 receiving no information. Under that information, there is an equilibrium where $\widehat{a}_1 = \theta$ and player 2 mixes with equal probabilities. Similarly, the payoff of $-1/2$ can be achieved by giving no information to

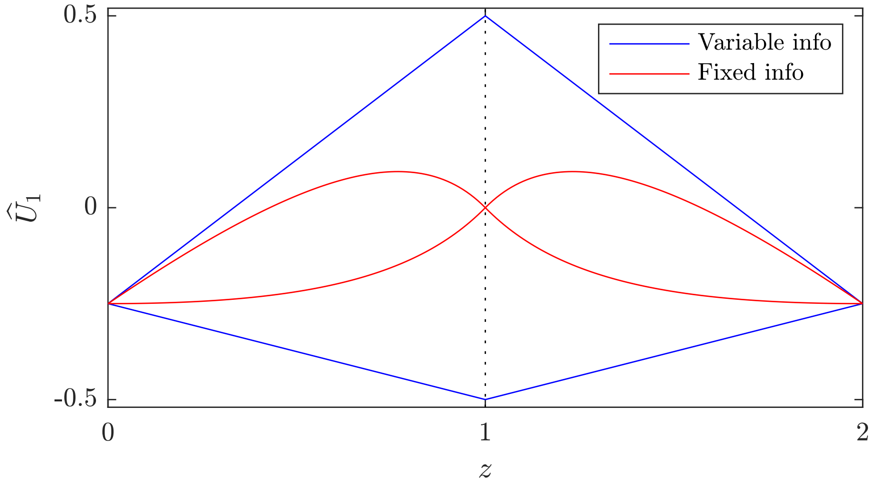

Figure 8: Counterfactual payoffs for player 1 in the zero-sum game.

player 1 and full information to player 2. In fact, it is a result of Peski (2008) that these are the information structures that achieve extreme welfare outcomes in any two-player zero-sum game, and it is not particular to our example.[13]

Note that when $z = 0$ or $z = 1$, then payoffs are zero in one state, so that it is effectively a game with a single state, and thus the value of the game is uniquely pinned down independent of the information.

When we fix information, the range of counterfactual outcomes is tighter. Indeed, when $z = 1$, there is a unique counterfactual prediction when the counterfactual game coincides with the observed game. Once again, this is a general insight that is not particular to our example. In any two-player zero-sum game, if there is an information structure $\mathcal{I}$ and equilibrium σ that rationalizes the observed actions and in which player 1's payoff is u_1, then it must be that the zero-sum game $(\mu, \mathcal{G}, \mathcal{I})$ has a value which is u_1, and hence all equilibria

[13] Here is a sketch of the proof. Player 1's payoff in $(\mathcal{G}, \mathcal{I})$ is at least their maxmin payoff, where the max and min are taken over player 1 and player 2's strategies, respectively. Player 2 has the option to use a strategy that does not depend on their private information t_2, so player 1's maxmin payoff would increase if we restricted player 2 to use only those constant strategies. This is what happens if player 2 has no information. Next, if we look at information structures where only player 1 gets information, then it must be that player 1's payoff is maximized by having as much information is possible. For, any strategy under partial information can be replicated under full information simply by "simulating" the noisy signal, so the effective strategy space is largest under full information. Finally, in the extreme case of full information/no information, the game is finite so the minimax theorem holds, and the maxmin payoff is player 1's equilibrium payoff.

have the same payoffs. This observation completes an analogue of Proposition 4 for zero-sum games:

Proposition 5 (Two-Player Zero-Sum Counterfactuals).
Consider a two-player zero-sum game in which players' observed payoffs are $(u_1, -u_1)$*. If the counterfactual and observed games are the same, then under fixed information, there is point identification of the players counterfactual payoffs, which must be* $(u_1, -u_1)$*. Under variable information, then a tight upper bound on player 1's payoff is given by what is attained when player 1 has full information and player 2 has no information, and a tight lower bound is what is attained when player 1 has no information and player 2 has full information.*

Thus, it is a general phenomenon that there are point predictions for local counterfactuals in two-player zero-sum games under fixed information, although there is generally a fat set of counterfactual predictions under variable information.

Returning now to the particular example, as z moves away from 1, the range of counterfactual payoffs expands, before contracting again as we approach the complete information extremes. Thus, the predictive power of fixed information is large when the counterfactual is closed to the observed game, and it degrades as the counterfactual environment diverges from that which generated the data.

The broad economic conclusion is that player 1 prefers moderate z, while player 2 prefers extreme values.[14] Specifically, when information is fixed and $|z - 1| > 0.58$, then we can unambiguously say that player 1 is worse off and player 2 is better off in the counterfactual than in the observed game. When $|z - 1| \leq 0.58$, then the change in welfare is ambiguous: player 1 may be better off or worse off, depending on the true information structure. A similar statement applies when information is variable, but the conditions for player 1 to be better off are more stringent, and we can unambiguously sign the change in welfare only when $|z - 1| > 2/3$.

[14] As we discuss further in Section 7, an equivalent interpretation is that if we hold $z = 1$ fixed and vary the prior μ, then player 1 prefers large uncertainty about θ ($\mu(\theta)$ close to 1/2 for both θ) and player 2 prefers small uncertainty ($\mu(\theta)$ close to either 0 or 1).

8.4 First-Price Auction

Our final example is a private-values first-price auction (cf. Section 4). This setting is similar to the one studied by Syrgkanis et al. (2018), except that we consider counterfactuals with fixed information, whereas they allow variable information.

Suppose that there are two bidders with values in $V = \{0, 1/9, \ldots, 8/9, 1\}$. We also restrict bids to be in the value grid, and we also assume that bidders do not bid more than their values. There is no reserve price in the auction. Bidders learn at least their own value, but may learn more. We assume that the values are iid uniform, and the econometrician observes either the BCE that minimizes the auction's revenue or the BCE that maximizes revenue (both of which are computed numerically). The counterfactual of interest is revenue as we vary the reserve price. In particular, does there exist a reserve price under which revenue unambiguously increases, relative to the observed game without a reserve price?

Let us first consider the case where the observed outcome was the revenue minimizing BCE. Figure 9 shows how the counterfactual prediction for revenue varies with the reserve price. In particular, the solid red curves represent maximum and minimum counterfactual revenue. There are two features to notice: First, even if the reserve price stays at zero, there is a fat set of counterfactual revenue levels. This indicates that there exist information structures that could induce the revenue minimizing BCE for which there are multiple equilibria, and that revenue varies across these equilibria. So, even if the reserve price does not change, revenue could in principle increase if the bidders coordinated on a different equilibrium. This multiplicity persists at higher reserve prices. However, for moderate reserve prices, the lower bound on revenue increases above the observed level. This lower bound is maximized at 5/9. At this reserve price, we can unambiguously say that regardless of the information and equilibrium, revenue would necessarily be higher than in the observed outcome. Note that since the lowest value is zero, it is necessarily the case that minimum revenue increases when the reserve price changes from 0 to 1/9, although it is not obvious that revenue should continue to increase in the reserve price beyond this point.

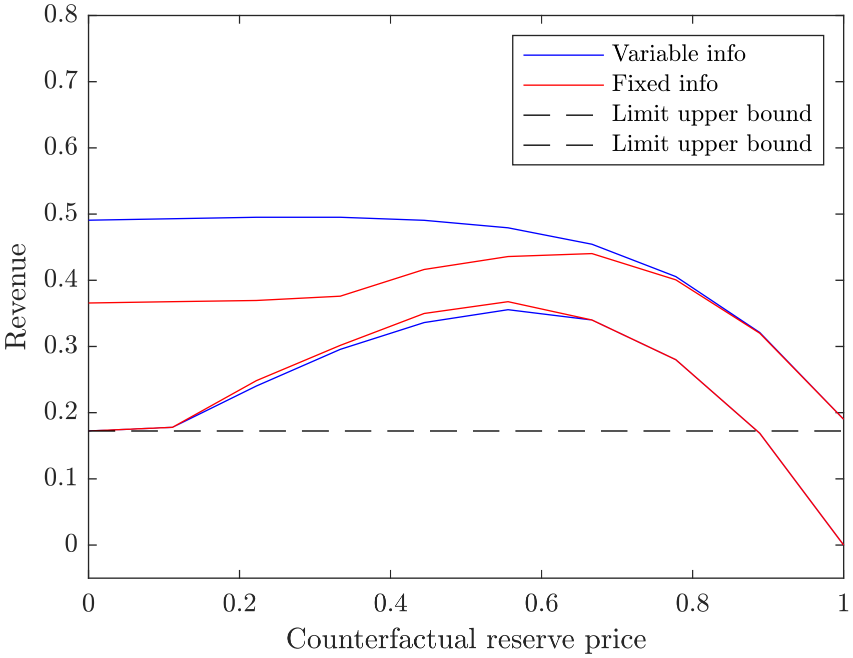

Figure 9: Counterfactual when observed outcome is the revenue minimizing BCE.

Figure 9 also shows how the counterfactual prediction if we allowed information vary, but held fixed the value distribution. For the lower bound, the predictions are not substantively different, although the upper bound on revenue is considerably more permissive. This is not surprising: The simulated data came from the revenue-minimizing information structure, so the fact that the lower red and blue curves nearly coincide is a reflection of the fact that the revenue-minimizing information does not vary significantly with the reserve price.

We next consider the case where the observed outcome is the revenue maximizing BCE. The corresponding counterfactual prediction is depicted in Figure 10. In this case, adding a reserve price cannot lead to a significant increase in revenue, and will necessarily cause revenue to decrease if it the reserve price is sufficiently high. Again, this prediction is substantively the same as what we would obtain with variable information, although in this case it is the lower bound on revenue that is more permissive with variable information. In fact, we can give an analytical justification for both the fact that maximum revenue is (nearly) decreasing in the reserve price, and also the fact that the fixed and variable information bounds coincide. As discussed in Bergemann, Brooks, and Morris (2017, Section 5.4), under the hypothesis that bidders do not bid more than their values, there is an elementary lower bound on bidder surplus, which is the maximum payoff a bidder could obtain if others were bidding their values. With two bidders whose values are exactly uniformly distributed on $[0,1]$, and when the reserve price is r, the lower bound for a bidder with value $v \geq r$ is the maximum of

$$\max_{b\in[r,v]} (v-b)\,v = \begin{cases} \frac{v^2}{4} & \text{if } v \geq 2r; \\ (v-r)\,r & \text{if } r \leq v < 2r. \end{cases}$$

(If $v < r$, the lower bound on bidder surplus is zero.) The lower bound on ex ante bidder surplus when $r \leq 1/2$ is therefore

$$2\left[\int_{v=r}^{2r} (v-r)\,r dv + \int_{v=2r}^{1} \frac{v^2}{4} dv\right] = \frac{1}{6} - \frac{r^3}{3},$$

and when $1/2 \leq r \leq 1$, the lower bound is

$$2\int_{v=r}^{1} (v-r)\,r dv = \left(v^2 - 2rv\right) r = r - 2r^2 + r^3.$$

At the same time, total surplus when the reserve price is r is at most

$$\int_{v=r}^{1} v d\left(v^2\right) = \frac{2}{3}\left(1-r^3\right).$$

Thus, an upper bound on revenue with a reserve price r is

$$\overline{R}\,(r) = \frac{2}{3}\left(1-r^3\right) + \begin{cases} \frac{r^3}{3} - \frac{1}{6} & \text{if } r < \frac{1}{2}; \\ 2r^2 - r - r^3 & \text{if } \frac{1}{2} \leq r \leq 1. \end{cases}$$

We have plotted $\overline{R}$ in green in Figure 10. It is straightforward to verify that this function is decreasing. Moreover, Bergemann, Brooks, and Morris (2017) show that the bound is tight

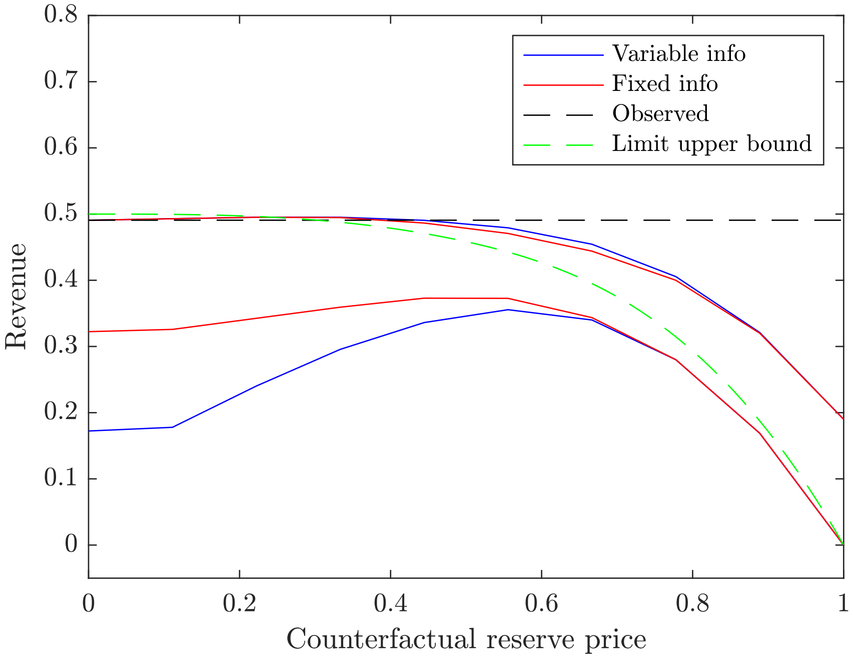

Figure 10: Counterfactual when observed outcome is the revenue-maximizing BCE.

when $r = 0$, meaning that there exists an information structure and equilibrium in which revenue is $\overline{R}(0)$, so that in the limit when the value and bid grids fill in all of $[0, 1]$, maximum revenue must be decreasing in the reserve price. We conjecture that this construction can be generalized to $r > 0$, so that in fact $\overline{R}(r)$ is maximum revenue across all BCE. As a final note, this counterfactual exercise extracts as much from the data as possible about players' information, as it pertains to this particular counterfactual prediction. We may contrast this approach with one suggested by us in our analysis of BCE of interdependent value first-price auctions (Bergemann, Brooks, and Morris, 2017). In that paper, we identified a tight lower bound on the winning bid distribution across all BCE consistent with a given ex post value distribution. We suggested using that bound to partially identify the value distributions that can rationalize observed winning bids. This partially identified set could then be used to generate counterfactual predictions. Such an exercise would allow information to vary between the observed and counterfactual auctions. In contrast, the methodology in the present paper holds information fixed between observation and counterfactual. Our focus is also less on the identification of values than on the identification of information, although we could have also treated the value distribution as a latent variable to be identified from the BCE, in which case we would have been using the entire observed bid distribution to implicitly restrict the value distribution, rather than just the distribution of the winning bid.

8.5 Innocuous Assumptions

Our model imposes a great deal of structure on the environment. In particular, we have assumed that information is described by a single information structure, utilities are known, the prior over the state is held fixed, and there is a single equilibrium that is played in the observed game and a single equilibrium in the counterfactual. At first glance, this structure seems restrictive for empirical applications in which the data is generated by many different instances from the observed game, and where conditions may vary from one instance to another. But, as we will now explain, these assumptions are without loss of generality and could be relaxed at the expense of a richer model.

1. All players receive signals from the same information structure. In practice, players with different characteristics, in different locations, or different points in time may receive qualitatively different forms of information. We may, however, consider these to be special cases of global description of players' information, where the heterogeneity in information is encoded as an extra dimension of signal. For example, suppose that for each $k = 1, \ldots, K$, a fraction $\beta_k \in [0,1]$ of the data is generated when the players have common knowledge that the information structure is $\mathcal{I}^k = \left\{S_1^k, \ldots, S_n^k, \pi^k\right\}$. We could equivalently represent this economy with a new information structure in which $S_i = \sqcup_{k=1}^{K} \{k\} \times S_i^k$, i.e., each player's set of signals is a disjoint union of the k information structures, and

$$\pi\left(X|\theta\right) = \begin{cases} \beta_k \pi^k\left(Y|\theta\right) & \text{if } X = \{k\} \times Y \text{ for some } k; \\ 0 & \text{otherwise.} \end{cases}$$

In words, with probability one, all players get signals in the same S^k, and each k has probability β_k. Our counterfactual prediction implicitly allows for information structures of this form.

2. The utility functions $u_i\left(a,\theta\right)$ are known to the analyst. Uncertainty about preferences can be incorporated by expanding the state space. For example, suppose we start with a state space Θ, a moment restriction $M = \left\{\phi\left(a,\theta\right)\right\}$, and two possible utility functions u^1 and u^2. Then we can expand the state space to $\tilde{\Theta} = \{1,2\} \times \Theta$, utility function $u\left(a,\left(k,\theta\right)\right) = u^k\left(a,\theta\right)$, and the moment restriction is

$$M = \left\{ \tilde{\phi} \in \Delta\left(A \times \tilde{\Theta}\right) \middle| \sum_{k=1,2} \tilde{\phi}\left(a,\left(k,\theta\right)\right) = \phi\left(a,\theta\right) \right\}.$$

Thus, the prevalence of u^1 and u^2 in the population is a free variable, and is partially identified from the data.

3. The distribution over states μ is held fixed in the counterfactual. In fact, we can allow a different distribution $\hat{\mu}$ in the counterfactual, as long as it is absolutely continuous with respect to μ, meaning that it can be written as $\hat{\mu}\left(\theta\right) = \eta\left(\theta\right)\mu\left(\theta\right)$ for some $\eta : \Theta \to \mathbb{R}_+$. For example, when we are only interested in varying the prior and the absolute continuity hypothesis is satisfied, then we can set the counterfactual utility to $\widehat{u}_i\left(a,\theta\right) = \eta\left(\theta\right)u_i\left(a,\theta\right)$, in which case equilibrium utility is simply

$$\begin{aligned}
\sum_{\theta\in\Theta}\int_{s\in S}\sum_{a\in A}\mu\left(\theta\right)\widehat{u}_i\left(a,\theta\right)\sigma\left(a|s\right)\pi\left(ds|\theta\right) &= \sum_{\theta\in\Theta}\int_{s\in S}\sum_{a\in A}\eta\left(\theta\right)\mu\left(\theta\right)u_i\left(a,\theta\right)\sigma\left(a|s\right)\pi\left(ds|\theta\right) \\
&= \sum_{\theta\in\Theta}\int_{s\in S}\sum_{a\in A}\hat{\mu}\left(\theta\right)u_i\left(a,\theta\right)\sigma\left(a|s\right)\pi\left(ds|\theta\right),
\end{aligned}$$

and the represented payoffs are equivalent to those that would obtain with the different prior. This is merely a reflection of the well-known indeterminacy of probabilities versus utilities in the subjective expected utility model, when utilities are state dependent

(Savage, 1954; Anscombe and Aumann, 1963). Indeed, this transformation was being used in the single-player analysis of Section 5, which can be reinterpreted as variations of the prior.

4. All players play the same equilibria of the observed and counterfactual games. This is also without loss of generality. Suppose that the information structure is $\mathcal{I}$, and a share β_k of the data is generated from players who play strategies σ^k for $k = 1, \ldots, K$. The same outcome can be induced with a single information structure $\widetilde{\mathcal{I}}$, in which $\widetilde{S}_i = \{1, \ldots, K\} \times S_i$, $\widetilde{\pi}\left(\{k\} \times X|\theta\right) = \beta_k \pi\left(X|\theta\right)$, and strategies are $\tilde{\sigma}_i\left(a|\left(k, t\right)\right) = \sigma_i^k\left(a|s\right)$. In effect, the first coordinate of the new signal $\widetilde{s}_i$ is a public randomization device which is equal to k with probability β_k. Strategies on the larger space say to play σ^k when $X = k$.