1-1-2019

Counterfactuals with Latent Information

Dirk Bergemann

Benjamin Brooks

Stephen Morris

COUNTERFACTUALS WITH LATENT INFORMATION

By

Dirk Bergemann, Benjamin Brooks, and Stephen Morris

January 2019

Revised February 2019

COWLES FOUNDATION DISCUSSION PAPER NO. 2162R

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§

February 16, 2019

Abstract

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. A counterfactual prediction is desired about behavior in another strategic setting, under the hypothesis that the distribution of 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 type space, is infinite dimensional.

KEYWORDS: Counterfactuals, Bayes correlated equilibrium, information structure, type space, linear program.

JEL CLASSIFICATION: C72, D44, D82, D83.

*We acknowledge financial support through NSF Grant SES 1459899. We are grateful for valuable discussions with Stéphane Bonhomme, Hanming Fang, Tibor Heumann, Vasilis Syrgkanis and Elie Tamer.

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

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

§Department of Economics, Princeton University, smorris@princeton.edu

1 Introduction

The usual approach to economic analysis is to ask how assumptions about exogenous variables—i.e., technology and agents' preferences—translate into implications for endogenous variables—i.e., agents' choices. But when there is uncertainty, agents' choices also depend on their information. When many agents interact strategically, agents' information about others agents' information, and higher-order information, will also matter. The usual approach is to also treat the *type space*, or the structure of information, as known. However, an alternative approach is to see what predictions can be made about agents' choices while being agnostic about agents' information. Bergemann and Morris (2013, 2016) have pursued this approach and report a characterization ("Bayes correlated equilibrium" (BCE)) of the set of outcomes that may arise without making assumptions about the type space.

The usual approach to identification reverses this problem. Given observed endogenous variables (agents' choices), what can we infer about about structural parameters (technology and agents' preferences)? The usual identification exercise is carried out under the assumption of a known type space. However, the type space is often not known. It is therefore natural to ask what partial identification is possible using the BCE characterization of possible outcomes. The possibility of such partial identification has been discussed in Bergemann and Morris (2013) and Bergemann, Brooks, and Morris (2017). This partial identification has been implemented and used to develop counterfactual implications in Magnolfi and Roncoroni (2017) and Syrgkanis, Tamer, and Ziani (2017). In particular, these papers partially identify structural parameters from the data using the BCE characterization, and then ask what BCE could have arisen with that partially identified set in an alternative game. Thus, they assume that structural parameters stay the same in the counterfactual but make no assumption about information. We might call this "counterfactuals with latent and variable information" since it is assumed that the econometrician does not know the type space and the information in the counterfactual is allowed to be different from the information that generated the data.

In this paper, we describe an alternative approach—"counterfactuals with latent and fixed information"—where we continue to assume that the econometrician knows nothing about the type space, but we ask what would have happened if we changed the payoffs but kept the information the same. This corresponds to the classical counterfactual exercise of asking what happens if we change one variable but leave everything else the same.

A potential problem with this approach (not present in the variable information approach) is that the space of type spaces is a complex object. In a single-agent context, information can be canonically represented as a distribution over distributions over states, which is an infinite dimensional vector space even when the number of states is finite. In the multi-agent context, it is canonically represented as a distribution over belief hierarchies in the universal type space. Given the high dimensionality of information, a common approach is to assume a particular functional form, e.g., affiliated private values in the context of auction models. The dimension reduction then facilitates identification and counterfactuals, at the cost of additional assumptions which may be more for tractability than realism or economic substance.

Despite this complexity, we explore a completely non-parametric approach to the (partial) identification of agents' information. We cannot escape the high dimensionality of information, and a direct description of the type spaces that rationalize observed behavior—i.e., explicit partial identification—would be impractical. Instead, we fix a specific counterfactual that is of interest. We argue that observed behavior can be treated as an implicit restriction on behavior in the counterfactual economy, so that the counterfactual prediction remains finite dimensional as long as the underlying action and state spaces are also finite.

Let us give a semi-formal discussion for the special case of a single-agent decision problem. Suppose that an agent must take 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. As θ is uncertain, so too may be the agent's information about θ. But we maintain the standard assumption

that the agent's 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 such agents, 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, we can 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 the agent *knew* 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 agent observes a signal t, and the distribution of t conditional on θ is known to follow a conditional probability law $\pi(t|\theta)$. By adding a prior over θ and applying Bayes' rule, this experiment will induce a distribution over interim beliefs. An experiment rationalizes the observed data if an expected utility maximizing agent who observed the signal would optimally behave in a way that results in the observed distribution ϕ.

Now suppose we wish to predict how the same agent 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 agent's information (i.e., the Blackwell experiment) remain the same.[1] Note that 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.

One 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 resulting $\widehat{\phi}$ in the

[1]The assumption of a constant prior distribution over θ is nearly without loss of generality, see the discussion in Section 5.

counterfactual. But in the Blackwell model, the signals are an abstract set, and we have not even assumed a particular space in which these signals should live. For single-agent decision problems, there are canonical signal spaces. For example, an equivalent representation of the set of experiments is the set of distributions over distributions over θ whose expectation is the prior. But the space of such distributions is infinite dimensional, and it is not obvious whether or how it can be approximated with finite structures.

Instead, we propose to skip the identification step and proceed directly to counterfactual predictions. This can be done as follows. Imagine that rather than performing an abstract thought experiment, the agent did in fact choose $\widehat{a}$ at the same time as a was chosen, and we simply did not observe it. The agent's payoffs were simply the sum across the two decision problems, so that there was no interaction between the two decisions 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 $\overline{\phi}$ which are consistent with Bayesian rationality. But this problem has already been solved: The solution concept of Bayes correlated equilibrium (Bergemann and Morris, 2016) 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 provides linear restrictions. These constraints can then be used to compute counterfactual welfare outcomes by simply solving linear programs.

Our discussion above considers the special case of a one player finite action game (i.e., a decision problem) where the distribution over fundamentals is observed. Our analysis shows that this logic goes through in general many player finite action games, where arbitrary data about player of the game and fundamentals is revealed. For example, it might be that fundamentals are not observed but the distribution of actions are observed. Or it might be that only some statistics of players' actions (such as the winning bid) is observed. The argument that set of feasible counterfactuals is characterized by a set of linear inequalities is completely general.

In order to illustrate the logic of the approach, we study some examples showing how holding information fixed tightens predictions relative the variable information approach used in existing empirical work. To sharply illustrate this, we study counterfactuals when fundamentals are known as a proof of concept for the method. We illustrate how in a single agent example and also in a two player zero sum game (which has a unique correlated equilibrium), there is continuity of counterfactuals under our fixed information approach. That is, small changes in parameters lead to a small range of counterfactual predictions. If one had taken the variable information approach in these examples, there a large set of counterfactuals even if there is no change in parameters, corresponding to new completely different information. On the other hand, in a two player two action game with multiple equilibria, there is already a fat set of counterfactuals even if there are no changes in parameters. This is because even with a fixed information structure about fundamentals, different correlation of actions can give rise to different equilibria. However, the set of counterfactuals remains much smaller under fixed information than under variable information.

We study counterfactuals with latent but fixed fundamentals. In related work, Heumann (2018) is offering "informationally robust" comparative statics in a fixed game with incomplete information where the information structure is unknown. Heumann (2018) carries out his analysis 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-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. Section 4 illustrates our theorem with three examples for the case where fundamentals are known. Section 5 is a discussion of the theorem, and Section 6 briefly concludes.

2 Preliminaries

The economy consists of N agents, indexed by $i = 1, \ldots, N$. Agents' 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 prior *distribution* over states is denoted $\mu \in \Delta(\Theta)$.

The players interact through a *game form*, 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 *type space*, denoted by $\mathcal{T}$, which consists of the following objects. Each player has a measurable set of *types* T_i, with $T = \times_{i=1}^{N} T_i$ denoting the set of type profiles, and there is a conditional probability measure $\pi : \Theta \to \Delta(T)$ over type profiles as a function of the state. Thus $\mathcal{T} = \left((T_i)_{i=1}^{N}, \pi \right)$.[2]

A *Bayesian game* is a tuple $(\mu, \mathcal{G}, \mathcal{T})$. A *strategy* for player i in the Bayesian game is a measurable mapping $\sigma_i : T_i \to \Delta(A_i)$. We write $\sigma_i(a_i|t_i)$ for the probability of an action a_i given the type t_i. A strategy profile is denoted $\sigma = (\sigma_1, \ldots, \sigma_N)$, and is associated with the product mapping $\sigma : T \to \Delta(A)$, where $\sigma(a|t) = \times_{i=1}^{N} \sigma_i(a_i|t_i)$. Player i's expected utility under the strategy profile σ is

$$U_i(\sigma) = \sum_{\theta \in \Theta} \int_{t \in T} \sum_{a \in A} u_i(a, \theta) \, \sigma(a|t) \, \pi(dt|\theta) \, \mu(\theta).$$

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 alternative 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 $(\mu, \mathcal{G}, \mathcal{T})$ if

$$\phi(a, \theta) = \int_{t \in T} \sigma(a|t) \, \pi(dt|\theta) \, \mu(\theta).$$

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

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

[2]Note that we allow the type space 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.

It is a theorem of Bergemann and Morris (2013, 2016) that ϕ is a BCE of $(\mu, \mathcal{G})$ if and only if there exists a type space $\mathcal{T}$and a Nash equilibrium σ of $(\mu, \mathcal{G}, \mathcal{T})$ such that ϕ is induced by σ.

We will sometime abuse notation by saying that an outcome ϕ is a Bayes correlated equilibrium of the game form $\mathcal{G}$ if it is a Bayes correlated equilibrium of the basic game $(\mu, \mathcal{G})$ for some μ.

3 Counterfactuals when Information is Latent and Fixed

We will use the preceding framework to study counterfactual predictions, in which the sets of possible states, actions and agents' preferences are assumed to be known, but the distribution μ, the type space $\mathcal{T}$ and the agents' strategies are latent parameters. In particular, we fix a state space Θ, and we suppose that the agents interact in two distinct game forms, which we will distinguish as the (partially) *observed game* $\mathcal{G}$ and the *unobserved game* $\widehat{\mathcal{G}}$. 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 $M \subseteq \Delta(A \times \Theta)$; (ii) it was generated under some prior μ and type space $\mathcal{T}$, and (iii) it was induced by a Nash equilibrium of $(\mu, \mathcal{G}, \mathcal{T})$. We ask which outcomes $\hat{\phi}$ could be induced by some equilibrium of $\left(\mu, \widehat{\mathcal{G}}, \mathcal{T}\right)$?

Formally, an outcome $\widehat{\phi} \in \Delta\left(\widehat{A} \times \Theta\right)$ is a *counterfactual prediction* if there exist μ, $\mathcal{T}$, and Nash equilibria σ and $\widehat{\sigma}$ of $(\mu, \mathcal{G}, \mathcal{T})$ and $\left(\mu, \widehat{\mathcal{G}}, \mathcal{T}\right)$, respectively, such that the outcome ϕ induced by σ is in M and such that $\widehat{\phi}$ is induced by $\widehat{\sigma}$. 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:

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 known. It is only the information 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.

Our main result is the following characterization of $\widehat{\Phi}(M)$. We denote by $\bar{\mathcal{G}}$ the following *linked game*, where player i's actions are $\overline{A}_i = A_i \times \widehat{A}_i$, and preferences are represented by

$$\overline{u}_i(\overline{a}, \theta) = u_i(a, \theta) + \widehat{u}_i(\widehat{a}, \theta),$$

where $\overline{a}_i = (a_i, \widehat{a}_i)$ for all i.[3]

We refer to $\mathcal{G}$ and $\widehat{\mathcal{G}}$ as the *component games* of the linked game. Note that a Bayes correlated equilibrium $\overline{\phi}$ of $\left(\Theta, \mu, \overline{\mathcal{G}}, \mathcal{T}\right)$ can be identified with a joint distribution in $\Delta\left(A \times \widehat{A} \times \Theta\right)$.

[3]Aumann and Dreze (2008) introduce a construction, termed "doubled game", in a game with complete information. By "doubled game" they refer to an artificial game where they introduce two copies of each player's action and use the correlated equilibria of this artificial game to characterize the set of interim payoffs consistent with common knowledge of rationality and the common prior assumptionsh. Our linked game instead has players choosing actions from each of two distinct games. Like Aumann and Dreze (2008), we use the (Bayes) correlated equilibria as a device to answer a novel substantive economic question, in this case characterizing counterfactuals.

Theorem 1 (Counterfactual Predictions)**.**

An outcome $\hat{\phi}$ is in $\widehat{\Phi}(M)$ if and only if there exists a Bayes correlated equilibrium $\overline{\phi}$ of $\overline{\mathcal{G}}$ for which (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 1. Fix a type space $\mathcal{T}$. Any strategy profile $\overline{\sigma}$ in the linked game can be identified with a pair of strategy profiles σ and $\widehat{\sigma}$ in the observed and unobserved game, where $\sigma_i(\cdot|t_i)$ is the marginal of $\overline{\sigma}(\cdot,\cdot|t_i)$ on A_i and $\widehat{\sigma}_i(\cdot|t_i)$ is the marginal on $\widehat{A}_i$.

Claim: $\bar{\sigma}$ is a Nash equilibrium of $\left(\mu, \overline{\mathcal{G}}, \mathcal{T}\right)$ if and only if σ and $\widehat{\sigma}$ are Nash equilibria of $(\mu, \mathcal{G}, \mathcal{T})$ and $\left(\mu, \widehat{\mathcal{G}}, \mathcal{T}\right)$, respectively. This follows from the identity

$$\begin{aligned}
\overline{U}_i(\overline{\sigma}) &= \sum_{\theta\in\Theta} \int_{t\in T} \sum_{\overline{a}\in\overline{A}} \overline{u}_i(\overline{a},\theta)\,\overline{\sigma}(\overline{a}|t)\,\pi(dt|\theta)\,\mu(\theta). \\
&= \sum_{\theta\in\Theta} \int_{t\in T} \left[\sum_{a\in A} u_i(a,\theta)\,\sigma(a|t) + \sum_{\widehat{a}\in\widehat{A}} \widehat{u}_i(\widehat{a},\theta)\,\widehat{\sigma}(\widehat{a}|t)\right] \pi(dt|\theta)\,\mu(\theta) \\
&= U_i(\sigma) + \widehat{U}_i(\widehat{\sigma}).
\end{aligned}$$

Thus, if $\bar{\sigma}$ is not a Nash equilibrium, then there exists $\bar{\sigma}_i'$, which is associated with marginal strategies σ_i' and $\widehat{\sigma}_i'$, such that

$$\begin{aligned}
U_i(\sigma) + \widehat{U}_i(\widehat{\sigma}) &= \overline{U}_i(\overline{\sigma}) \\
&< \overline{U}_i(\overline{\sigma}', \overline{\sigma}_{-i}) \\
&= U_i(\sigma_i', \sigma_{-i}) + \widehat{U}_i(\widehat{\sigma}_i', \widehat{\sigma}_{-i}),
\end{aligned}$$

so that at least one of σ_i' or $\widehat{\sigma}_i'$ must be a profitable deviation. Similarly, if there is a profitable deviation in one of the component games, say to σ_i' for player i in the observed game, then the product strategy defined by $\overline{\sigma}_i'(a_i, \widehat{a}_i|t_i) = \sigma_i'(a_i|t_i)\,\widehat{\sigma}_i(\widehat{a}_i|t_i)$ is a profitable deviation in the linked game.

With the claim in hand, we now prove the if direction of the theorem. As is known, $\overline{\phi}$ is a Bayes correlated equilibrium of the game form $\bar{\mathcal{G}}$ if and only if there exists a μ, $\mathcal{T}$, and Nash

equilibrium $\overline{\sigma}$ of $\left(\mu, \overline{\mathcal{G}}, \mathcal{T}\right)$ that induces $\overline{\phi}$. As a result, the marginal strategies σ and $\widehat{\sigma}$ of $\overline{\sigma}$ are also Nash equilibria of the component games. Moreover, they induce Bayes correlated equilibria ϕ and $\widehat{\phi}$, which are simply marginals of $\overline{\phi}$. Thus, if the marginal of $\overline{\phi}$ on $A \times \Theta$ is in M, then ϕ is in M, so that $\widehat{\phi} \in \widehat{\Phi}(M)$.

Now the only if direction. If for some μ and $\mathcal{T}$ there are Nash equilibria σ and $\widehat{\sigma}$, which induce outcomes $\phi \in M$ and $\widehat{\phi}$, respectively, then the latter is in $\widehat{\Phi}(M)$. Then the aforementioned product strategy $\overline{\sigma}$ of σ and $\widehat{\sigma}$ is a Nash equilibrium of the linked game, and it induces an Bayes correlated equilibrium $\overline{\phi}$ whose marginals are ϕ and $\widehat{\phi}$. This completes the proof. □

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, then $\widehat{\Phi}(M)$ is also a polytope, namely the projection 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 marginal 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 programming problems. We will use this fact in the examples in the next section.

4 Three Examples

We now present three examples to illustrate the content of our theorem. In order to focus on the logic of our approach, our examples all concern the case where the distribution over fundamentals is revealed by the data and the interesting question is how the data constrains information in the counterfactual. Our examples show that fixing the information structure

dramatically reduces the set of possible counterfactuals, relative to a variable information approach. We start with a single agent decision problem, then consider two-person zero-sum gane, and then consider an entry game. We leave for future work the question of how holding the information structures fixed reduces the set of possible counterfactuals in more realistic empirical settings where the distribution of fundamentals is also not known.

4.1 Single Agent Decision Problem

Let us begin with a single-agent decision problem. This is the special case we discussed in the Introduction. The possible states are $\Theta = \{0, 1\}$, and both states are equally likely. The set of actions is similarly $A = \{0, 1\}$. In the observed single-agent game, the agent gets a payoff of 1 if and only if the action matches the state, 0 otherwise:

$$u(a, \theta) = \begin{cases} 1 & \text{if } a = \theta; \\ 0 & \text{otherwise.} \end{cases}$$

In the counterfactual game, the agent gets a payoff of $2 - z$ from matching state 0, and a payoff of z from matching state 1, where $z \in [0, 2]$:

$$\hat{u}(a, \theta) = \begin{cases} 2 - z & \text{if } a = \theta = 0; \\ z & \text{if } a = \theta = 1; \\ 0 & \text{otherwise.} \end{cases}$$

Thus, in both games, the agent wants to match the state, and z controls the relative value of matching state 0 versus state 1 in the counterfactual. The parameter value $z = 1$ corresponds to the observed game. If the agent knew the state perfectly, then we would see $a = \widehat{a} = \theta$ with probability one and a payoff of 1 (in both observed and counterfactual games). If the agent knew nothing about the state, then the agent strictly prefers $\widehat{a} = 0$ if $z < 1$, strictly

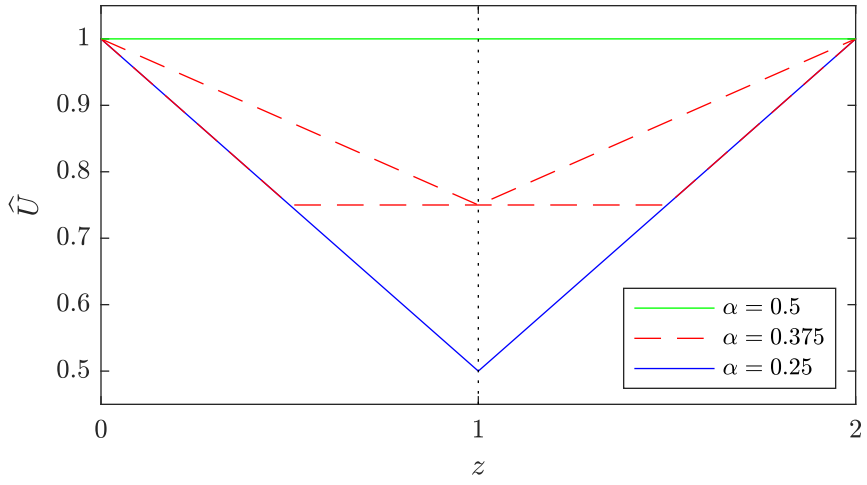

Figure 1: Counterfactual welfare in the binary example.

prefers $\widehat{a} = 1$ if $z > 1$, and is indifferent at $z = 1$. The payoff under no information is therefore $\max\{(2-z)/2, z/2\}$.

We suppose that the joint distribution of (a, θ) is perfectly observed and is given by the symmetric distribution:

$$\phi(a, \theta) = \begin{cases} \frac{1}{2}\alpha & \text{if } a = \theta; \\ \frac{1}{2} - \alpha & \text{if } a \neq \theta. \end{cases}$$

Note that we must have $\alpha \in [1/2, 1]$: if $\alpha < 1/2$, the agent could achieve a higher average payoff in the observed game by picking one action and playing it all the time.

We ask, what are the possible payoffs for the agent in the counterfactual game? First, consider what counterfactual predictions we can make if we allow information to be different in the counterfactual. Then the only thing we can learn from the data is that both states are equally likely, and thus α contains no useful information. Since welfare in single-agent decision problems is monotonically increasing in the Blackwell order, we know that full information and no information achieve the maximum and minimum possible payoffs, respectively. Intermediate information must result in a payoff between these bounds. There is no more we can say. These are the highest and lowest solid lines in Figure 1.

Next, consider what happens if we hold information fixed. We compute the convex set of possible counterfactual welfare outcomes for various values of α and z. The results are plotted in Figure 1. Suppose first that $z = 1$, so that the observed and unobserved decision problems are the same. This is a kind of "local" counterfactual exercise. In this case, there is a point prediction for welfare. In fact, it is the same welfare as in the observed game. Why? The agent has the option to use information the same way in the counterfactual game as in the observed game, and guarantee a weakly higher payoff. A symmetric argument says that observed welfare must be at least the counterfactual welfare, so the two are equal. But observed welfare is known exactly from ϕ.

Now consider $z \neq 1$. When $\alpha = 0.5$, there is a point prediction for all z, given by the top-most line in Figure 1. For in this case, observed actions and states are perfectly correlated. This is only *feasible* if the agent learns the state perfectly. Then in any counterfactual, since the agent knows the state and prefers to match, we must see that states and actions are still perfectly correlated, and the agent achieves the optimal payoff of 1. Similarly, when $\alpha = 0.25$, observed actions and states are uncorrelated, which can only be rationalized by the agent having no information. In the counterfactual, the agent can do no better than play a pure action, which results in the minimum payoff of $\max\{(2-z)/2, z/2\}$.

The counterfactual prediction for the intermediate case of $\alpha = 0.375$ is the interval between the dashed lines (the lower dashed line coincides with the no-information prediction for z that are close to either 0 or 1). There is generally a fat set of welfare outcomes, except for $z \in \{0, 1, 2\}$. The reason is that there is a range of information structures that can rationalize the agent's behavior, and these have different implications in the counterfactual. Here are two possible explanations for the fact that the agent guesses the state correctly 3/4 of the time:

(a) Half of the time, the agent perfectly learns the state and plays the correct action all the time, and the other half of the time, the agent learns nothing and randomizes with equal probabilities.

(b) With probability one, the agent gets a noisy observation of the state that is correct with probability 3/4, and the agent plays an action equal to the realized signal.

Both of these experiments rationalize the data, but they have different implications for the counterfactual. In fact, they attain the numerically computed bounds. With information as in (a), the optimal strategy is to match the state when it is revealed, attaining 1, and to play the optimal pure action when the state is not revealed, attaining $\max\{(2-z)/2, z/2\}$. The average of these two yields the upper bound. With information as in (b), as long as z is close to 1, it is still optimal to play the action equal to the signal. But for extreme z, it becomes better to play a pure action, thus giving the payoff of $\max\{(2-z)/2, z/2\}$.

The takeaway from this example is that information can have a great deal of predictive power in single-agent counterfactual analysis, especially when the counterfactual environment does not differ too much from the one that was observed. In fact, the tightness of the prediction in local counterfactuals, and the gap between fixed and variable information, will be true in any single-agent decision problem. We summarize this observation in the following result:

Proposition 1 (Single-Agent Counterfactuals)**.**
Consider a single-agent decision problem where the agent's observed payoff is u. If the counterfactual and observed decision problems are the same, then under fixed information, there is point identification of the agent's counterfactual payoffs, which must be u. Under variable information, then a tight upper bound on the agent's payoff is given by what is attained under full information, and a tight lower bound is what is attained with no information.

In the next two examples, we show how the same basic idea can be applied to multi-agent games where not just beliefs about θ but also higher order beliefs are implicitly identified in the process of making counterfactual predictions.

4.2 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 $z \in [0, 2]$. The observed game corresponds to $z = 1$. Since the game is zero sum, the only 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 2 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 type space. 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,

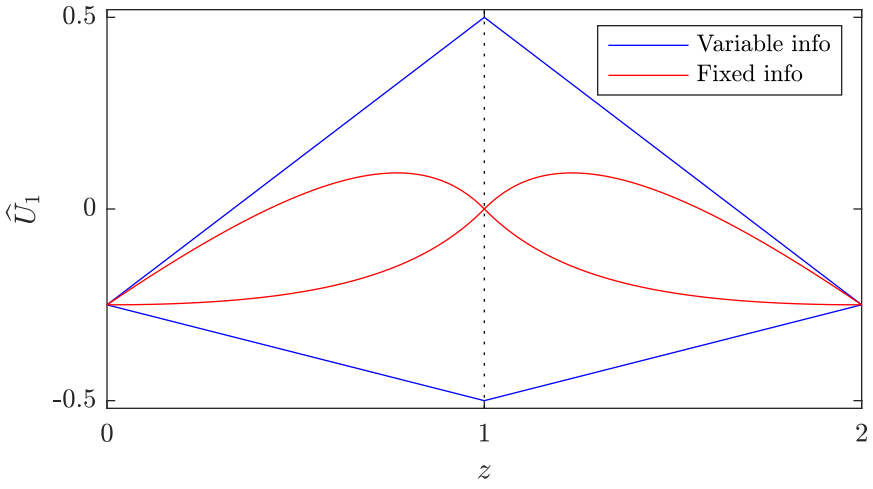

Figure 2: Counterfactual payoffs for player 1 in the zero-sum game.

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 player 1 and full information to player 2. In fact, it is a result of Peski (2008) that these are the type spaces that achieve extreme welfare outcomes in any two-player zero-sum game, and it is not particular to our example.[4]

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 a type space $\mathcal{T}$ and equilibrium σ

[4]Here is essence of the proof. Player 1's payoff in $(\mathcal{M}, \mathcal{T})$ is at least his 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 his 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 type spaces 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, this maxmin is only a lower bound on player 1's payoff. But in the extreme case of full info/no info, the bound is clearly unimprovable, since the game is finite so an equilibrium exists, which must have a value equal to the maxmin (Osborne and Rubinstein, 1994, Proposition 22.2).

that rationalizes the observed actions and in which player 1's payoff is u_1, then it must be that the zero-sum game $(\mu, \mathcal{G}, \mathcal{T})$ has a value which is u_1, and hence all equilibria have the same payoffs. This observation completes an analogue of Proposition 1 for zero-sum games:

Proposition 2 (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, but that predictive power 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.[5] 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 type space. 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 $|z-1| > 2/3$.

[5]As we discuss further in Section 5, 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).

4.3 Entry Game

For our last example, we consider a simple entry game. Two firms choose whether or not to enter a market. The firms that enter compete in quantities and face the downward sloping demand curve $P = \max\{\theta - Q\}$, where P is the market price, Q is aggregate output, and $\theta \in \Theta = \{1, 2, 3\}$ is an unknown state of demand. All demand intercepts are equally likely. Production cost is zero, but there is a fixed cost $F \geq 0$ to enter the market.[6]

If no firms enter, both firms' profits, consumer surplus, and total welfare are all zero.

If one firm enters, the monopoly outcome obtains, in which the entering firm sets $q_i = \theta/2$, the market price is $\theta/2$, the entering firm's profit is $\theta^2/4 - F$, the firm that does not enter has a payoff of 0, consumer surplus is $\theta^2/8$, and total welfare is $3\theta^2/8 - F$.

If two firms enter, there the subgame has a unique equilibrium in which $q_1 = q_2 = \theta/3$, the market price is $P = \theta/3$, both firms earn profits $\theta^2/9 - F$, consumer surplus is $4\theta^2/9$, and total welfare is $2\theta^2/3 - 2F$. We assume that this equilibrium is played in the event that both firms enter, which is essentially imposing sequential rationality.

We assume that $F = 3/4$ in the observed game. As a result, it is profitable for one or two firms to enter when $\theta = 3$, it is profitable for one but not two firms to enter when $\theta = 2$, and it is unprofitable for any firms to enter when $\theta = 1$. We again assume that ϕ is observed, and has the following form: Both firms enter when $\theta = 2$; exactly one firm enters when $\theta = 2$, and both are equally likely to be the entering firm; and no firms enter when $\theta = 1$. Each firm's observed profit is $1/8$, observed consumer surplus is $3/2$, and observed total welfare is $7/4$. This outcome is consistent with the firms having complete information about θ, and playing a symmetric pure-strategy equilibrium when $\theta = 2$.

We numerically computed maximum and minimum welfare, consumer surplus, and profit, for a range of counterfactual entry fees between 0 and $5/2$. The results are depicted in Figure

[6]Note that the uncertainty here is about a common demand shock, and firms have complete information about the symmetric entry cost. In the entry games analyzed in much of the literature, e.g., Ciliberto and Tamer (2009) and Magnolfi and Roncoroni (2017), the uncertainty is about firms' entry costs, which are privately known. Our example is chosen for its simplicity, but private entry costs can easily be incorporated into our framework, as we discuss in Section 5.

3. Once again, we have included the benchmark of variable information to see how the bounds improve when we impose fixed information in the counterfactual.

Note that if $F > 9/4$, then it is unprofitable to enter even as a monopolist in the best state, so that all outcomes must be zero irrespective of information. Similarly, if $F < 1/9$, then it is profitable to enter even if $\theta = 1$ and there is another firm already in the market, so that both firms always enter and there is again a point prediction for welfare, in both the fixed and variable information regimes.

For $F \in [1/9, 9/4]$, there is generally a fat set of welfare outcomes, even for fixed information. This is true even when the counterfactual entry cost is $3/4$, i.e., the counterfactual game is the same as the observed game. This is at least partly due to the fact that the entry game generally has multiple equilibria which are not payoff equivalent. For example, even under complete information (which is one way to rationalize the data), there are pure strategy equilibria which could generate the data, but there is also a mixed strategy equilibrium in which the firms enter when $\theta = 2$ with independent probabilities. Our counterfactual prediction ranges over *all* equilibria, and hence there will be multiple counterfactual welfare outcomes. This kind of multiplicity is similar to that which arises in the literature on entry games with unknown equilibrium selection, such as Ciliberto and Tamer (2009). Our methodology gives rise to an additional source of multiplicity, which is that we implicitly give players access to rich correlating devices. Thus, even in a complete information game, our counterfactual prediction would range over all *correlated* equilibria, as well as Nash equilibria. This also occurs in Magnolfi and Roncoroni (2017) and Syrgkanis, Tamer, and Ziani (2017).

Even so, the counterfactual prediction is far from vacuous. For example, we may ask, how low does the entry fee have to go before we are *certain* that total welfare will be higher than what we observed? Under fixed information, the answer is that if $F < 0.54$, then expected welfare must be at least $7/4$ in all equilibria and in all type spaces that can rationalize the data, where $7/4$ is total welfare in the hypothetical data. If we allowed information to vary, the corresponding threshold is 0.41. Similarly, if the entry fee rose above 0.81, then total

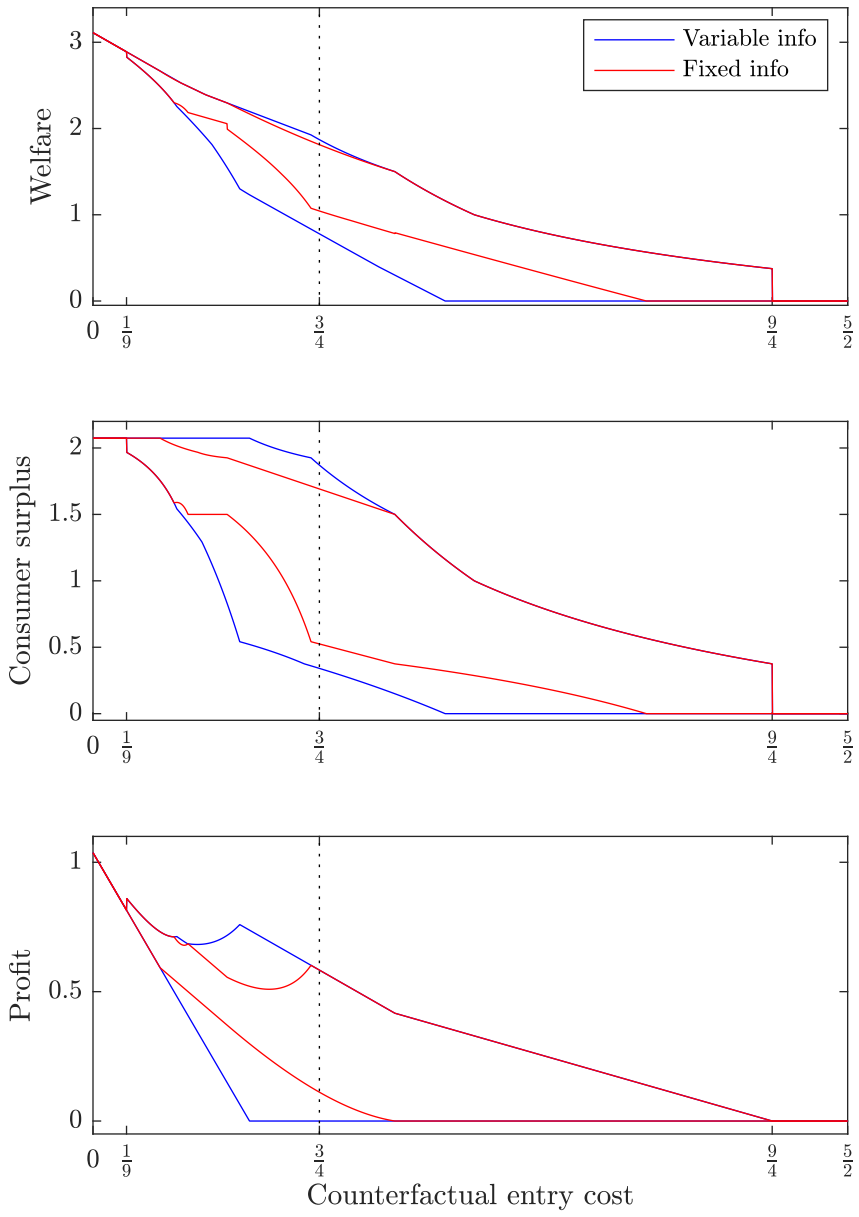

Figure 3: Counterfactual welfare in the entry game.

welfare must unambiguously fall, regardless of the type space or equilibrium. This number is the same whether we hold information fixed or allow it to vary.

5 Discussion

We now turn to a discussion of Theorem 1. In particular, we will discuss the benefits from the implicit identification approach versus more direct partial identification. We will compare our work to the related literature on counterfactuals with variable information. We discuss how many of the assumptions of our model can be relaxed or are simply normalizations that are without loss of generality. And finally, we discuss how additional bounds of information can be incorporated into the model.

5.1 Complexity and the Partially Identified Set

As we mentioned in the introduction, it is important to note that while these counterfactual exercises are finite dimension, the latent parameter of the type space is infinite dimensional. Indeed, we have not even specified a particular space in which the types live, so we cannot even construct a "set of all type spaces" that rationalize the data.[7] Nonetheless, in the single-agent case, all that matters for behavior is the agent's interim beliefs about θ. We could without loss of generality restrict attention to type spaces in which $T = \Delta(\Theta)$ and the types are normalized to be equal to the interim belief. Even so, the agent'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 distribution of beliefs, we can always find a 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-agent decision problems with continuous utilities can be well approximated in this topology. We could in principle compute the set of rationalizing type

[7] This would be a bit like a "set of all sets." If a set of all type spaces can be constructed, does it contain a type space whose types are the type spaces which are not types in themselves?

spaces whose support lives in a finite grid in $\Delta(\Theta)$ and presumably have a fair approximation of behavior in counterfactual games.

In the multi-agent case, the canonical space for types 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-agent 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 type spaces, is dense in the universal type space in the product topology. Finite types, however, 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 type spaces that rationalize multi-agent outcomes. Moreover, the universal type space only encodes agents' 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.

5.2 Counterfactuals when Information is Latent and Variable

As we briefly mentioned in the introduction, Magnolfi and Roncoroni (2017) and Syrgkanis, Tamer, and Ziani (2017) also uses BCE for partial identification and counterfactual prediction. Translated to our language, they partially identify the prior distribution μ, the distribution over states, which is in turn used for counterfactual prediction. They do not, however, use the information revealed about the type space in the counterfactual. This makes their counterfactual predictions more robust. But if we take the usual meaning of "counterfactual"—what would have happened if we had changed an aspect of the game, leaving everything else fixed—it is not clear why one would want to allow the information to

change. Because we hold the type space fixed, our counterfactual prediction is necessarily tighter.

One way of contrasting the approaches is consider the case where μ is known. Our fixed information approach is still useful, since data reveals information about the type space that can be used to narrow down predictions in the counterfactual. In the variable information approach, data from the first problem does not restrict behavior in the counterfactual. We illustrated this case in our examples in the previous section.

It remains true in the variable information case that it is not necessary to be explicit about the partial identification of the prior distribution μ. Syrgkanis, Tamer, and Ziani (2017) emphasize that identifying counterfactuals reduces to solving a linear programming problem, and they heavily use this fact for computation and estimation. However, if we are only interested in partially identifying μ, the partially identified set is itself a finite-dimensional polytope Syrgkanis, Tamer, and Ziani (2017). In this sense, the computational advantage of bypassing explicit partial identification is less significant in the variable information case.

5.3 Nominal Assumptions that are Without Loss of Generality

Our model assumes a great deal of structure on the environment. These assumptions are considerably weaker than they appear at first glance. We discuss them in turn.

1. All agents receive signals from the same type space. In practice, agents 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 variations of "representative" 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 agents have common knowledge that the type space is $\mathcal{T}^k = \left\{T_1^k, \ldots, T_n^k, \pi^k\right\}$. We could equivalently represent this economy with a new type space in which $T_i = \cup_{k=1}^K \{k\} \times T_i^k$, i.e., each player's set of

types is a disjoint union of the k type spaces, 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 agents get signals in the same T^k, and each k has probability β_k. Our counterfactual prediction implicitly allows for type spaces 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 $M=\{\phi\left(a,\theta\right)\}$, 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

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

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 $\hat{u}_i(a,\theta) = \eta(\theta)\, u_i(a,\theta)$, in which case equilibrium utility is simply

$$\begin{aligned}\sum_{\theta\in\Theta}\int_{t\in T}\sum_{a\in A}\mu(\theta)\,\hat{u}_i(a,\theta)\,\sigma(a|t)\,\pi(dt|\theta) &= \sum_{\theta\in\Theta}\int_{t\in T}\sum_{a\in A}\eta(\theta)\,\mu(\theta)\,u_i(a,\theta)\,\sigma(a|t)\,\pi(dt|\theta)\\ &= \sum_{\theta\in\Theta}\int_{t\in T}\sum_{a\in A}\hat{\mu}(\theta)\,u_i(a,\theta)\,\sigma(a|t)\,\pi(dt|\theta),\end{aligned}$$

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-agent and zero-sum counterfactuals of Sections 4.1 and 4.2, which can be reinterpreted as variations of the prior.

4. All agents play the same equilibria of the observed and counterfactual games. This is also without loss of generality. Suppose that the type space is $\mathcal{T}$, and a share β_k of the data is generated from agents who play strategies σ^k for $k = 1,\ldots,K$. The same outcome can be induced with a single type space $\tilde{\mathcal{T}}$, in which $\tilde{T}_i = \{1,\ldots,K\}\times T_i$, $\tilde{\pi}(\{k\}\times X|\theta) = \beta_k\pi(X|\theta)$, and strategies are $\tilde{\sigma}_i(a|(k,t)) = \sigma_i^k(a|t)$. In effect, the first coordinate of the new signal $\tilde{t}_i$ 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 $X = k$.

5. There is a single data source, from the game $\mathcal{G}$. In practice, there could be more than one game form for which we have data. This could be easily incorporated into the framework by expanding the "linked" game into a "$K+1$-tupled" game, where the first K games correspond to observed outcomes, and the $K+1-$th game is the counterfactual. The key feature would be that payoffs are additively separable across games, although there is no harm in imposing correlation in outcomes across the observed games.

5.4 Bounds on Information

As in Bergemann and Morris (2013, 2016), we may also consider counterfactual predictions under stronger assumptions about what information is available to the agents. For example, in an auction setting, we may wish to impose that each bidder knows their own value for the good being sold but has only partial information, of an unknown form, about others' values and information.

More broadly, we may suppose that the type space is at least as informative as some $\underline{\mathcal{T}}$, in the sense described in Bergemann and Morris (2016). That paper gives a richer definition of BCE with respect to $\underline{\mathcal{T}}$, which is essentially a joint distribution in $\Delta\left(A \times \underline{T} \times \Theta\right)$, where $\underline{T}$ is the set of minimal signal profiles, such that for each i and (a_i, t_i), a_i is a best response to the conditional distribution of (a_{-i}, θ) given (a_i, t_i). Returning to the private value auctions example, we could take $\Theta = \mathbb{R}^N$, the set of value profiles, $\underline{T}_i = \mathbb{R}$, where $\pi\left(\cdot|v\right)$ puts probability one on $t = v$, so that each bidder's type is equal to their value. This is the structure imposed in Syrgkanis, Tamer, and Ziani (2017). Magnolfi and Roncoroni (2017) use a similar structure, where agent i's coordinate is interpreted as a private cost of entering a market. We can incorporate such a lower bound on information into our counterfactual prediction by simply requiring that $\overline{\phi}$ be a BCE of the linked game with respect to $\underline{\mathcal{T}}$. The proof of our theorem can be replicated almost verbatim with the minor modification of integrating over types in $\underline{\mathcal{T}}$.

Ideally, one would like to incorporate upper bounds on information as well, for example, imposing that players are not too informed about the state. This could either take the form "the type space $\mathcal{T}$ is not at least as informative as $\overline{\mathcal{T}}$" or "the type space $\overline{\mathcal{T}}$ is at least as informative as $\mathcal{T}$." Unfortunately, we know of no simple way to incorporate such restrictions while preserving the linear structure and low dimensionality of the counterfactual. It is possible, however, to incorporate upper bounds on information in the form of conjectures about how agents will behave in certain situations. For example, we could reinterpret the single-agent example of Section 4.1 as follows. There is no data, and no observation. There

is only a conjecture that *if* the agent were to play the game with $z = 1$, he or she would guess the state correctly with a probability α. This represents both a lower bound on information (the agent knows enough about θ to be correct with probability at least α) and an upper bound on information (the agent does not know so much about θ to be correct with probability greater than α). We could have instead dropped the former assumption and set M to be the set of outcomes such that the agent is correct with probability no greater than α. Our counterfactual predictions would then respect a crude upper bound on information. By expanding on this idea, it is possible to generate fairly flexible upper bounds on information, e.g., in the auction context, one could impose that bidders are unable to guess others' values with a high degree of accuracy. In light of our point above about having multiple data sources, there is no conceptual difficulty in combining multiple such conjectures with data sources.

6 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 quite large, especially compared to what can be predicted if we do not fix information between observation and counterfactual.

It is easy to think of reasons why information should *not* be held fixed in a counterfactual. 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? Nonetheless, the fixed information benchmark is a useful starting point, 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 agents' 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.

Aumann, R. and J. Dreze (2008): "Rational Expectations in Games," *American Economic Review*, 98, 72–86.

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.

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

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

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.

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

Heumann, T. (2018): "Informationally Robust Comparative Statics in Incomplete Information Games," Tech. rep., HEC Montréal.

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

Magnolfi, L. and C. Roncoroni (2017): "Estimation of Discrete Games with Weak Assumptions on Information," Tech. rep.

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

Osborne, M. and A. Rubinstein (1994): *A Course in Game Theory*, Cambridge: MIT Press.

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

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.

Syrgkanis, V., E. Tamer, and J. Ziani (2017): "Inference on Auctions under Weak Assumptions on Information," Tech. rep., Harvard University.