On the "Limited Feedback" Foundation of Boundedly Rational Expectations*

Ran Spiegler†

November 9, 2015

Abstract

A common justification for boundedly rational expectations is that agents receive partial feedback about the equilibrium distribution. I formalize this idea in the context of the "Bayesian network" representation of boundedly rational expectations, presented in Spiegler (2015). According to this representation, the decision maker forms his beliefs as if he fits a subjective causal model - captured by a directed acyclic graph (DAG) over the set of variables - to the objective distribution. When the causal model is misspecified, the belief systematically distorts the objective distribution's correlation structure. I show that when the DAG is perfect, the representation is the outcome of extrapolating from limited feedback - in the form of a long spreadsheet with randomly missing values - using a variant on a familiar imputation technique. When the DAG is imperfect, this foundation breaks down.

*An earlier version of the paper was circulated under the title "Bayesian Networks and Missing-Data Imputation". This paper has benefitted from ESRC grant no. ES/L003031/1. I am grateful to Noga Alon, Yair Antler, Simon Byrne, Philip Dawid, Kfir Eliaz, Ehud Lehrer and many seminar participants for helpful comments.

†Tel Aviv University and University College London. URL: http://www.tau.ac.il/~rani. E-mail: rani@post.tau.ac.il.

1 Introduction

Equilibrium models in economics typically assume that agents have "rational expectations" - i.e. they have full understanding of statistical regularities in the steady state. Recent years have seen intensified interest in equilibrium models with "boundedly rational expectations", in which agents' subjective beliefs systematically distort the correlation structure of the steady-state distribution. The distortions take various forms: "coarse" beliefs that neglect correlations, because agents omit certain variables from their subjective model or because they clump contingencies into "analogy classes" (Piccione and Rubinstein (2003), Jehiel (2005), Koessler and Jehiel (2008), Mullainathan et al. (2008) and Eyster and Piccione (2013), Woodford (2013)); failure to realize how correlation between actions and consequences would change if the agent played an off-equilibrium action (Esponda (2008) and Esponda and Pouzo (2014)); belief in spurious correlations due to naive extrapolation from small samples (Osborne and Rubinstein (1998); attributing fluctuations of a certain variable to the wrong cause (Eyster and Rabin (2005), Ettinger and Jehiel (2010)); etc.

A common justification for the idea of equilibrium with boundedly rational expectations is that agents receive *partial feedback* as they try to learn statistical regularities in their environment. Even if the social learning process reaches a steady state, agents have not received enough feedback that would enable them to develop a perfect understanding of the correlation structure of the steady-state distribution; as a result, their subjective beliefs distort the correlation structure. In some cases (e.g. Osborne and Rubinstein (1998), Jehiel (2005)), this idea is formally built into the definition of equilibrium. Other papers (e.g. Schwartzstein (2014), Esponda and Pouzo (2015)) incorporate limited feedback into dynamic learning models.

This paper offers a new perspective into the "limited feedback" justification for the notion of boundedly rational expectations as systematic distortion of objective correlations. Rather than modeling a dynamic learning

process, I formalize the notion of limited feedback statically, as a *dataset with missing values*. The dataset consists of observations that are drawn from the steady-state distribution, yet these observations are not complete and contain (randomly generated) missing values. The decision maker's expectations are the outcome of a procedure he employs to extrapolate a probabilistic belief from the dataset. Systematic distortions of the true correlation structure thus originate from the process that generates missing values in the dataset, as well as the extrapolation procedure.

1.1 A Motivating Example: Vitamin Pills

A decision maker (DM henceforth) considers taking vitamin pills to improve his health. He wishes to base his decision on statistical evidence. The data that is available to him comes in the form of a long spreadsheet that contains (infinitely) many joint observations of vitamin-pill consumption (denoted c) and blood-vitamin level (denoted v), as well as (infinitely) many joint observations of blood-vitamin level and a health indicator (h). Crucially, the spreadsheet contains *no* joint observations of c and h. This type of data limitation is common in the areas of pharmaceutics and nutrition.

Suppose that every observation in the spreadsheet reflects an independent realization of a true steady-state probability distribution p over c, v, h. because the spreadsheet contains infinitely many observations of (c, v) and (v, h), the DM ends up learning the objective marginal distributions $p(c, v)$ and $p(v, h)$. However, this knowledge does not pin down p - typically, the known marginals are consistent with infinitely many joint distributions over c, v, h. How does the DM respond to this indeterminacy?

I assume that the DM attempts to extend his partial knowledge to a *fully specified* probability distribution over observable variables. One reason behind this objective is that the DM wants to base his decision on rudimentary statistical analysis: scatterplots, regressions, etc. These tools can only be applied to *rectangular* datasets, namely observations without missing values.

Indeed, applied statisticians and econometricians regularly confront datasets with missing values, and their usual modus operandi is to "*rectangularize their dataset*", such that the processed dataset does not contain incomplete observations. There are various methods for achieving this goal, from omitting observations with missing values (an impractical method in our example, where *all* observations contain missing values) to employing systematic procedures for *imputing* the missing values. For a textbook on statistical analysis with missing data, see Little and Rubin (2002).[1]

In the context of our story, the following procedure is an intuitive method for imputing missing values in the DM's spreadsheet. In any observation where h (c) is missing, impute a random draw from $p(h \mid v)$ ($p(c \mid v)$). This conditional distribution is derived from the learned marginal $p(v, h)$ ($p(c, v)$).

Here is a schematic illustration of the DM's spreadsheet before and after the imputation:

c	v	h
+	+	-
-	+	+

Before

c	v	h
+	+	*
*	+	+

After

That is, the original spreadsheet consists of two blocks of observations - one in which values of h are missing, and another in which values of c are missing (plus and minus signs represent available and missing values, respectively). The rectangularized spreadsheet contains no missing values. The imputed values of c (h) are generated by extrapolating the distribution over c, v (v, h) in the top (bottom) block to the bottom (top) block (the star sign represents imputed values).

[1] An alternative response to the indeterminacy is to conduct worst-case analysis, treating the set of distributions that are consistent with the known marginals as the set of priors in Gilboa and Schmeidler's (1989) max-min expected utility theory.

The imputation procedure is a non-parametric variant on a familiar method known as "*stochastic regression*" (Little and Rubin (2002, Ch. 4)). However, I wish to emphasize that my motivation in this exercise is not prescriptive but *behavioral*. This is the main feature that differentiates this paper from the statistics literature on inferring from partial datasets. I do not ask how a professional statistician *should* handle datasets with missing values. Instead, I ask how a non-professional who nevertheless wishes to ground his decision in rudimentary data analysis *would intuitively* extrapolate from partial data. And I am particularly interested in whether these intuitive methods of extrapolation might lead to "boundedly rational expectations" that display systematic distortion of objective correlations. Lacking empirical evidence on how non-professionals handle datasets with missing values, I settle for the thought experiment given by our "vitamins" example. The fact that the intuitive extrapolation method that I propose resembles a "certified" method is interesting, but not essential for the example's appeal.

The DM's belief - i.e., the distribution he would use for decision analysis - is given by the frequencies of c, v, h in the rectangularized spreadsheet. Let us calculate these frequencies. By construction, the distribution in the top block (where values of h were originally missing) is $p(v,h)p(c \mid v)$, whereas the distribution in the bottom block (where values of c were originally missing) is $p(c,v)p(h \mid v)$. Using the definition of conditional probabilities and a bit of algebra, we can see that the two expressions are in fact identical. One way of rewriting them is $q(c,v,h) = p(c)p(v \mid c)p(h \mid v)$. This expression factorizes the true distribution p as if it were consistent with a (probabilistic) causal chain $c \rightarrow v \rightarrow h$ - i.e., vitamin-pill consumption is a primary cause of blood-vitamin level, which in turn is the sole immediate cause of health.

Of course, this is an "as if" argument: the DM's belief is consistent with imposing a particular causal structure on p, but it does not follow that the DM truly believes in this or any other causal mechanism. From a statistical point of view, q merely embodies the property that h and c

are independent conditional on v. This should come as no surprise: the DM arrived at the belief q by extrapolating the missing values of c and h independently of each other, conditioning only on observed realizations of v. If the true distribution p violates this conditional independence property, then q systematically distorts the correlation structure of p, in a way that looks *as if* the DM is trying to fit a causal model (represented by the diagram $c \rightarrow v \rightarrow h$) to the true distribution p. This is an example of what I will refer to as a "*Bayesian-network representation*" of boundedly rational expectations.

The particular form of the DM's subjective belief has potentially interesting behavioral implications. The following argument is taken from Spiegler (2015), where it appears under the title "The Dieter's Dilemma" - I refer the reader to that paper for the full analysis. Suppose that the objective distribution p has the property that c and h are statistically independent. Thus, vitamin pills have no health effects. However, both variables are correlated with v: blood-vitamin deficiency is associated with both poor health and low vitamin-pill intake. This correlation pattern is consistent with the causal scheme $c \rightarrow v \leftarrow h$ (i.e., c and h are independent causes of v). In this sense, the DM's belief q exhibits *reverse causation*: the causal model implicit in q posits that v causes h, whereas the causal model implicit in p posits a causal link in the opposite direction.

If the DM had rational expectations, he would realize that c has no effect on h, and would not consume costly vitamin pills. In contrast, the subjective belief that the DM extrapolates from his dataset implies that c *is* correlated with h. Moreover, higher vitamin-pill consumption is observationally associated with higher blood-vitamin levels, which in turn are associated with better health. And since the causal model implicit in q interprets these estimated correlations in terms of the causal chain $c \rightarrow v \rightarrow h$, the DM concludes that consuming vitamin pills will have an indirect positive effect on his health. If the pills are not too costly, the DM's erroneous belief will

impel him to sub-optimally consume vitamin pills.

Furthermore, the strength of the estimated indirect effect of c on h is sensitive to consumption frequencies in the dataset. Think of each observation in the dataset as a description of the behavior and health outcomes of other agents. The more these agents consume pills, the weaker the estimated correlation between v and h, and therefore the smaller the inferred health benefits of vitamin-pill consumption. This suggests that we should think of individual behavior in this example as an *equilibrium* notion: in equilibrium, agents' consumption patterns (given by $p(c)$) are consistent with individual optimization with respect to the subjective belief q that is extrapolated from the dataset induced by the equilibrium distribution itself. Spiegler (2015) showed that the equilibrium can be unique and "mixed" - i.e., although we are dealing with single-agent decision making, there could be a unique equilibrium in which the DM mixes over consumption levels.

Let us summarize the lessons from our example. Our DM had limited feedback regarding steady-state correlations, in the form of missing values in a large spreadsheet. He extrapolated his partial data into a fully specified probability distribution, using an imputation procedure akin to the "stochastic regression" method. The resulting belief had a "Bayesian-network representation" - i.e., it looked as if the DM were trying to fit a (potentially misspecified) causal model to the true distribution. And the resulting belief distortions led to sub-optimal actions and non-trivial equilibrium effects.

1.2 Plan of the Paper

In general, a Bayesian-network representation is based on the notion of a *directed acyclic graph* (DAG henceforth), defined over a set of nodes which is the set of variable labels. For instance, in the vitamin-pills example, the nodes would represent the three variables c, v, h. Following Pearl (2009), we can interpret the DAG as a causal model: a directed link from one node to another captures a direct causal relation between the variables that these nodes

represent. A Bayesian-network representation (also referred to as a *DAG representation*) is a subjective distribution over all variables that factorizes the true distribution p according to the DAG - i.e., it fits the causal model given by the DAG to p. Thus, the distribution $p(c)p(v \mid c)p(h \mid v)$ factorizes p according to the DAG $c \to v \to h$, while the distribution $p(c)p(h)p(v \mid c, h)$ factorizes p according to the DAG $c \to v \leftarrow h$. The vitamin-pills example established that the former representation can be justified as the outcome of a natural method of extrapolation from partial data. My objective in this paper is to examine the generality of this proposition.

Why should we be interested in DAG representations of boundedly rational expectations? First, I find the idea that people arrive at their subjective beliefs by filtering empirical regularities through the prism of a subjective causal model interesting and highly intuitive by itself. Indeed, Sloman (2009) provides experimental evidence in favor of the psychological validity of this idea. Second, Spiegler (2015) shows that several existing notions of equilibrium with boundedly rational expectations (including a subset of the references at the beginning of the paper) can be reformulated as instances (or refinements) of a DAG representation (I refer the reader to Section 5 in Spiegler (2015) for details, in order to avoid duplication). In other words, the DAG representation functions as a unifying framework for equilibrium models with boundedly rational expectations. .

In Spiegler (2015), I adopt a literal interpretation of Bayesian-network representations: the DM has an explicit subjective causal model, which he actively fits to the objective distribution without testing for model misspecification. The focus of that companion paper is on decision making under this representation of subjective beliefs. However, as the vitamins example demonstrates, the representation also has an "as if" interpretation: the DM in our story had no explicit causal model; his attempt to extrapolate a subjective belief from limited feedback produces a belief distortion that only *looks* as if he reasons in terms of a causal model. This paper is an attempt

to develop this "as if" interpretation.

Generalizing the vitamin-pills example to environments with arbitrarily many variables and more complex patterns of limited feedback involves two components. First, I extend the notion of a dataset with missing values. The DM has access to infinitely many draws from the objective distribution p, but in each draw he gets to observe a subset of the variables. The missing-value process that determines this subset is independent of p - this assumption is known in statistics as "Missing Completely at Random". Second, I extend the "stochastic regression" imputation procedure by applying it *iteratively*. In each round of the iterative procedure, the DM rectangularizes *part* of the spreadsheet, using the same method as in the vitamin-pills example. As the procedure evolves, the DM expands the rectangularized portion of the spreadsheet, until it contains no missing values.

The main result is that whenever the DAG is *perfect* (i.e., all direct parents of an individual node are mutually connected), the DAG representation has a "limited feedback foundation" - that is, there is a missing-value process such that if we apply the iterative imputation procedure to the dataset induced by the missing-value process and any objective distribution p, we obtain a final belief that factorizes p according to the DAG. Conversely, if the DAG is imperfect, this foundation breaks down: for every missing-value process we can find an objective distribution p, such that the imputation procedure's final output will not factorize p according to the DAG. I use this result to discuss the limited-feedback foundation of various types of boundedly rational expectations.

Of course, the foundation is limited to the notions of limited feedback and extrapolation that I assume in this paper. Therefore, it is interesting to check whether the conclusions would be different if we modified these notions. An existing result from the Artificial Intelligence literature (reformulated to fit our present purposes) implies that when we apply a different extrapolation method known as "maximal entropy", the limited-feedback foundation for

perfect-DAG representaitons persists.

At any rate, this paper is an *example* of a larger research agenda: find tractable, interpretable representations of boundedly rational expectations, and check whether they can be justified as the outcome of "extrapolation from limited feedback". Hopefully the example will stimulate other researchers to develop this direction.

2 Datasets and Imputation

Let $X = X_1 \times \cdots \times X_n$ be a finite set of *states*, where $n \geq 2$. I refer to x_i as a *variable*. Let N be the set of variable *indices*. For many purposes, it will be simplest to set $N = \{1, ..., n\}$. However, in examples, it is often useful to notate indices such that their relation to the variables is more transparent. (In some cases, I will use the variable labels themselves as indices, somewhat absuing notation, in order to make causal diagrams easier to read.)

Let $p \in \Delta(X)$ be an *objective probability distribution* over states. For every $S \subseteq N$, denote $x_S = (x_k)_{k \in S}$ and $X_S = \times_{k \in S} X_k$, and let p_S denote the marginal of p over X_S. Consider a decision maker (DM henceforth) who obtains an infinite sequence of independent draws from p. However, for each draw, he only gets to see the realized values of a subset of variables $S \subset N$, where S is independently drawn from a probability distribution $\sigma \in \Delta(2^N)$, referred to as the *missing-values process*. The pair (p, σ) constitutes the DM's *dataset*, and an observation is a random draw from $p \circ \sigma$. Denote the support of σ by $\mathcal{S}$, and $|\mathcal{S}| = m$. The infinite-sample assumption ensures that the DM gets to learn p_S for every $S \in \mathcal{S}$. I impose two restrictions on $\mathcal{S}$. First, $\mathcal{S}$ is a *cover* of N - that is, all variables are observable (I discuss this assumption below). Second, no two subsets $S, S' \in \mathcal{S}$ contain one another (this assumption is made purely because it simplifies notation at certain points).[2]

[2]We could formalize the notion of a dataset more explicitly, as an infinite sequence

An iterative imputation procedure

The DM's task is to extend any dataset (p, σ) into a fully specified subjective probability distribution over X. He performs this task using an iterative extension of the "stochastic regression" method described in Section 1.1. The iterative imputation procedure consists of $m-1$ rounds. The initial condition of round $k = 1, ..., m-1$ is a pair (B^{k-1}, q^{k-1}), where $B^{k-1} \subseteq N$ and $q^{k-1} \in \Delta(X_{B^{k-1}})$. In particular, B^0 is an arbitrary member of $\mathcal{S}$, and $q^0 = p_{B^0}$. Round k consists of two steps:

Step 1: Select $S^k \in \arg\max_{S \in \mathcal{S} - \{B^0, ..., S^{k-1}\}} \left|S \cap B^{k-1}\right|$. *Let* $B^k = B^{k-1} \cup S^k$.

Step 2: Define two auxiliary distributions over X_{B^k}*:*

$$\begin{aligned} q_1^k(x_{B^k}) &\equiv q^{k-1}(x_{B^{k-1}}) \cdot p(x_{S^k - B^{k-1}} \mid x_{S^k \cap B^{k-1}}) \\ q_2^k(x_{B^k}) &\equiv p(x_{S^k}) \cdot q^{k-1}(x_{B^{k-1} - S^k} \mid x_{S^k \cap B^{k-1}}) \end{aligned}$$

and then, define $q^k \in \Delta(X_{B^k})$ *as follows:*

$$q^k \equiv \frac{\sum_{i=1}^{k-1} \sigma(S^i)}{\sum_{i=1}^{k} \sigma(S^i)} q_1^k + \frac{\sigma(S^k)}{\sum_{i=1}^{k} \sigma(S^i)} q_2^k$$

If $k = m-1$*, the procedure is terminated and its output (namely, the DM's final belief) is* $q^{m-1} \in \Delta(X)$*. If* $k < m-1$*, switch to round* $k+1$*.*

This procedure describes a process by which the DM gradually completes his dataset. By the end of round $k-1$, the DM has "rectangularized" the part of the original dataset that consisted of observations of x_S, $S \in$

(x^t, S^t), where x^t is a random draw from p; S^t is an independent random draw from σ; and the value of x_i^t is missing if and only if $i \notin S^t$. This alternative definition would capture more directly the idea of a dataset as a long spreadsheet with missing values. However, for our purposes, identifying the dataset with the reduced form (p, σ) is w.l.o.g and more convenient to work with. The more elaborate definition would be appropriate for extensions of the model, e.g. when the DM's sample is finite.

$\{B^0, S^1..., S^{k-1}\}$. He has thus transformed the (infinitely many) observations of the variable sets $B^0, S^1..., S^{k-1}$ into an infinite collection of "observations" that induce a fully specified distribution over $X_{B^{k-1}}$, where B^{k-1} is the union of $B^0, S^1..., S^{k-1}$. In step 1 of round k, the DM looks for a *new* set of variables $S^k \in \mathcal{S}$ having *maximal overlap* with B^{k-1}. The rationale for this "maximal overlap" criterion is that the DM tries to extrapolate as little as possible and make the most of observed correlations.

In Step 2 of round k, the DM extends the "rectangularization" of the dataset from the variable set B^{k-1} to the variable set $B^k = B^{k-1} \cup S^k$, using the same method as in the vitamin-pills example. That is, he extrapolates q^{k-1} to the observations of x_{S^k} (where the variables $B^{k-1} - S^k$ are missing), and he extrapolates p_{S^k} to the "observations" of $x_{B^{k-1}}$ (where the variables $S^k - B^{k-1}$ are missing).

This two-part process can be described in terms of the "spreadsheet" metaphor. *(i)* Consider a row in the spreadsheet, in which the set of variables with reported values is B^{k-1}, and the value is some $x_{B^{k-1}}$. Note that the value is not entirely raw data, since it may contain imputed values from previous rounds of the iterative procedure. The variables in $S^k - B^{k-1}$ are missing in this observation. The DM imputes a value $x_{S^k - B^{k-1}}$ for them; this value is a random draw from $p(x_{S^k - B^{k-1}} \mid x_{S^k \cap B^{k-1}})$. The resulting frequency of x_{B^k} across all rows of this kind is given by q_1^k. *(ii)* Consider a row in the spreadsheet, in which the set of variables with reported values is S^k, and the value is some x_{S^k}. This value *is* raw data. The variables in $B^{k-1} - S^k$ are missing in this observation. The DM imputes a value $x_{B^{k-1} - S^k}$ for them; this value is a random draw from $q^{k-1}(x_{B^{k-1} - S^k} \mid x_{S^k \cap B^{k-1}})$. The resulting frequency of x_{B^k} across all rows of this kind is given by q_2^k. The DM produces the distribution q^k over x_{B^k} as a weighted average of q_1^k and q_2^k, according to the relative number of rows of each kind.

The iterative imputation procedure is in my opinion the most natural extension of the intuitive method discussed in Section 1.1. In each round k,

the DM extrapolates two blocks into a rectangular component of the spreadsheet, just as in the vitamin-pills example. One of these blocks (covering the variable set S^k) consists of raw data; for $k > 1$, the other block (covering the variable set B^{k-1}) is the outcome of previous rounds of the procedure and thus contains previously imputed values. The variable set S^k is selected to maximally intersect B^{k-1}. The "maximal overlap" criterion is the only restriction on the order in which the procedure unfolds.

In Section 1.1, I pointed out that the imputation procedure employed by the DM is akin to a familiar method known as "stochastic regression". The literature also contains iterative versions of this method, which extend it to multivariate environments. However, the iterative versions I am familiar with proceed one variable at a time (e.g., Gelman and Hill (2006, Ch. 25), whereas the present procedure proceeds one observation-block at a time, in a sequence that obeys the maximal overlap criterion.

The following examples illustrate the procedure.

Example 2.1: Partitional datasets
Suppose that $\mathcal{S}$ is a *partition* of N. Then, in any round k, $B^{k-1} \cap S = \varnothing$ for every $S \notin \{B^0, S^1, ..., S^{k-1}\}$. Therefore, S^k can be selected arbitrarily. The auxiliary distributions q_1^k and q_2^k are both equal to $q^{k-1}(x_{B^{k-1}})p(x_{S^k})$. It follows that in any round k, $q^k(x_{B^k}) = p(x_{B^0})p(x_{S^1}) \cdots p(x_{S^k})$. The procedure's final output is

$$q^{m-1}(x) = \prod_{S \in \mathcal{S}} p(x_S) \tag{1}$$

We see that the belief that emerges from the DM's procedure is a product of the marginal objective distributions he gets to learn.

Example 2.2: Vitamins revisited
Consider the following elaboration of the vitamins example from the Introduction. In addition to the variables c, v, h, there is a fourth variable s, representing a "surrogate marker" that may be correlated with the health indicator. The DM's dataset consists of many joint observations of the following

variable pairs: (c, v), (v, s) and (s, h). In this example, I use δ_y to denote the index of any variable $y = c, v, s, h$. Thus, $\mathcal{S} = \{\{\delta_c, \delta_v\}, \{\delta_v, \delta_s\}, \{\delta_s, \delta_h\}\}$. Here is a schematic illustration of the original spreadsheet:

c	v	s	h
+	+	-	-
-	+	+	-
-	-	+	+

Let us implement the iterative imputation procedure. In round 1, select the initial condition arbitrarily to be $B^0 = \{\delta_c, \delta_v\}$. By the maximal-overlap criterion of Step 1, The only legitimate continuation is $S^1 = \{\delta_v, \delta_s\}$, such that $B^1 = \{\delta_c, \delta_v, \delta_s\}$. Imputing the missing values of c and s is done exactly as in the original example. Using the "spreadsheet" metaphor, by the end of round 1 the DM has "rectangularized" the part of his spreadsheet in which he originally observed c, v or v, s. He has replaced these raw observations with "manufactured observations" of the triple c, v, s, and the joint distribution over this block can be written as $q^1(c, v, s) = p(c)p(v \mid c)p(s \mid v)$ - or, equivalently, as $q^1(c, v, s) = p(s)p(v \mid s)p(c \mid v)$.

The following is a schematic illustration of how the spreadsheet looks at the end of each of the two rounds.

c	v	s	h
+	+	*	-
*	+	+	-
-	-	+	+

Round 1

c	v	s	h
+	+	*	**
*	+	+	**
**	**	+	+

Round 2

In the second and final round, $S^2 = \{\delta_s, \delta_h\}$, such that $B^2 = N$, and

$$\begin{aligned} q_2^2(c,v,s,h) &= p(s,h) \cdot q^1(c,v \mid s) = p(s,h)\frac{q^1(c,v,s)}{q^1(s)} \\ &= p(s,h)\frac{p(s)p(v \mid s)p(c \mid v)}{\sum_{c'}\sum_{v'} p(s)p(v' \mid s)p(c' \mid v')} \\ &= p(s,h)\frac{p(s)p(v \mid s)p(c \mid v)}{p(s)} = p(s,h)p(v \mid s)p(c \mid v) \\ &= p(s)p(v \mid s)p(c \mid v)p(h \mid s) = q^1(c,v,s) \cdot p(h \mid s) = q_1^2(c,v,s,h) \end{aligned}$$

Thus, $q_1^2 = q_2^2 = q^2$, which can be written as

$$q^2(c,v,s,h) = p(c)p(v \mid c)p(s \mid v)p(h \mid s) \tag{2}$$

This expression is consistent with the following causal chain: vitamin-pill consumption causes blood-vitamin level, which is turn causes the surrogate marker, which in turn causes health.

Discussion

I conclude this section with three comments. First, suppose that we relax the assumption that $\mathcal{S}$ is a cover of N, such that some variables are unobservable. In this case, the procedure's final output q^{m-1} will be a probability distribution over the union of the members of $\mathcal{S}$. The DM's final belief will neglect all unobservable variables. In this sense, the assumption that $\mathcal{S}$ is a cover of N is without loss of generality: we can simply redefine $x_1, ..., x_n$ as the collection of *observable* variables.

Second, the imputation procedure is entirely non-parametric and invariant to the variables' meaning. This makes sense when X lacks intrinsic structure. If, however, variables get real values and the DM has a-priori reasons to hypothesize a monotone relation between two variables, then it would be plausible to incorporate this hypothesis into the extrapolation process. Thus, while the procedure's generality makes it applicable to any dataset, it also means that the procedure will not necessarily be the most plausible method

of extrapolation in certain contexts.

Finally, to illustrate the role of the "maximal overlap" criterion in the imputation procedure, suppose that in the extended vitamin-pills example, the DM ignored the "maximal overlap" criterion and picked $S^1 = \{\delta_s, \delta_h\}$ in the first round. The output of this round would be $B^1 = N$ and $q^1(c, v, s, h) = p(c, v)p(s, h)$. In the second round, we would have $q_1^2 = q^1$, and $q_2^2(c, v, s, h) = p(v, s)q^1(c, h \mid v, s)$. As a result, the procedure's final output would be

$$q^2 = (\sigma\{\delta_c, \delta_v\} + \sigma\{\delta_s, \delta_h\})q^1 + \sigma\{\delta_v, \delta_s\}q_2^2$$

This expression can be written as follows:

$$q^2(c, v, s, h) = q^1(c, h \mid v, s)\left[\sigma\{\delta_v, \delta_s\}p(v, s) + (\sigma\{\delta_c, \delta_v\} + \sigma\{\delta_s, \delta_h\})q^1(v, s)\right]$$

Consider an objective distribution p under which both c and h are independently distributed variables, whereas the variables v and s are mutually correlated. Then, $q^1(c, v, s, h) = p(c)p(v)p(s)p(h)$. Therefore,

$$q^2(c, v, s, h) = p(c)p(h)\left[\sigma\{\delta_v, \delta_s\}p(v, s) + (\sigma\{\delta_c, \delta_v\} + \sigma\{\delta_s, \delta_h\})p(v)p(s)\right]$$

This expression underestimates the objective correlation between v and s. By comparison, under the same assumptions on p, expression (2) would reduce to $p(c)p(h)p(v, s)$, which fully accounts for the objective correlation between v and s. This example demonstrates that the maximal-overlap criterion matters for the evolution of the iterative imputation procedure.

3 Bayesian Networks

I now formally introduce the notion of a "Bayesian-network representation". The exposition is standard (e.g., see Cowell et al. (1999)). Occasionally I use different terminology that is more familiar to economists.

Let (N, R) be a directed graph, where N (the set of variable indices) is the set of nodes and R is the set of directed links. I use the notations jRi and $j \to i$ interchangeably. The graph is *acyclic* if it does not contain any directed path from a node to itself. From now on, I refer to R itself as a directed acyclic graph (DAG). For every $i \in N$, denote $R(i) = \{j \in N \mid jRi\}$. The *skeleton* of R, denoted $\tilde{R}$, is its non-directed version - that is, $i\tilde{R}j$ if iRj or jRi. A subset of nodes $C \subseteq N$ is a *clique* in R if $i\tilde{R}j$ for every $i, j \in C$. A clique is *maximal* if it is not a strict subset of another clique. A clique C is ancestral if $R(i) \subset C$ for every $i \in C$.

Fix a DAG R. For every objective distribution $p \in \Delta(X)$, define

$$p_R(x) \equiv \prod_{i \in N} p(x_i \mid x_{R(i)}) \tag{3}$$

The distribution p_R is said to *factorize* p *according to* R. The DAG R and the set of distributions that can be factorized by R constitute a *Bayesian network*.

For instance, if $R : 1 \to 2 \to 3 \leftarrow 4$, then $p_R(x) = p(x_1)p(x_4)p(x_2 \mid x_1)p(x_3 \mid x_2, x_4)$. In the original vitamin-pills example, the final belief factorizes p according to the DAG $c \to v \to h$. In the extended vitamin-pills example, the final belief factorizes p according to the DAG $c \to v \to s \to h$. The R.H.S of (1) factorizes p according to a DAG that consists of disjoint cliques.

Different DAGs can be equivalent in terms of the distributions they factorize. For example, the DAGs $1 \to 2$ and $2 \to 1$ are equivalent, since $p(x_1)p(x_2 \mid x_1) \equiv p(x_2)p(x_1 \mid x_2)$.

Definition 1 (Equivalent DAGs) *Two DAGs R and Q are* ***equivalent*** *if $p_R = p_Q$ for every $p \in \Delta(X)$.*

For instance, all fully connected DAGs are equivalent: in this case, the factorization formula (3) reduces to a textbook chain rule. Verma and Pearl

(1991) provided a complete characterization of equivalent DAGs, which will be useful in the sequel. Define the *v-structure* of a DAG R to be the set of all triples of nodes i, j, k such that iRk, jRk, $i\not Rj$ and $j\not Ri$.

Proposition 1 (Verma and Pearl (1991)) *Two DAGs R and Q are equivalent if and only if they have the same skeleton and the same v-structure.*

To illustrate this result, the DAGs $1 \rightarrow 3 \leftarrow 2$ and $1 \rightarrow 3 \rightarrow 2$ have identical skeletons but different v-structures. Therefore, these DAGs are not equivalent: there exist distributions that can be factorized by one DAG but not by the other. In contrast, the DAGs $1 \rightarrow 3 \rightarrow 2$ and $1 \leftarrow 3 \leftarrow 2$ are equivalent because they have the same skeleton and the same (vacuous) v-structure.

The following class of DAGs will play an important role in our analysis.

Definition 2 (Perfect DAGs) *A DAG R is **perfect** if $R(i)$ is a clique for every $i \in N$.*

A DAG is perfect if and only if it has a vacuous v-structure: for every triple of variables i, j, k for which jRi and kRi, it is the case that $j\tilde{R}k$. For instance, the DAG $3 \leftarrow 1 \rightarrow 2 \rightarrow 4$ is perfect, whereas the DAG $3 \leftarrow 1 \rightarrow 2 \leftarrow 4$ is imperfect. In the two vitamin-pills examples, the causal chains $c \rightarrow v \rightarrow h$ and $c \rightarrow v \rightarrow s \rightarrow h$ are perfect DAGs, whereas the "true DAG" $c \rightarrow v \leftarrow h$, invoked in the original vitamin-pills example, is imperfect.

In what follows, I refer to p_R as a DAG representation. When R is perfect, I refer to p_R as a perfect-DAG representation. Proposition 1 implies the following result.

Remark 1 *Two perfect DAGs are equivalent if and only if they have the same set of maximal cliques. In particular, we can set any one of these cliques to be ancestral w.l.o.g.*

The DAG representation p_R generally distorts the objective distribution p: unless R is fully connected, there exists an objective distribution p for which $p_R \neq p$. However, certain marginal distributions are not distorted by p_R. The following proposition, which will be useful in the next section, characterizes these cases. The proof is relegated to an appendix.[3]

Proposition 2 *Let R be a DAG and let $C \subseteq N$. Then, $p_R(x_C) \equiv p(x_C)$ for every p if and only if C is an ancestral clique in some DAG in the equivalence class of R.*

Thus, the marginal distribution over X_C induced by p_R never distorts the true marginal if C is an ancestral clique in R, or in some DAG that is equivalent to R. The intuition for this result can be conveyed through the causal interpretations of DAGs, and by considering a set C that consists of a single node i. When i is an ancestral node, it is a "primary cause". The belief distortions that arise from a misspecified DAG concern variables that are either independent of i or (possibly indirect) effects of i. Therefore, these distortions are irrelevant for calculating the marginal distribution of x_i. In contrast, suppose that i cannot be represented as an ancestral node in any DAG that is equivalent to R. Then, there must be two other variables j, k that function as (possibly indirect) causes of i, and the DAG deems j independent of k or i. This failure to account for the full dependencies among i and its multiple causes can lead to distorting the marginal distribution over x_i.

4 Analysis

This section analyzes the connection between DAG representations and the iterative imputation procedure. In particular, it establishes that perfect-DAG

[3]For other properties of DAG representations, see Dawid and Lauritzen (1993).

representations can always be justified as the final output of the iterative imputation procedure. In contrast, imperfect-DAG representations lack this foundation.

Theorem 1 *Suppose that a DAG R is perfect. Then, there exists a missing-data process σ such that for every objective distribution p, applying the iterative imputation procedure to the dataset (p, σ) yields the final output $q^{m-1} = p_R$. In particular, we can select any missing-data process σ for which $\mathcal{S}$ is the set of maximal cliques of R.*

This result provides a constructive "limited feedback" foundation for perfect-DAG representations. Let R be perfect, and consider any missing-data process σ whose support coincides with the set of maximal cliques of R. Then, for any objective distribution p, when we apply the iterative imputation procedure to the dataset (p, σ), we obtain a final output q^{m-1} that coincides with the DAG representation p_R - i.e., q^{m-1} will factorize the objective distribution p according to R.

The following result is a converse to Theorem 1.

Theorem 2 *Suppose that a DAG R is imperfect. Then, for every missing-data process σ there is an objective distribution p, such that applying the iterative imputation procedure to the dataset (p, σ) yields a final output $q^{m-1} \neq p_R$.*

Thus, the iterative imputation procedure does not provide a foundation for imperfect-DAG representations of non-rational expectations. The simplest example of an imperfect-DAG representation is given by the DAG $R : 1 \rightarrow 3 \leftarrow 2$, such that $p_R(x) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$. It is easy to see from this expression why it lacks a limited-feedback foundation. In order to estimate the term $p(x_3 \mid x_1, x_2)$, the DM must have access to joint observations of all three variables. But this would also enable him to correctly estimate whatever correlation exists between x_1 and x_2, whereas p_R treats them as mutually independent.

Example 3.1: Observing variable pairs

The following is the simplest example of a missing-data process for which the iterative imputation procedure fails to output a DAG-representation. Consider a missing-data process σ, in which $\mathcal{S} = \{\{1,2\},\{1,3\},\{2,3\}\}$. Apply the iterative imputation procedure. The maximal-overlap criterion is vacuous in this case, because any ordering of the members of $\mathcal{S}$ satisfies it. Let us select $S^2 = \{2,3\}$ (without loss of generality, for our purposes). Then, $q^1(x_1,x_2,x_3) = p(x_2)p(x_1 \mid x_2)p(x_3 \mid x_1)$, as in the original vitamins example. Turning to round 2, note that $q_1^2 = q^1$, whereas

$$q_2^2(x) = p(x_2,x_3)q^1(x_1 \mid x_2,x_3) = q^1(x)\frac{p(x_2,x_3)}{q^1(x_2,x_3)}$$

Then,

$$q^2(x) = q^1(x)\left[\sigma\{2,3\}\frac{p(x_2,x_3)}{q^1(x_2,x_3)} + (1-\sigma\{2,3\})\right]$$

To see why q^2 has no DAG representation, consider first any objective distribution p that is consistent with $R : 2 \rightarrow 1 \rightarrow 3$. Then, $q^1 = p_R = p$, such that $q^1(x_2,x_3) = p(x_2,x_3)$. It follows that q^2 does not distort such objective distributions, because they are consistent with R. But now consider an objective distribution p for which x_1 is an independent variable whereas x_2 and x_3 are correlated. Such a distribution is inconsistent with R, because R postulates that x_2 and x_3 are independent conditional on x_1. In this case, $q^1(x_2,x_3) = p(x_2)p(x_3) \neq p(x_2,x_3)$, such that $q^2 \neq q^1$, and therefore q^2 does not factorize such distributions p according to R (or any DAG that is equivalent to R). Moreover, for such a distribution p, $q^1(x)$ is reduced to $p(x_1)p(x_2)p(x_3)$, such that

$$q^2(x) = p(x_1)\left[\sigma\{2,3\}p(x_2,x_3) + (1-\sigma\{2,3\})p(x_2)p(x_3)\right]$$

This expression does not factorize p according to *any* DAG, again because $p(x_2)p(x_3) \neq p(x_2,x_3)$.

4.1 Discussion

Theorems 1 and 2 create a distinction between types of correlation distortions that can be understood as the outcome of "extrapolation from limited feedback" and those that cannot - at least modulo the notion of limited feedback and extrapolation method that I have assumed. In this sub-section I discuss the significance of this distinction.

Recall that by Remark 1, any node in a perfect DAG R can be ancestral in some DAG in the equivalence class of R. This means that if iRj, there exists an equivalent DAG R' such that $jR'i$. In other words, perfect DAGs do not postulate identifiable causal links. In this sense, the causal interpretation of perfect DAGs is arbitrary. In contrast, every imperfect DAG R contains at least one identifiable causal link, in the sense that the link remains unreversed in any DAG in the equivalence class of R. Thus, the main results imply that only causal models that make unidentifiable causal assumptions can be extrapolated from limited data (via the iterative imputation procedure). This is consistent with the familiar motto "correlation does not imply causation": the DM's dataset is purely statistical, and contains no information about causation; the extrapolation method that the DM employs does not create meaningful beliefs about causality out of thin air.

The following examples illustrate the distinction that emerges from the main results.

Fixed-lag causal models

Suppose that the DM's subjective belief q distorts the objective distribution p by forcing a "fixed lag" causal structure on it. That is, q treats each of the n variables as a stochastic function of its L immediate predecessors:

$$q(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i \mid x_{\max(1,i-L)}, ..., x_{\max(1,i-1)})$$

For instance, when $L = 1$, $q(x_1, ..., x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1})$, as

in our two vitamin-pills examples.

The "fixed lag" belief factorizes p according to a DAG R defined by $R(i) = \{\max(1, i-L), ..., \max(1, i-1)\}$. This DAG is perfect: if jRi, kRi and $j < k$, then $i - L \leq j, k \leq i - 1$, and therefore j is among the L immediate predecessors of k, such that jRk. Theorem 1 then implies that any missing-data process with $\mathcal{S} = \{\{\max(1, i-L), ..., i\}\}_{i \in N}$ will give rise to the fixed-lag belief, via the iterative imputation procedure. Thus, subjective beliefs that distort p according to a fixed-lag causal model are justified as the outcome of extrapolation from limited feedback.

Correlation neglect
Let $n = 3$, and consider the DAG R

$$\theta \rightarrow z \leftarrow a$$

For instance, θ represents a state of Nature that affects the valuation of an object, a represents the bidding behavior of a player in an auction, and z represents the auction's outcome. Then, $p_R(\theta, a, z) = p(\theta)p(a)p(z \mid \theta, a)$. The DAG R regards the variables θ and a as independent. In terms of the auction example, p_R posits that the player does not condition his bid on the object's valuation. Therefore, there exist objective distributions p for which θ and a are correlated, such that $p_R \neq p$. In this case, we may say that the subjective belief given by p_R exhibits "correlation neglect" (of a sort studied by Eyster and Rabin (2005), Jehiel and Koessler (2008) and others).

Is it possible for an observer of partial data about $p(\theta, a, z)$ to "infer" the (potentially false) causal structure given by R? As we saw in the discussion that immediately followed Theorem 2, the answer is negative, as R is imperfect. There is no missing-data process that will generate a final belief that factorizes every p according to R. Thus, the form of correlation neglect captured by p_R cannot be justified by our model of extrapolation from limited feedback.

This conclusion is crucially sensitive to the specification of the observable variables. Suppose that the variable z is not observable. Correspondingly, remove the node that represents z from R. Then, the DAG representation is reduced to $p_R(\theta, a) = p(\theta)p(a)$, which obviously has a limited-feedback justification, in the form of any missing-data process in which the DM has access to observations of θ alone and observations of a alone. The lesson is that whether a certain type of belief distortion has a limited-feedback foundation crucially depends on the precise specification of the observable variables.

Price taking vs. quantity competition
Let $N = \{1, 2, 3\}$. Suppose that x_1 represents the price of a product, while x_2 and x_3 represent the quantities produced by two sellers. Consider two alternative DAGs:

$$\begin{aligned} R^t &: 2 \leftarrow 1 \rightarrow 3 \\ R^c &: 2 \rightarrow 1 \leftarrow 3 \end{aligned}$$

The DAG R^t is a "price taking" causal model: the price x_1 fluctuates according to some exogenous forces outside the model; and each seller reacts independently to the price. The DAG R^c is a "quantity competition" causal model: the two sellers choose their production quantities independently, and the price is determined as a consequence of their quantity choices.

Can an outside observer "infer" any of these models from limited data? The DAG R^t is perfect. Therefore, by Theorem 1, p_{R^t} has a "limited feedback foundation". Any missing-data process with $\mathcal{S} = \{\{1, 2\}, \{1, 3\}\}$ will generate p_{R^t} as the outcome of a single-round imputation, as in the original vitamins example. Note that in the present context, the limited data has a simple interpretation: the outside observer measures each seller's "supply function", namely the correlation between his production quantity and the product price. These "supply functions" are extrapolated into a joint dis-

tribution that is consistent with the "price taking" model. In contrast, the DAG R^c is imperfect. By Theorem 2, p_{R^c} lacks a "limited feedback foundation": there exists no missing-data process that would generate a final output that coincides with p_{R^c} for all objective distributions p. In this sense, the "quantity competition" model cannot be extrapolated from limited feedback.

Game forms

Let $N = \{1, 2, 3, 4\}$. Each variable x_i represents the action of a different player in an extensive game. Each of the following two DAGs represents a game form in which each player moves once and the order of moves is fixed. Moreover, a link $i \to j$ means that j always observes i's move. In other words, DAGs represent what Eyster and Rabin (2014) refer to as "observability structures".

The question is whether an outside observer would "infer" these game forms from some partial feedback regarding the joint distribution over players' actions. The left-hand DAG is imperfect. Therefore, there is no missing-values process that could generate a final belief that factorizes every objective distribution according to this graph ($R(4) = (2, 3)$, and yet 2 are 3 are not linked). In this sense, the game form described by the left-hand DAG cannot be "inferred" from limited feedback. In contrast, the right-hand DAG is perfect, and therefore "inferable" from limited feedback, in the sense that any missing-values process with $\mathcal{S} = \{\{1, 2, 3\}, \{1, 4\}\}$ will generate a final belief that can be written as $q^2 = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1)$.

4.2 Proofs

The basic idea behind the proofs of the main results is as follows. I first introduce a property of collection of sets, called RIP*, and observe that the set of maximal cliques of a perfect DAG satisfies this property. I then use a proof by induction on m to show that for any missing-data process for which S satisfies RIP*, the final output of the iterative imputation procedure q^{m-1} has a perfect-DAG representation - i.e., it factorizes any objective distribution according to a perfect DAG. Moreover, the set of maximal cliques of the perfect DAG is S (which means that the DAG is essentially unique). Finally, I show that if S violates RIP*, q^{m-1} lacks a DAG representation - i.e., there is no DAG R such that $q^{m-1} = p_R$ for all objective distributions p. The two main results follow from this reasoning. If R is perfect, we know what $\mathcal{S}$ needs to be in order to give p_R a limited-feedback foundation. In contrast, if R is imperfect, it is not equivalent to any perfect DAG; and since the iterative imputation can only produce a perfect-DAG representation, this means that no missing-data process can give p_R a limited-feedback foundation.

Definition 3 *A sequence of sets $S_1, ..., S_m$ satisfies the **running intersection property (RIP)** if for every $k = 2, ..., m$, $S_k \cap (\cup_{i<k} S_i) \subseteq S_j$ for some $j < k$. The collection $\mathcal{S}$ satisfies **RIP*** if its elements can be ordered in a sequence that satisfies RIP.*

RIP is a familiar concept in the Bayesian-network literature (see Cowell et al. (1999, p. 54)). A sequence of sets satisfies RIP if the intersection between any set along the sequence and the union of its predecessors is weakly contained in one of these predecessors. We will be interested in whether $\mathcal{S}$ (the support of the missing-values process σ) satisfies RIP*. The property holds trivially for $m = 2$. To illustrate the definition, note that the sequence $\{1,3\}, \{2,4\}, \{1,2\}$ violates RIP because $\{1,2\} \cap (\{1,3\} \cup \{2,4\}) = \{1,2\}$, which is not contained in any of the first two

sets in the sequence. In contrast, the sequence $\{1,3\},\{1,2\},\{2,4\}$ satisfies RIP because $\{2,4\} \cap (\{1,3\} \cup \{1,2\}) = \{2\}$, which is contained in the second set in the sequence. Therefore, the collection $\mathcal{S} = \{\{1,3\},\{1,2\},\{2,4\}\}$ satisfies RIP*. In contrast, the collection $\mathcal{S} = \{\{1,3\},\{1,2\},\{2,3\}\}$ violates RIP*, because there is no way to order its members in a way that satisfies RIP.

Remark 2 (Cowell et al. (1999, p. 54)) *The set of maximal cliques in any perfect DAG satisfies RIP*.*

Proof of Theorem 1

The following lemma due to Noga Alon will be instrumental in the proof. A sequence of sets $S^0, S^1, ..., S^K$ is **expansive** if for every $k \geq 1$, $\left|S^k \cap (\cup_{j<k} S^j)\right| \geq |S^i \cap (\cup_{j<k} S^j)|$ for all $i > k$.

Lemma (Alon (2014)). Suppose that $\mathcal{S}$ satisfies RIP*. Then, every expansive ordering of $\mathcal{S}$ satisfies RIP.

I will now show that if $\mathcal{S}$ satisfies RIP*, the iterative imputation procedure generates a perfect-DAG representation. Suppose that $\mathcal{S}$ satisfies RIP*. Consider the sequence of sets S^k that are introduced in Step 1 of each round k. By Step 1 of the procedure, the sequence $B^0, S^1, ..., S^{m-1}$ is expansive. By the lemma, it satisfies RIP. My task is to show that for every p and every round $k = 1, ..., m-1$, the belief $q^k \in \Delta(X_{B^k})$ has a perfect-DAG representation, where the DAG R^k is defined over B^k, and its set of maximal cliques is $\{B^0, S^1, ..., S^k\}$. The proof is by induction on k.

Let $k = 1$. Recall that $B^1 = B^0 \cup S^1$. By assumption, $\mathcal{S}$ does not include sets that contain one another. Therefore, $S^1 - B^0$ and $B^0 - S^1$ are non-empty. The auxiliary beliefs q_1^1 and q_2^1 defined over X_{B^1} are given by

$$\begin{aligned} q_1^1(x_{B^1}) &= q(x_{B^0}) p(x_{S^1 - B^0} \mid x_{S^1 \cap B^0}) \\ p_2^1(x_{B^1}) &= p(x_{S^1}) q(x_{B^0 - S^1} \mid x_{S^1 \cap B^0}) \end{aligned}$$

By the basic rules of conditional probability, we have

$$q(x_{B^0})p(x_{S^1-B^0} \mid x_{S^1\cap B^0}) = p(x_{S^1})q(x_{B^0-S^1} \mid x_{S^1\cap B^0})$$

and therefore q^1 is consistent with a DAG R^1 defined over B^1, where $i\tilde{R}^1 j$ if and only if $i,j \in B^0$ or $i,j \in S^1$; and $j\not{R}^1 i$ for every $i \in B^0, j \in S^1$. The DAG R^1 has two maximal cliques, B^0 and S^1.

Consider the initial condition (B^{k-1}, q^{k-1}) of any round $k > 1$. The auxiliary beliefs q_1^k and q_2^k over $B^k = B^{k-1} \cup S^k$ are given by

$$\begin{aligned} q_1^k(x_{B^k}) &= q^{k-1}(x_{B^{k-1}})p(x_{S^k-B^{k-1}} \mid x_{S^k\cap B^{k-1}}) \\ q_2^k(x_{B^k}) &= p(x_{S^k})q^{k-1}(x_{B^{k-1}-S^k} \mid x_{S^k\cap B^{k-1}}) \end{aligned} \tag{4}$$

Consider the expression for q_1^k. The inductive hypothesis is that q^{k-1} has a perfect-DAG representation, where the DAG R^{k-1} is defined over B^{k-1}, and its set of maximal cliques is $\{B^0, S^1, ..., S^{k-1}\}$. By RIP, $S^k \cap B^{k-1}$ is weakly contained in one of the sets $B^0, S^1, ..., S^{k-1}$. Extend R^{k-1} to a DAG R^k over B^k, simply by adding a directed link between every $i,j \in S^k$ (without destroying acyclicity). The DAG R^k is perfect, and its set of maximal cliques is $\{B^0, S^1, ..., S^k\}$. Thus, q_1^k is a perfect-DAG representation, where the DAG is R^k.

It remains to show that q_2^k coincides with q_1^k. Note that q_1^k and q_2^k can be written as

$$\begin{aligned} q_1^k(x_{B^k}) &= p(x_{S^k})q^{k-1}(x_{B^{k-1}}) \cdot \frac{1}{p(x_{S^k\cap B^{k-1}})} \\ q_2^k(x_{B^k}) &= p(x_{S^k})q^{k-1}(x_{B^{k-1}}) \cdot \frac{1}{q^{k-1}(x_{S^k\cap B^{k-1}})} \end{aligned} \tag{5}$$

Since q^{k-1} is a perfect-DAG representation, where the DAG is R^{k-1}, and $S^k \cap B^{k-1}$ is a clique in R^{k-1}, Remark 1 implies that w.l.o.g it is an ancestral clique. Proposition 2 then implies that $q^{k-1}(x_{S^k\cap B^{k-1}}) \equiv p(x_{S^k\cap B^{k-1}})$.

Therefore, q_2^k coincides with q_1^k.

We have thus established that if $\mathcal{S}$ satisfies RIP*, the iterative imputation procedure generates a final output q^{m-1} with a perfect-DAG representation, where the set of maximal cliques of the DAG is $\mathcal{S}$. This means that if we are given a perfect-DAG representation p_R, then for any missing-data process for which $\mathcal{S}$ is the set of maximal cliques of R, applying the iterative imputation procedure to (p, σ) will yield a final output $q^{m-1} = p_R$ for any objective distribution p. This completes the proof.

Proof of Theorem 2

I will show that if $\mathcal{S}$ violates RIP*, the iterative imputation fails to generate a DAG representation. Suppose that $\mathcal{S}$ violates RIP*. Then, any ordering of its sets will violate RIP. Note that this means $m \geq 3$. Recall that the sequence B^0, S^1 trivially satisfies RIP. Let $k > 1$ be the earliest round for which $S^k \cap B^{k-1}$ is *not* weakly contained in any of the sets $B^0, S^1, ..., S^{k-1}$. By the proof of Proposition **??**, $q^{k-1} \in \Delta(X_{B^{k-1}})$ is a perfect-DAG representation, where the perfect DAG R^{k-1}, defined over B^{k-1}, is characterized by the set of maximal cliques $\{B^0, S^1, ..., S^{k-1}\}$. It follows that $S^k \cap B^{k-1}$ is not a clique in R^{k-1}. By Proposition 2, there exist distributions p for which $q^{k-1}(x_{S^k \cap B^{k-1}}) \neq p(x_{S^k \cap B^{k-1}})$. Therefore, by (5), there exists p for which q_1^k and q_2^k do not coincide. I now construct a family of such distributions. Since $S^k \cap B^{k-1}$ is not a clique in R^{k-1}, it must contain two nodes, denoted w.l.o.g 1 and 2, that are not linked by R^{k-1}. Let p be an arbitrary objective distribution for which x_i is independently distributed for every $i > 2$, whereas x_1 and x_2 are mutually correlated. Then,

$$q^{k-1}(x_{B^{k-1}}) = \prod_{i \in B^{k-1}} p(x_i)$$

It follows that q_1^k and q_2^k can be written as

$$\begin{aligned} q_1^k(x_{B^k}) &= p(x_1)p(x_2) \cdot \prod_{i \in B^k - \{1,2\}} p(x_i) \\ q_2^k(x_{B^k}) &= p(x_1)p(x_2 \mid x_1) \cdot \prod_{i \in B^k - \{1,2\}} p(x_i) \end{aligned}$$

such that

$$q^k(x_{B^k}) = \left(\prod_{i \in B^k - \{2\}} p(x_i) \right) \cdot \left(\frac{\sum_{i \leq k-1} \sigma(S^i)}{\sum_{i \leq k} \sigma(S^i)} p(x_2) + \frac{\sigma(S^k)}{\sum_{i \leq k} \sigma(S^i)} p(x_2 \mid x_1) \right)$$

Since all the variables $i \in N - \{1, 2\}$ are independently distributed under p, the continuation of the iterative imputation procedure will eventually produce a final belief of the form

$$q^{m-1}(x) = \left(\prod_{i \neq 2} p(x_i) \right) \cdot [\alpha p(x_2) + (1 - \alpha) p(x_2 \mid x_1)]$$

where $\alpha \in (0, 1)$. Since $p(x_2 \mid x_1) \neq p(x_2)$ for some x_1, x_2, q^{m-1} does not have a DAG representation.

We have thus proved that if $\mathcal{S}$ violates RIP*, q^{m-1} lacks a DAG representation. And in the proof of Theorem 1, we saw that if $\mathcal{S}$ satisfies RIP*, q^{m-1} has a *perfect*-DAG representation, hence it cannot have an imperfect-DAG representation. By Remark 1, if R is imperfect, it is not equivalent to any perfect DAG. It follows that there is no missing-data process σ such that applying the iterative imputation procedure to (p, σ) will generate a final output with an imperfect-DAG representation.

4.3 Restricting the Domain of Objective Distributions

The exercise we conducted in this section imposed no restriction on the domain of objective distributions. However, in many applications of interest, we have some a-priori reason to rule out certain classes of objective distributions. For example, if we know that the context is a Bayesian game, then it is plausible to impose the restriction that players' actions are independent conditional on their signals. The question is, how does our exercise change when the domain of objective distributions, denoted P, is some strict subset of $\Delta(X)$? In this sub-section I partially explore this question, by focusing on a single type of domain restriction, which deems some variables to be independent.

Consider the extreme case in which P consists of all distributions for which all variables are independent - i.e., $p(x_1, ..., x_n) = p(x_1) \cdots p(x_n)$ for every $p \in P$. In this case, the final output of the iterative imputation procedure will satisfy $q^{m-1} = p$. Therefore, q^{m-1} will be consistent with *any* DAG representation, including the empty DAG. The reason is that the belief distortion that the procedure generates is that it falsely extrapolates observed correlations. However, if the objective distribution exhibits no correlations, there is nothing the procedure can distort.

Let $M \subseteq N$, and define $P(M)$ as the set of all distributions for which x_i is an independent variable for all $i \notin M$. For any DAG R and any non-empty $C \subseteq N$, let R^C denote the DAG's projection on C - that is, R^C is defined over the set of nodes C, such that for every $i, j \in C$, iR^Cj if and only if iRj.

Proposition 3 *Let R be a DAG, and suppose that R^C is perfect for some non-empty $C \subseteq N$. Then, there is a missing-data process σ such that for every $p \in P(N-C)$, applying the iterative imputation procedure to the dataset (p, σ) generates a final output $q^{m-1} = p_R$.*

Proof. For every $p \in P(N - C)$, we can write

$$p_R(x) = p_{R^C}(x_C) \cdot \prod_{i \notin C} p(x_i)$$

The R.H.S of this equation factorizes p according to a DAG R' over N that consists of R^C as well as a collection of isolated nodes $N - C$. Clearly, R' is perfect. By Theorem 1, This DAG over N has a limited-feedback foundation: there exists a missing-data process σ, such that for every $p \in \Delta(X)$, applying the iterative imputation procedure to (p, σ) will generate a final belief $q^{m-1} = p_{R'}$. In particular, we can select any missing-data process for which $\mathcal{S}$ consists of all maximal cliques of R^C, as well as all singletons $\{i\}$, $i \notin C$. Since $p_{R'}$ and p_R coincide over $P(N - C)$, this concludes the proof. ■

This result is a simple extension of Theorem 1. It basically says that even if R is imperfect, it will have a limited-feedback foundation as long as we restrict the variables that cause imperfections to be independent across the domain of objective distributions. For example, when $R : 1 \rightarrow 3 \leftarrow 2$, p_R has a limited-feedback foundation if impose the domain restriction that x_1 or x_2 are mutually independent.

5 Relation to the MaxEnt Problem

The iterative imputation procedure is a "behaviorally motivated" procedure for extending a dataset with missing values into a fully-specified probability distribution over X. We could consider other, more "normatively motivated" extrapolation methods. One such criterion is *maximal entropy*. Suppose the DM faces the dataset (p, σ) and derives from it the marginal distributions of p over X_S, $S \in \mathcal{S}$. The DM's problem is to find a probability distribution $q \in \Delta(X)$ that maximizes entropy subject to the constraint that $q(x_S) \equiv p(x_S)$ for every $S \in \mathcal{S}$. A more general version of this problem was originally stated by Jaynes (1957) and has been studied in the Machine Learning literature,

where it is known as the MaxEnt problem.

The maximal-entropy criterion generalizes the "principle of insufficient reason" (recall that unconstrained entropy maximization yields the uniform distribution). The idea behind it is that the DM wishes to be "maximally agnostic" about the aspects of the distributions he has not learned, while being entirely consistent with the aspects he has learned. For instance, suppose that the DM only manages to learn the marginal distributions over all individual variables - i.e., $\mathcal{S} = \{\{1\}, ..., \{n\}\}$. Then, the maximal-entropy extension of these marginals is $p(x_1) \cdots p(x_n)$.

The following is an existing result, reformulated to suit our present purposes.

Proposition 4 ((Hajek et al. (1992))) *Suppose that R is perfect. Then, p_R is a maximal-extension entropy of the marginals of $(p_S)_{S \in \mathcal{S}}$, where $\mathcal{S}$ is the set of maximal cliques of R.*

This result establishes a connection between the iterative imputation procedure and the MaxEnt problem: the former can be viewed as an algorithm for implementing the latter, whenever $\mathcal{S}$ satisfies RIP*. For a more recent study of information-theoretic methods of extrapolation from limited data, see Miller and Liu (2002).

6 A Mixed-DAG Representation

The following is a natural extension of DAG representations. Let λ be a probability distribution over all DAGs R defined over N. Define

$$p_\lambda = \sum_R \lambda(R) p_R$$

I refer to p_λ as a "mixed DAG representation". This extended representation can capture a DM who is uncertain about causal relations among the relevant

variables, and therefore "mixes" between various causal models when trying to fit objective data. In this section I discuss the possibility of giving mixed DAG representations a limited-feedback foundation, in the context of the concrete example of "*partial cursedness*" (Eyster and Rabin (2005)).

Consider the following three-variable model: x_1 and x_2 represent the signals that the DM and another agent (referred to as the DM's opponent) receive, and x_3 represents the opponent's action. Consider the following two DAGs:

$$\begin{aligned} R &: 1 \to 2 \to 3 \\ Q &: 3 \leftarrow 1 \to 2 \end{aligned}$$

Note that R and Q are not equivalent. The DAGs R and Q differ in their causal attribution of the opponent's action: R regards the agent's information as the cause of the agent's action, whereas Q regards the DM's own information as the cause of the agent's action. In other words, Q projects the DM's own information on the agent.

If p results from a conventional economic model in which the DM and the agent move simultaneously, then R is obviously the correct model, whereas Q captures an "attribution error" that can be interpreted as a "curse of knowledge". In this case, the DAG representation p_Q captures a "fully cursed" DM, who fails to imagine that the opponent need not share the DM's own information.

Now consider a mixed-DAG representation given by $\lambda(Q) = \chi$ and $\lambda(R) = 1 - \chi$, such that

$$p_\lambda(x) = p(x_1, x_2)\left[\chi p(x_3 \mid x_1) + (1-\chi)p(x_3 \mid x_2)\right]$$

This representation matches Eyster and Rabin's (2005) notion of a χ-cursed DM. The parameter χ captures the magnitude of the DM's tendency to attribute the opponent's action to the DM's own information. The question is:

does this mixed-DAG representation possess a limited-feedback foundation?

It should be clear that any missing-data process that could possibly generate p_λ must have a support $\mathcal{S} = \{\{1,2\},\{1,3\},\{2,3\}\}$ (any other specification of $\mathcal{S}$ would satisfy RIP* and therefore generate a "pure" DAG representation). In Example 3.1 of Section 4, we saw that when x_1 is an independent variable according to the objective distribution p, applying the iterative imputation procedure to such a missing-data process (with $S^2 = \{2,3\}$) produces a final belief which can be written as

$$q^2(x) = p(x_1, x_2)\left[(1 - \sigma\{2,3\})p(x_3) + \sigma\{2,3\}p(x_3 \mid x_2)\right]$$

Note that in this case, q^2 is consistent with p_λ, with $\chi = 1-\sigma\{2,3\}$. However, for distributions in which x_1 is not an independent variable, this equivalence no longer holds. It can also be shown that if we change the order of the iterative imputation procedure such that $S^2 \neq \{2,3\}$, the procedure's final output is inconsistent with the partially cursed representation.

Thus, if the "partial cursedness" representation is applied to the restricted domain of objective distributions for which the DM's signal is an independent variable, it has a limited-feedback foundation (at least for a particular sequencing of B^0, S^1, S^2). However, this foundation breaks down for the unrestricted domain that allows signals to be correlated.

7 Conclusion

This paper presented an example of a broad research program: establishing a "limited learning feedback" foundation for representations of "boundedly rational expectations". The basic idea is that DMs extrapolate their belief from some feedback they receive about a prevailing probability distribution. The form of their belief - and the systematic biases it may be exhibit in relation to the true distribution - will reflect the structure of their feedback and the method of extrapolation they employ. In this paper, limited feed-

back took the form of an infinitely large sample subjected to an independent missing-values process; the extrapolation method was defined by the iterative imputation procedure. And we saw that when subjective beliefs have a perfect-DAG representation, they can be justified by this notion of extrapolation from limited feedback.

This view of our exercise suggests natural directions for extension. First, it may be interesting to explore further the mixed-DAG representation, as well as other extensions of the basic DAG representation (e.g., allowing for non-directed graphs). Second, the link to the literature on statistical inference with missing data can be developed: allowing p and σ to be correlated, finite samples, datasets that contain both passive observations and active experimentation, coarse or aggregate data, etc.

The link between learning feedback and non-rational expectations was studied in two recent papers. Esponda and Pouzo (2015) proposed a general game-theoretic model, in which each player has a "subjective model", which is a set of stochastic mappings from his action a to a primitive set of payoff-relevant consequences y he observes during his learning process. This learning feedback is limited because in the true model, other "latent" variables may affect the action-consequence mapping. Esponda and Pouzo define an equilibrium concept, in which each player best-replies to a subjective distribution (of y conditional on a), which is the closest in his subjective model to the true equilibrium distribution. Distance is measured by a weighted version of Kullback-Leibler divergence. Esponda and Pouzo justify this equilibrium concept as the steady-state of a Bayesian learning process.

Schwartzstein (2014) studied a dynamic model in which a DM tries to predict a variable y as a function of two variables x and z. At every period, he observes the realizations of y and x. In contrast, he pays attention to the realization of z only if his belief at the beginning of the period is that z is sufficiently predictive of y. When the DM chooses not to observe z, he imputes a constant value. Schwartzstein examines the long-run belief that

emerges from this learning process, and in particular the DM's failure to perceive correlations among the three variables.

References

[1] Alon, N. (2014), "Problems and Results in Extermal Combinatorics - III," Tel Aviv University, mimeo.

[2] Cowell, R., P. Dawid, S. Lauritzen and D. Spiegelhalter (1999), *Probabilistic Networks and Expert Systems,* Springer, London.

[3] Dawid, P. and S. Lauritzen (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *Annals of Statistics*, 21, 1272–1317.

[4] Esponda, I. (2008), "Behavioral Equilibrium in Economies with Adverse Selection," *The American Economic Review,* 98, 1269-1291.

[5] Esponda, I. and D. Pouzo (2014), "Conditional Retrospective Voting in Large Elections," mimeo.

[6] Esponda, I. and D. Pouzo (2015), "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models," mimeo.

[7] Ettinger, D. and P. Jehiel (2010), "A Theory of Deception," *American Economic Journal: Microeconomics* 2, 1-20.

[8] Eyster, E. and M. Piccione (2013), "An Approach to Asset Pricing Under Incomplete and Diverse Perceptions," *Econometrica,* 81, 1483-1506.

[9] Eyster, E. and M. Rabin (2005), "Cursed Equilibrium," *Econometrica,* 73, 1623-1672.

[10] Eyster, E. and M. Rabin (2014), "Extensive imitation is irrational and harmful," *Quarterly Journal of Economics*, 129, 1861-1898.

[11] Gelman, A. and J. Hill (2006), *Data Analysis using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

[12] Gilboa, I. (2014), "Rationality and the Bayesian Paradigm," mimeo.

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

[14] Hajek, P., T. Havranek and R. Jirousek (1992), *Uncertain Information Processing in Expert Systems*, CRC Press.

[15] Jaynes, E. T. (1957), "Information Theory and Statistical Mechanics," *Physical Review,* 106, 620-630.

[16] Jehiel, P. (2005), "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory,* 123, 81-104.

[17] Jehiel, P. and F. Koessler (2008), "Revisiting Games of Incomplete Information with Analogy-Based Expectations," *Games and Economic Behavior,* 62, 533-557.

[18] Little, R. and D. Rubin (2002), *Statistical Analysis with Missing Data,* Wiley, New Jersey.

[19] Miller, D. and W. Liu (2002), "On the Recovery of Joint Distributions from Limited Information," *Journal of Econometrics* 107, 259-274.

[20] Mullainathan, S. J. Schwartzstein and A. Shleifer (2008), "Coarse Thinking and Persuasion," *Quarterly Journal of Economics* 123, 577-619.

[21] Osborne, M. and A. Rubinstein (1998), "Games with Procedurally Rational Players," *American Economic Review,* 88, 834-849.

[22] Pearl, J. (2009), *Causality: Models, Reasoning and Inference,* Cambridge University Press, Cambridge.

[23] Piccione, M. and A. Rubinstein (2003), "Modeling the Economic Interaction of Agents with Diverse Abilities to Recognize Equilibrium Patterns," *Journal of the European Economic Association*, 1, 212-223.

[24] Schwartzstein, J. (2014), "Selective Attention and Learning," *Journal of European Economic Association,* 12, 1423-1452.

[25] Sloman, S. (2009), Causal Models: How People Think about the World and its Alternatives, Oxford University Press.

[26] Spiegler, R. (2015), "Bayesian Networks and Boundedly Rational Expectations," mimeo.

[27] Verma, T. and J. Pearl (1991), "Equivalence and Synthesis of Causal Models," *Uncertainty in Artificial Intelligence,* 6, 255-268.

[28] Woodford, M. (2013), "Macroeconomic Analysis without the Rational Expectations Hypothesis," *Annual Review of Economics*, forthcoming.

Appendix: Proof of Proposition 2

For convenience, label the variables in C by $1, ..., m$. Let us write down the explicit expression for $p_R(x_C)$:

$$\begin{aligned} p_R(x_C) &= \sum_{x'_{m+1},...,x'_n} p_R(x_1, ..., x_m, x'_{m+1}, ..., x'_n) \qquad (6) \\ &= \sum_{x'_{m+1},...,x'_n} \prod_{i \in C} p(x_i \mid x_{R(i) \cap C}, x'_{R(i)-C}) \prod_{i \notin C} p(x'_i \mid x_{R(i) \cap C}, x'_{R(i)-C}) \end{aligned}$$

(i). Assume C is an ancestral clique in R. Then,

$$\prod_{i\in C} p(x_i \mid x_{R(i)\cap C}, x'_{R(i)-C}) = p(x_1)\prod_{i=2}^{m} p(x_i \mid x_1, ..., x_{m-1}) = p(x_C)$$

Expression (6) can thus be written as

$$p(x_C)\sum_{x'_{m+1},...,x'_n}\left(\prod_{i=m+1}^{n} p(x'_i \mid x_{R(i)\cap C}, x'_{R(i)-C}\right) = p(x_C)$$

Therefore, $p_{R'}(x_C) \equiv p(x_C)$ for every R' that is equivalent to R.
(ii). Let us distinguish between two cases.

Case 1: C is not a clique in some DAG R'. Then, C contains two variables, labeled w.l.o.g 1 and 2, such that $1 \not R 2$ and $2 \not R 1$ for every R in the equivalence class of R'. Consider an objective distribution p, for which every x_i, $i > 2$, is distributed independently, whereas x_1 and x_2 are mutually correlated. Then, expression (6) is simplified into

$$\prod_{i=1}^{m} p(x_i)\sum_{x'_{m+1},...,x'_n}\prod_{i=m+1}^{n} p(x'_i) = \prod_{i=1}^{m} p(x_i)$$

whereas the objective distribution can be written as

$$p(x_C) = p(x_1)p(x_2 \mid x_1)\prod_{i=3}^{m} p(x_i)$$

The two expressions are different because x_2 and x_1 are not independent.

Case 2: C is a clique which is not ancestral in any DAG in the equivalence class of R. Suppose that for every node $j \in C$, j has no "unmarried parents" - i.e., if there exist nodes k, k' such that kRj and $k'Rj$, then kRk' or $k'Rk$. In addition, if there is a directed path from some $i \notin C$ to j, then i has no unmarried parents either. Transform R into another DAG R' by inverting

every link along every such path. The DAGs R and R' share the same skeleton and v-structure. By Proposition 1, they are equivalent. By construction, C is an ancestral clique in DAG that is equivalent to R, a contradiction.

It follows that R has the following structure. First, there exist three distinct nodes, denoted w.l.o.g $1, 2, 3$, such that $1, 2 \notin C$, $1R3$, $2R3$, $1\not R2$ and $2\not R1$. Second, there is a directed path from 3 to some node $s \in C$, $s \geq 3$. For convenience, denote the path by $(3, 4, ..., s)$ - i.e., the immediate predecessor of any $j > 3$ along the path is $j - 1$. It is possible that $s = 3$, in which case the path is degenerate. W.l.o.g, we can assume that $i \notin C$ for every $i \neq s$ along this path (otherwise, we can take s to be the lowest-numbered node that belongs to C along the path).

Consider any p which is consistent with a DAG R^* that has the following structure: first, $1R^*2R^*3$ and $1R^*3$; second, for every $j \in \{4, ..., s\}$, $R^*(j) = \{j-1\}$; and $R^*(j) = \varnothing$ for every $j \notin \{2, ..., s\}$. (Note that the latter property means that every x_j, $j \notin \{2, ..., s\}$ is independently distributed. Then,

$$p(x) = p_{R^*}(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdot \prod_{i=4}^{s} p(x_i \mid x_{i-1}) \cdot \prod_{j>4} p(x_j)$$

In contrast,

$$p_R(x) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2) \cdot \prod_{i=4}^{s} p(x_i \mid x_{i-1}) \cdot \prod_{j>4} p(x_j)$$

By definition, every $i = 4, ..., s-1$ does not belong to C. Denote

$$D(x') = p(x'_1)p(x'_3 \mid x'_1, x'_2) \left(\prod_{i=4}^{s-1} p(x'_i \mid x'_{i-1}) \right) p(x_s \mid x'_{s-1})$$

Therefore,

$$\begin{aligned} p(x_C) &= \prod_{j \in C-\{s\}} p(x_j) \sum_{x'} p(x'_2 \mid x'_1) D(x') \\ p_R(x_C) &= \prod_{j \in C-\{s\}} p(x_j) \sum_{x'_1,\ldots,x'_{s-1}} p(x'_2) D(x') \end{aligned}$$

It is easy to see from these expressions that we can find a distribution p which is consistent with R^* such that $p_R(x_C) \neq p(x_C)$ for some x.