Journal of Machine Learning Research 10 (2009) 1759-1799 Submitted 10/07; Revised 3/09; Published 8/09

Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning

Halbert White HWHITE@UCSD.EDU
Department of Economics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093, USA

Karim Chalak CHALAK@BC.EDU
Department of Economics
Boston College
140 Commonwealth Avenue
Chestnut Hill, MA 02467, USA

Editor: Michael Jordan

Abstract

Judea Pearl's Causal Model is a rich framework that provides deep insight into the nature of causal relations. As yet, however, the Pearl Causal Model (PCM) has had a lesser impact on economics or econometrics than on other disciplines. This may be due in part to the fact that the PCM is not as well suited to analyzing structures that exhibit features of central interest to economists and econometricians: optimization, equilibrium, and learning. We offer the settable systems framework as an extension of the PCM that permits causal discourse in systems embodying optimization, equilibrium, and learning. Because these are common features of physical, natural, or social systems, our framework may prove generally useful for machine learning. Important features distinguishing the settable system framework from the PCM are its countable dimensionality and the use of partitioning and partition-specific response functions to accommodate the behavior of optimizing and interacting agents and to eliminate the requirement of a unique fixed point for the system. Refinements of the PCM include the settable systems treatment of attributes, the causal role of exogenous variables, and the dual role of variables as causes and responses. A series of closely related machine learning examples and examples from game theory and machine learning with feedback demonstrates some limitations of the PCM and motivates the distinguishing features of settable systems.

Keywords: causal models, game theory, machine learning, recursive estimation, simultaneous equations

1. Introduction

Judea Pearl's work on causality, especially as embodied in his landmark book *Causality* (Pearl, 2000), represents a rich framework in which to understand, analyze, and explain causal relations. This framework has been adopted and applied in a broad array of disciplines, but so far it has had a lesser impact in economics. This may be due in part to the fact that the Pearl causal model (PCM) is not as explicit about or well suited to analyzing structures that exhibit features of central interest to economists and econometricians: optimization, equilibrium, and learning. Here, we offer the

settable systems framework as an extension of the PCM that permits causal discourse in systems embodying these features.

Because optimization, equilibrium, and learning are features not only of economic systems, but also of physical, natural, or social systems more generally, our extended framework may prove useful elsewhere, especially in areas where empirical analysis, whether observational or experimental, has a central role to play. In particular, settable systems offer a number of advantages relative to the PCM for machine learning. To show this, we provide a detailed examination of the features and limitations of the PCM relevant to machine learning. This examination provides key insight into the PCM and helps to motivate features of the settable systems framework we propose.

Roughly speaking, a settable system is a mathematical framework describing an environment in which multiple agents interact under uncertainty. In particular, the settable systems framework is explicit about the principles underlying how agents make decisions, the equilibria (if any) resulting from agents' decisions, and learning from repeated interactions. Because it is explicit about agents' decision making, the settable systems framework extends the PCM by providing a decision-theoretic foundation for causal analysis (see, e.g., Heckerman and Shachter, 1995) in the spirit of influence diagrams (Howard and Matheson, 1984). However, unlike influence diagrams, the settable systems framework preserves the spirit of the PCM and its appealing features for empirical analysis, including its use of response functions and the causal notions that these support.

As Koller and Milch (2003, pp. 189-190) note in motivating their study of multi-agent influence diagrams (MAIDs), "influence diagrams [...] have been investigated almost entirely in a single-agent setting." The settable systems framework also permits the study of multiple agent interactions. Nevertheless, a number of settable systems features distinguishes them from MAIDs, as we discuss in Section 6.4. Among other things, settable systems permit causal discourse in systems with multi-agent interactions.

Some features of settable systems are entirely unavailable in the PCM. These include (1) accommodating an infinite number of agents; and (2) the absence of a unique fixed point requirement. Other features of settable systems rigorously formalize and refine or extend related PCM features, thereby permitting a more explicit causal discourse. These features include (3) the notion of attributes, (4) definitions of interventions and direct effects, (5) the dual role of variables as causes and responses, and (6) the causal role of exogenous variables.

For instance, for a given system, the PCM's common treatment of attributes and background variables rules out a causal role for background variables. Specifically, this rules out structurally exogenous causes, whether observed or unobserved. This also limits the role of attributes in characterizing systems of interest. Because the status of a variable in the PCM is relative to the analysis and is entirely up to the researcher, a background variable may be treated as an endogenous variable in an alternative system if deemed sensible by the researcher, thereby permitting it to have a causal role. Nevertheless, the PCM is silent about how to distinguish between attributes, background variables, and endogenous variables. In contrast, in settable systems one or more governing principles, such as optimization or equilibrium, provide a formal and explicit way to distinguish between structurally exogenous and endogenous variables, permitting explicitly causal roles not only for endogenous but also for exogenous variables. Attributes are unambiguously defined as constants (numbers, sets, functions) associated with the system units that define fundamental aspects of the decision problem represented by the settable system.

The Rubin treatment effect approach to causal inference (e.g., as formalized by Holland, 1986) also relates to settable systems. We leave a careful study of the relations between these two ap-

proaches to other work in order to keep a sharp and manageable focus for this paper. Thus, our goal here is to compare and contrast our approach with the PCM, which, along with structural equation systems in econometrics, comprise the frameworks that primarily motivate settable systems. Nevertheless, some brief discussion of the relation of settable systems to Rubin's treatment effect approach is clearly warranted. In our view, the main feature that distinguishes settable systems (and the PCM) from the Rubin model is the explicit representation of the full causal structure. This has significant implications for the selection of covariates and for providing primitive conditions that deliver unconfoundedness conditions as consequences in settable systems, rather than introducing these as maintained assumptions in the Rubin model. Explicit representation of the full causal structure also has important implications for the analysis of "simultaneous" systems and mutually causal relations, which are typically suppressed in the Rubin approach. Finally, the allowance for a countable number of system units, the partitioning device of settable systems, and settable systems' more thorough exploitation of attributes also represent useful differences with Rubin's model.

The plan of this paper is as follows. In Section 2, we give a succinct statement of the elements of the PCM and of a generalization due to Halpern (2000) relevant for motivating and developing our settable systems extension.

Section 3 contains a series of closely related machine learning examples in which we examine the features and limitations of the PCM. These in turn help motivate features of the settable systems framework. Our examples involve least squares-based machine learning algorithms for simple artificial neural networks useful for making predictions. We consider learning algorithms with and without weight clamping and network structures with and without hidden units. Because learning is based on principles of optimization (least squares), our discussion relates to decision problems generally.

Our examples in Section 3 show that although the PCM applies to key aspects of machine learning, it also fails to apply to important classes of problems. One source of these limitations is the PCM's unique fixed point requirement. Although Halpern's (2000) generalization does not impose this requirement, it has other limitations. We contrast these with settable systems, where there is no fixed point requirement, but where fixed points may help determine system outcomes. The feature of settable systems delivering this flexibility is partitioning, an analog of the submodel and do operator devices of the PCM.

The examples of Section 3 do not involve randomness. We introduce randomness in Section 4, using our machine learning examples to discuss heuristic aspects of stochastic settable systems. We compare and contrast these with aspects of Pearl's probabilistic causal model. An interesting feature of stochastic settable systems is that attributes can determine the governing probability measure. In contrast, attributes are random variables in the PCM. Straightforward notions of counterfactuals, interventions, direct causes, direct effects, and total effects emerge naturally from stochastic settable systems.

Section 5 integrates the features of settable systems motivated by our examples to provide a rigorous formal definition of stochastic settable systems.

In Section 6 we use a series of examples from game theory to show how settable systems apply to groups of interacting and strategically competitive decision-making agents. Game theoretic structures have broad empirical relevance; they also present interesting opportunities for distributed and emergent computation of important quantities, such as prices. The decision-making agents may be consumers, firms, or government entities; they may also be biological systems or artificial intelligences, as in automated trading systems. Our demonstrations thus provide foundations for causal

analysis of systems where optimization and equilibrium mechanisms both operate to determine system outcomes. We relate our results to multi-agent influence diagrams (Koller and Milch, 2003) in Section 6.4.

In Section 7 we close the loop by considering examples from a general class of machine learning algorithms with feedback introduced by Kushner and Clark (1978) and extended by Chen and White (1998). These systems contain not only learning methods involving possibly hidden states, such as the Kalman filter (Kalman, 1960) and recurrent neural networks (e.g., Elman, 1990; Jordan, 1992; Kuan, Hornik, and White, 1994), but also systems of groups of strategically interacting and learning decision makers, as shown by Chen and White (1998). These systems exhibit optimization, equilibrium, and learning and map directly to settable systems, providing foundations for causal analysis in such systems.

Section 8 contains a summary and a discussion of research relying on the foundations provided here as well as discussion of directions for future work. An Appendix contains supplementary material; specifically, we give a formal definition of nonstochastic settable systems.

In a recent review of Pearl's book for economists and econometricians, Neuberg (2003) expresses a variety of reservations and concerns. Nevertheless, Neuberg (2003, p. 685) recommends that "econometricians should read *Causality* and start contributing to the cross-disciplinary discussion of the subject that Pearl has begun. Hopefully mutual enlightenment will be the effect of our reading and talking about the book among ourselves and with the Bayesian causal network thinkers." By examining aspects of what can and cannot be accommodated within Pearl's framework, and by proposing settable systems as an extension of this framework designed to accommodate features of central interest to economists, namely optimization, equilibrium, and learning, we offer this paper as part of this dialogue.

2. Pearl's Causal Model

Pearl's definition of a *causal model* (Pearl, 2000, Def. 7.1.1, p. 203) provides a formal statement of the elements essential to causal reasoning. According to this definition, a causal model is a triple $M := (u, v, f)$, where $u := \{u_1, ..., u_m\}$ is a collection of "background" variables determined outside the model, $v := \{v_1, ..., v_n\}$ is a collection of "endogenous" variables determined within the model, and $f := \{f_1, ..., f_n\}$ is a collection of "structural" functions that specify how each endogenous variable is determined by the other variables of the model, so that $v_i = f_i(v_{(i)}, u), \quad i = 1, ..., n$. Here $v_{(i)}$ denotes the vector containing every element of v except v_i. The integers m and n are finite. We refer to the elements of u and v as system "units."

Finally, the definition requires that f yields a unique fixed point for each u, so that there exists a unique collection $g := \{g_1, ..., g_n\}$ such that for each u,

$$v_i = g_i(u) = f_i(g_{(i)}(u), u), \quad i = 1, ..., n.$$

The unique fixed point requirement is a crucial aspect of the PCM, as this ensures existence of the *potential response function* (Pearl, 2000, Def. 7.1.4). This provides the foundation for discourse about causal relations between endogenous variables; this discourse is not possible in the PCM otherwise. A variant of the PCM analyzed by Halpern (2000) does not require a fixed point, but if any exist, there may be multiple collections of functions g yielding a fixed point. We refer to such a model as a Generalized Pearl Causal Model (GPCM). We note that GPCMs do not possess an analog of the potential response function, due to the lack of a unique fixed point.

In presenting the elements of the PCM, we have adapted Pearl's original notation somewhat to facilitate the discussion to follow, but all essential elements of the definition are present and complete.

3. Machine Learning, the PCM, and Settable Systems

We now consider how machine learning can be viewed in the context of the PCM. We consider machine learning examples that fundamentally involve optimization, a feature of a broad range of physical, natural, and social systems.[1] Specifically, optimization lies at the heart of most decision problems, as these problems typically involve deciding which of a range of possible options delivers the best expected outcome given the available information. When machine learning is based on optimization, it represents a prototypical decision problem. As we show, certain important aspects of machine learning map directly to the PCM. This permits us to investigate which causal questions are meaningful for machine learning within the PCM, and it motivates the modifications and refinements that lead to settable systems and the more extensive causal discourse possible there.

3.1 A Least-Squares Learning Example

Our first example considers predicting a random variable Y using a single random predictor X and an artificial neural network. In particular, we study the causal consequences for the optimal network weights of interventions to certain parameters of the joint distribution of the randomly generated X and Y.

More specifically, the output of an artificial neural network having a simple linear architecture is given by

$$f(X;\alpha,\beta) = \alpha + \beta X.$$

We suppose that Y and X are randomly generated according to a joint distribution F_γ indexed by a vector of parameters γ belonging to the parameter space Γ. We thus view γ as a variable whose values may range over Γ. For clarity, we suppose that γ is not influenced by our prediction (e.g., a weather forecast or an economic growth forecast).

We evaluate network performance in terms of expected squared prediction error loss,

$$\begin{aligned} L(\alpha,\beta,\gamma) &:= E_\gamma([Y - f(X;\alpha,\beta)]^2) \\ &= \int [y - f(x;\alpha,\beta)]^2 dF_\gamma(x,y), \end{aligned}$$

where $E_\gamma(\cdot)$ denotes expectation taken with respect to the distribution F_γ. Our goal is to obtain the best possible predictions according to this criterion. Accordingly, we seek loss-minimizing network weights, which solve the optimization problem

$$\min_{\alpha,\beta} L(\alpha,\beta,\gamma).$$

This makes it explicit that the governing principle in this example is optimization.

Under mild conditions, least squares-based machine learning algorithms converge to the optimal weights as the size of the training data set grows. For clarity, we work for now with the optimal network weights.

1. The great mathematician Leonhard Euler once wrote, "nothing at all takes place in the Universe in which some rule of maximum or minimum does not appear" (as quoted in Marsden and Tromba, 2003).

For our linear network, the first order conditions necessary for an optimum are

$$\begin{aligned}(\partial/\partial\alpha)L(\alpha,\beta,\gamma) &= -2E_\gamma([Y-\alpha-\beta X]) = 0,\\ (\partial/\partial\beta)L(\alpha,\beta,\gamma) &= -2E_\gamma(X[Y-\alpha-\beta X]) = 0.\end{aligned}$$

Letting $\mu_X := E_\gamma(X)$, $\mu_Y := E_\gamma(Y)$, $\mu_{XX} := E_\gamma(X^2)$, and $\mu_{XY} := E_\gamma(XY)$, we can conveniently parameterize F_γ in terms of the moments $\gamma := (\mu_X, \mu_Y, \mu_{XX}, \mu_{XY})$. (This parameterization need not uniquely determine F_γ; that is, there may be multiple distributions F_γ for a given γ. Nevertheless, this γ is the only aspect of the distribution that matters here.) We can then express the first order conditions equivalently as

$$\begin{aligned}\mu_Y - \alpha - \beta\mu_X &= 0,\\ \mu_{XY} - \alpha\mu_X - \beta\mu_{XX} &= 0.\end{aligned}$$

Now consider how this system fits into Pearl's causal model. Pearl's model requires a system of equations in which the left-hand side variables are structurally determined by the right-hand side variables. The first order conditions are not in this form, but, provided $\mu_{XX} - \mu_X^2 > 0$, they can be transformed to this form by solving jointly for α and β :

$$\begin{aligned}\alpha^* &= \mu_Y - [\mu_{XX} - \mu_X^2]^{-1}(\mu_{XY} - \mu_X\mu_Y)\mu_X,\\ \beta^* &= [\mu_{XX} - \mu_X^2]^{-1}(\mu_{XY} - \mu_X\mu_Y).\end{aligned} \tag{1}$$

We write (α^*, β^*) to distinguish optimized values from generic values (α, β).

This representation demonstrates that the PCM applies directly to this machine learning problem. The equations in (1) form a system in which the background (or "structurally exogenous") variables $u := (u_1, u_2, u_3, u_4) = (\mu_X, \mu_Y, \mu_{XX}, \mu_{XY}) =: \gamma$ determine the endogenous variables $v := (v_1, v_2) = (\alpha^*, \beta^*)$. The structural functions (f_1, f_2) are defined by

$$\begin{aligned}f_1(u) &= u_2 - [u_3 - u_1^2]^{-1}(u_4 - u_1u_2)u_1,\\ f_2(u) &= [u_3 - u_1^2]^{-1}(u_4 - u_1u_2).\end{aligned}$$

We observe that by the conventions of the PCM, the background variables u do not have formal status as causes, as we further discuss below.

In discussing the PCM, Pearl (2000, p. 203) notes that the background variables are often unobservable, but this is not a formal requirement of the PCM. In our example, we may view the γ variables as either observable or unobservable, depending on the context. For example, suppose we are given a linear least-squares learning machine as a black box: we know that it is a learning machine, but we don't know of what kind. To attempt to determine what is inside the black box, we can conduct computer experiments in which we set γ to various known values and observe the resulting values of (α^*, β^*). In this case, γ is observable.

Alternatively, we may have a least-squares learning machine that we apply to a variety of data sets obeying the distribution F_γ for differing unknown values of γ. In each case, γ is unobservable, but we can generate as much data as we want from F_γ.

Intermediate cases are also possible, in which some elements of γ are known and others are not. For example, in the multiple data set example, we could have knowledge of a subvector of γ, for example, we might know μ_X.

3.2 Learning with Clamping

Next, we study the effects on one of the optimal network weights of interventions to the other weight. For this, we consider the optimal network weights that arise when one or the other of the network weights is clamped, that is, set to an arbitrary fixed value. Specifically, consider

$$\min_{\alpha} L(\alpha,\beta,\gamma) \text{ and } \min_{\beta} L(\alpha,\beta,\gamma).$$

Clamping is useful in "nested" or multi-stage optimization, as

$$\begin{aligned} \min_{\alpha,\beta} L(\alpha,\beta,\gamma) &= \min_{\beta}[\min_{\alpha} L(\alpha,\beta,\gamma)] \text{ and} \\ \min_{\alpha,\beta} L(\alpha,\beta,\gamma) &= \min_{\alpha}[\min_{\beta} L(\alpha,\beta,\gamma)]. \end{aligned}$$

See, for example, Sergeyev and Grishagin (2001). Clamping is a central feature of a variety of powerful machine learning algorithms, for example, the restricted Boltzmann machine (e.g., Ackley et al., 1985; Hinton and Sejnowski, 1986; Hinton et al., 2006; Hinton and Salakhutdinov, 2006). Learning in stages is particularly useful in cases involving complex optimizations, as in the EM algorithm (Dempster, Laird, and Rubin, 1977).

The first order condition necessary for the β-clamped optimum $\min_\alpha L(\alpha,\beta,\gamma)$ is

$$(\partial/\partial\alpha)L(\alpha,\beta,\gamma) = -2E_\gamma([Y-\alpha-\beta X]) = 0.$$

Equivalently, $\mu_Y - \alpha - \beta\mu_X = 0$. Solving for the optimal α weight gives

$$\tilde{\alpha}^* = \mu_Y - \beta\mu_X. \tag{2}$$

We use the tilde notation to distinguish between the optimal weights with clamping and the jointly optimal weights obtained above.

Similarly, the first order condition necessary for the α-clamped optimum $\min_\beta L(\alpha,\beta,\gamma)$ is

$$(\partial/\partial\beta)L(\alpha,\beta,\gamma) = -2E_\gamma(X[Y-\alpha-\beta X]) = 0.$$

Equivalently, $\mu_{XY} - \alpha\mu_X - \beta\mu_{XX} = 0$. Given $\mu_{XX} > 0$, the optimal weight with clamping is

$$\tilde{\beta}^* = \mu_{XX}^{-1}(\mu_{XY} - \alpha\mu_X). \tag{3}$$

Writing Equations (2) and (3) as a system, we have

$$\tilde{\alpha}^* = \mu_Y - \beta\mu_X \qquad \tilde{\beta}^* = \mu_{XX}^{-1}(\mu_{XY} - \alpha\mu_X). \tag{4}$$

This resembles a structural system in the form of the PCM, except that here $\tilde{\alpha}^*$ and $\tilde{\beta}^*$ appear on the left, instead of α and β. This difference is significant; we address this shortly.

Nevertheless, suppose for the moment that we ignore this difference and modify the system above to conform to the PCM by replacing $\tilde{\alpha}^*$ and $\tilde{\beta}^*$ with α and β :

$$\alpha = \mu_Y - \beta\mu_X \qquad \beta = \mu_{XX}^{-1}(\mu_{XY} - \alpha\mu_X). \tag{5}$$

We take $u = \gamma$ as above, but in keeping with our conforming modification, we now take $(v_1, v_2) = (\alpha, \beta)$. The structural functions become

$$\tilde{f}_1(u, v_2) = u_2 - v_2 u_1 \qquad \tilde{f}_2(u, v_1) = u_3^{-1}(u_4 - v_1 u_1).$$

This system falls into the PCM, with consequent causal status for v, provided there is a unique fixed point for each u.

Unfortunately, this fixed point requirement fails here. As is apparent from the equations in (5), the only necessary restriction on u is that $u_3 = \mu_{XX} > 0$. This is the requirement that X is not equal to 0 with probability one. Nevertheless, it is readily verified that even with this restriction, the fixed point requirement fails for all u such that

$$u_3 - u_1^2 = \mu_{XX} - \mu_X^2 = 0.$$

This is the condition that $X = \mu_X$ with probability one, and μ_X can take any value, not just zero. When this condition holds, there is an uncountable infinity of fixed point solutions to the equations in (5). Stated another way, the solution to the system is set-valued in this circumstance.

Because of the lack of a fixed point, the PCM does not apply and therefore cannot provide causal meaning for such a system. The inability of the PCM to apply to this simple example of machine learning with clamping is an unfortunate limitation. Because Halpern's (2000) GPCM does not require a unique fixed point, it does apply here. Nevertheless, the lack of the potential response function in the GPCM prevents the desired causal discourse.

3.3 Settable Systems and Learning with Clamping

We now consider how these issues can be addressed. Our intent is to encompass this example while preserving the spirit of the PCM. This motivates and helps illustrate various features of our settable systems framework.

3.3.1 SETTABLE VARIABLES

We begin by taking seriously the difference in roles between (α, β) and $(\tilde{\alpha}^*, \tilde{\beta}^*)$ appearing in the equations in (4). In the simplest sense, the difference is that $(\tilde{\alpha}^*, \tilde{\beta}^*)$ and (α, β) appear on different sides of the equal signs: (α, β) appears on the right and $(\tilde{\alpha}^*, \tilde{\beta}^*)$ on the left. In the PCM, this difference is fundamentally significant, in that causal relations are asymmetric, with structurally determined (endogenous) variables on the left and all other variables on the right. In settable systems, we formalize these dual roles by defining *settable variables* as mappings $\mathcal{X}$ with a dual aspect:

$$\begin{aligned} \mathcal{X}_1(0) &:= \tilde{\alpha}^*, \quad \mathcal{X}_1(1) := \alpha, \\ \mathcal{X}_2(0) &:= \tilde{\beta}^*, \quad \mathcal{X}_2(1) := \beta. \end{aligned} \tag{6}$$

We call the $0-1$ argument of the settable variables $\mathcal{X}$ the "role indicator." When this is 0, the value of the variable is that determined by its structural equation. We call these values *responses*. In contrast, when the role indicator is 1, the value is not determined by its structural equation, but is instead set to one of its admissible values. We call these values *settings*. We require that a setting has *more than one* admissible value. That is, settings are variable.

Formally distinguishing between responses and settings makes explicit the dual roles played by variables in a causal system, entirely in the spirit of the PCM. Settable variables represent a formal

implementation, alternative to that of the "do operator" in the PCM, of the "wiping out" operation first proposed by Strotz and Wold (1960) and later used by Fisher (1970).

Once we make explicit the dual roles of the system variables, several benefits become apparent. First, the equal sign no longer has to serve in an asymmetric manner. This makes possible *implicit* representations of causal relations in settable systems that are either not possible in the PCM, because the required closed-form expressions do not exist; or that are possible in the PCM only under restrictions permitting application of the implicit function theorem. Such implicit representations are often natural for responses satisfying first order conditions arising from optimization. To illustrate, consider how explicit representation of the dual roles of system variables modifies the learning with clamping system. The first order condition necessary for the β-clamped optimum $\min_\alpha L(\alpha,\beta,\gamma)$ is now

$$\mu_Y - \tilde{\alpha}^* - \beta\mu_X = 0.$$

That for the α-clamped optimum $\min_\beta L(\alpha,\beta,\gamma)$ is now

$$\mu_{XY} - \alpha\mu_X - \tilde{\beta}^*\mu_{XX} = 0.$$

The structural system thus has the implicit representation

$$\mu_Y - \tilde{\alpha}^* - \beta\mu_X = 0, \tag{7}$$
$$\mu_{XY} - \alpha\mu_X - \tilde{\beta}^*\mu_{XX} = 0. \tag{8}$$

3.3.2 SETTABLE SYSTEMS AND THE ROLE OF FIXED POINTS

A second benefit of making explicit the dual roles of the system variables is that unique fixed points do not have a crucial role to play in settable systems. This enables us to dispense with the unique fixed point requirement prohibiting the PCM from encompassing our learning with clamping example. This is not to say that fixed points have no role to play. Instead, that role is removed from the structural representation of the system and, to the extent relevant, operates according to the governing principle, for example, optimization or equilibrium. We discuss this further below.

To illustrate, consider the learning with clamping system above where the dual roles of the system variables are made explicit. Now there is no necessity of finding a fixed point for Equations (7) and (8). Each equation stands on its own, representing its associated clamped optimum.

The simplest case is that for $\tilde{\alpha}^*$. For every μ_X, μ_Y, and β, there is a unique solution,

$$\tilde{\alpha}^* = \mu_Y - \beta\mu_X =: \tilde{r}_1(\beta,\gamma).$$

We call $\tilde{r}_1$ the *response function* for $\mathcal{X}_1$.

Next consider $\tilde{\beta}^*$. Provided $\mu_{XX} > 0$, Equation (8) determines a unique value for $\tilde{\beta}^*$,

$$\tilde{\beta}^* = \mu_{XX}^{-1}(\mu_{XY} - \alpha\mu_X).$$

But what happens when $\mu_{XX} = 0$? This further implies $\mu_X = \mu_{XY} = 0$. Consequently, *any* value will do for $\tilde{\beta}^*$, as any value of $\tilde{\beta}^*$ delivers the best possible prediction. To arrive at a unique value for $\tilde{\beta}^*$, we can apply criteria supplemental to predictive optimality. For example, we may choose a value that has the simplest representation. This reduces the viable choices to $\tilde{\beta}^* \in \{0,1\}$, as either of these requires only one bit to represent. Finally, by selecting $\tilde{\beta}^* \in \{0\}$, so that we set $\tilde{\beta}^* = 0$ when

$\mu_{XX} = 0$, we achieve a prediction, $f(X;\alpha,\tilde{\beta}^*) = \alpha$, that requires the fewest operations to compute. Formally, this gives

$$\tilde{\beta}^* = 1_{\{\mu_{XX}>0\}}\mu_{XX}^{-1}(\mu_{XY} - \alpha\mu_X) =: \tilde{r}_2(\alpha,\gamma),$$

where $1_{\{\mu_{XX}>0\}}$ is the indicator function taking the value one when $\mu_{XX} > 0$, and zero otherwise. We call $\tilde{r}_2$ the response function for $\mathcal{X}_2$.

This example demonstrates that even when structural equations conforming to the PCM (i.e., Equation 5) do not have a fixed point, we can find unique response functions for each settable variable of the analogous settable system. We do this by applying the governing principle for the system (e.g., optimization), supplemented when necessary by further appropriate principles (e.g., parsimony of memory and computation).

Applying the settable variable representation in the equations in (6), we obtain a settable variables representation for our learning with clamping example:

$$\mathcal{X}_1(0) = \tilde{r}_1(\mathcal{X}_2(1),\gamma), \qquad \mathcal{X}_2(0) = \tilde{r}_2(\mathcal{X}_1(1),\gamma).$$

So far, the variable γ has not been given status as a settable variable. Although it does not have a dual aspect, it can be set to any of several admissible values (those in Γ), so it does have the aspect of a setting. Accordingly, we can define $\mathcal{X}_0(1) := \gamma$. To ensure that $\mathcal{X}_0$ is a well-defined settable variable, we must also specify a value for $\mathcal{X}_0(0)$. By convention, we simply put $\mathcal{X}_0(0) := \mathcal{X}_0(1)$. We call $\mathcal{X}_0$ *fundamental* settable variables. As these are determined outside the system, they are structurally exogenous.

We can now give an explicit settable system representation for our present example, that is, a representation solely in terms of settable variables:

$$\mathcal{X}_1(0) = \tilde{r}_1(\mathcal{X}_2(1),\mathcal{X}_0(1)) \qquad \mathcal{X}_2(0) = \tilde{r}_2(\mathcal{X}_1(1),\mathcal{X}_0(1)).$$

3.4 Causes and Effects: Settable Systems and the PCM

This section introduces causal notions appropriate to settable systems.

3.4.1 DIRECT CAUSALITY

We begin by considering our learning with clamping example, where

$$\tilde{\alpha}^* = \tilde{r}_1(\beta,\gamma), \quad \tilde{\beta}^* = \tilde{r}_2(\alpha,\gamma).$$

In particular, consider the equation $\tilde{\beta}^* = \tilde{r}_2(\alpha,\gamma)$. In settable systems, settings are variable, that is, they can take any of a range of admissible values. We view this as sufficient to endow them with potential causal status. Thus, we call α and γ *potential causes* of $\tilde{\beta}^*$.

We say that a given element of (α,γ) *does not directly cause* $\tilde{\beta}^*$ if $\tilde{r}_2(\alpha,\gamma)$ defines a function constant in the given element for all admissible values of the other elements of (α,γ). Otherwise, that element is a *direct cause* of $\tilde{\beta}^*$. According to this definition, μ_Y does not directly cause $\tilde{\beta}^*$, whereas $\mu_X, \mu_{XX}, \mu_{XY}$, and α are direct causes of $\tilde{\beta}^*$.

3.4.2 INTERVENTIONS AND DIRECT EFFECTS IN SETTABLE SYSTEMS

In settable systems, an *intervention to a settable variable* is a pair of distinct admissible setting values. In our clamped learning example, let α_1 and α_2 be different admissible values for α.

Then $\alpha_1 \to \alpha_2 := (\alpha_1, \alpha_2)$ is an intervention to α, or, more formally, to $\mathcal{X}_1$. Similarly, $(\alpha_1, \gamma_1) \to (\alpha_2, \gamma_2) := ((\alpha_1, \gamma_1), (\alpha_2, \gamma_2))$ is an intervention to (α, γ) (i.e., to $(\mathcal{X}_1, \mathcal{X}_0)$). The *direct effect* on a given settable variable of a specified intervention is the response difference arising from the intervention. In our clamped learning example, the direct effect on $\mathcal{X}_2$ of the intervention $\alpha_1 \to \alpha_2$ is

$$\begin{aligned} \Delta\tilde{r}_2(\alpha_1, \alpha_2; \gamma) &:= \tilde{r}_2(\alpha_2, \gamma) - \tilde{r}_2(\alpha_1, \gamma) \\ &= 1_{\{\mu_{XX} > 0\}} \mu_{XX}^{-1} (\alpha_1 - \alpha_2) \mu_X . \end{aligned}$$

We emphasize that interventions are always well defined, as settings necessarily have more than one admissible value. Indeed, a key reason that we require settings to be variable is precisely to ensure that interventions to settable variables are always meaningful.

PCM notions related to the settable systems notion of intervention are the do operator and the "effect of action" defined in definition 7.1.3 of Pearl (2000); these specify a submodel associated with a given realization *x* for a given subset of the endogenous variables *v*.

3.4.3 EXOGENOUS AND ENDOGENOUS CAUSES

The notion of causality just defined contrasts in an interesting way with that formally given in the PCM. We have just seen that γ can serve in the settable system as a direct cause of $\tilde{\beta}^*$. Above, we saw that γ corresponds to background variables *u* in the PCM. In the PCM, the formal concept of *submodel* and the *do operator* necessary to define causal relations are meaningful only for endogenous variables *v*. None of these concepts are defined for *u*; that is, *u* is not subject to counterfactual variation in the PCM.[2] Consequently, *u* does not have formal causal status in the PCM as defined in Pearl (2000, Chap. 7).

In the PCM, *u* thus has four explicit distinguishing features: it is (i) a vector of variables that (ii) are determined outside the system, (iii) determine the endogenous variables, and (iv) are not subject to counterfactual variation. An optional but common feature of *u* is: (v) it is unobservable. As a result, background variables cannot act as causes in the PCM; in particular, for a given system, the PCM formally rules out structurally exogenous unobserved causes.

In settable systems, we drop requirement (iv) for structurally exogenous variables. Thus, we allow for observed structurally exogenous causes such as a treatment of interest in a controlled experiment, which is typically directly set (and observed) by the researcher. We also allow for unobservable structurally exogenous causes, ensuring a causal framework that is not relative to the capabilities of the observer, as is appropriate to the macroscopic, non-quantum mechanical systems that are the strict focus of our attention here. Unobserved common causes are particularly relevant for the analysis of confounding, that is, the existence of hidden causal relations that may prevent the identification of causal effects of interest (see Pearl, 2000, Chap. 3.3-3.5). Also, unobserved structurally exogenous causes are central to errors-in-variables models where a structurally exogenous cause of interest cannot be observed. Instead, one observes an version of this cause contaminated by measurement error. These models are the subject of a vast literature in statistics and econometrics (see, e.g., van Huffel and Lemmerling, 2002, and the references there).

Dropping (iv) in settable systems creates no difficulties in defining causal relations, as direct causality is a property solely of the response function on its domain. Moreover, by requiring that settings have more than one admissible value, we ensure that these domains contain at least two

2. We are grateful to two of the referees for emphasizing this.

points, making possible the interventions supporting definitions of effects in settable systems. We will return to this point shortly.

In the PCM, endogenous variables are usually observable, although this is not formally required. Structurally endogenous settable variables X may also be observable or not.

Fortunately, the PCM treats a variable as a background variable or an endogenous variable relative to the analysis. If the effects of a variable are of interest, it can be converted to an endogenous variable in an alternative PCM. Nevertheless, the PCM does not provide guidance on whether to treat a variable as a background variable or an endogenous one. This decision is entirely left to the researcher's discretion. For example, the "disturbances" in the Markovian PCM "represent background variables that the investigator chooses not to include in the analysis" (Pearl, 2000, p. 68), but the PCM does not specify how an investigator chooses to include variables in the analysis. Nor is it clear that background variables are necessary to the analysis in the first place. For example, Dawid (2002, p. 183) states that "when the additional variables are pure mathematical fictions, introduced merely so as to reproduce the desired probabilistic structure of the domain variables, there seems absolutely no good reason to include them in the model."

Settable systems permit but do not require background variables. Further, and of particular significance, in a settable system a governing principle such as optimization provides a formal way to distinguish between fundamental settable variables (exogenous variables) and other settable variables (endogenous variables). In particular, the decision problem determines if a variable is exogenous or endogenous. For instance, in our clamped learning example, the optimal network weights $\tilde{\alpha}^*$ and $\tilde{\beta}^*$ minimize the loss function $L(\alpha, \beta, \gamma)$. On the other hand, although the elements of γ are variables, our learning example does not specify a decision problem that determines how these are generated. This distinction endows the variables $\tilde{\alpha}^*$ and $\tilde{\beta}^*$ with the status of endogenous variables and the elements of γ with the status of structurally exogenous variables.

Thus, carefully and explicitly specifying the decision problems and governing principles in settable systems provides a systematic way to distinguish between exogenous and endogenous variables. This formalizes and extends the distinctions between the PCM endogenous and exogenous variables.

The PCM has been fruitfully applied in the sciences (e.g., Shipley, 2000). Nevertheless, because the PCM is agnostic concerning the status of variables, two researchers may employ two possibly inconsistent PCMs to study the same scientific phenomena. To resolve such inconsistencies, one may use the fact that under suitable assumptions, causal relations imply empirically testable conditional independence relations among system variables (Pearl, 2000; Chalak and White, 2008b). This yields procedures for falsifying causal structures that are inconsistent with data. Such procedures at best identify a class of observationally equivalent causal models, so resolution of inconsistencies by this means is not guaranteed. On the other hand, specifying the decision problems underlying the phenomena of interest may, among other things, offer guidance as to which (if either) model is more suitable to the analysis. The settable systems framework provides the foundation necessary for this in the context of optimally interacting agents under uncertainty. We emphasize that agents and their decision problems may be defined in such a way as to apply even to physical or biological systems not usually thought of in these terms; any system involving optimizing (e.g., least energy, maximum entropy) and/or equilibrium falls into this framework.

3.5 Unclamped Learning and Settable Systems

Now consider how our original unclamped learning example is represented using settable systems. We begin by recognizing that the solution to a given optimization problem need not be unique, but is in general a set. When the solution depends on other variables, the solution is in general a correspondence, not a function (see, e.g., Berge, 1963). Thus, we write the solution to the unclamped learning problem as

$$(\mathbb{A}^*(\gamma), \mathbb{B}^*(\gamma)) := \arg\min_{\alpha,\beta} L(\alpha, \beta, \gamma),$$

where $\mathbb{A}^*(\gamma)$ and $\mathbb{B}^*(\gamma)$ define correspondences.

Due to the linear network architecture, we can explicitly represent $\mathbb{A}^*(\gamma)$ and $\mathbb{B}^*(\gamma)$ as

$$\begin{aligned} \mathbb{A}^*(\gamma) &= \{\alpha : [\mu_{XX} - \mu_X^2](\alpha - \mu_Y) + (\mu_{XY} - \mu_X\mu_Y)\mu_X = 0\}, \\ \mathbb{B}^*(\gamma) &= \{\beta : [\mu_{XX} - \mu_X^2]\beta - (\mu_{XY} - \mu_X\mu_Y) = 0\}. \end{aligned}$$

When $\mu_{XX} - \mu_X^2 > 0$, $\mathbb{A}^*(\gamma)$ and $\mathbb{B}^*(\gamma)$ each have a unique element, namely

$$\begin{aligned} \alpha^* &= \mu_Y - [\mu_{XX} - \mu_X^2]^{-1}(\mu_{XY} - \mu_X\mu_Y)\mu_X, \\ \beta^* &= [\mu_{XX} - \mu_X^2]^{-1}(\mu_{XY} - \mu_X\mu_Y). \end{aligned}$$

When $\mu_{XX} - \mu_X^2 = 0$, we can select a unique value from each of $\mathbb{A}^*(\gamma)$ and $\mathbb{B}^*(\gamma)$. Choosing the simplest representation and the simplest computation of the prediction yields $\alpha^* = \mu_Y$ and $\beta^* = 0$. We thus represent optimal weights using response functions r_1 and r_2 as

$$\begin{aligned} \alpha^* &= r_1(\gamma) := \mu_Y - 1_{\{\mu_{XX} - \mu_X^2 > 0\}}[\mu_{XX} - \mu_X^2]^{-1}(\mu_{XY} - \mu_X\mu_Y)\mu_X, \\ \beta^* &= r_2(\gamma) := 1_{\{\mu_{XX} - \mu_X^2 > 0\}}[\mu_{XX} - \mu_X^2]^{-1}(\mu_{XY} - \mu_X\mu_Y). \end{aligned}$$

These response functions do represent fixed points of the equations in (5). This illustrates the role that fixed points can play in determining the response functions. Observe, however, that we do not require a *unique* fixed point.

Applying the settable system definition of direct causality, we have that a given element of γ, say γ_i, does not directly cause α^* (resp. β^*) if $r_1(\gamma)$ (resp. $r_2(\gamma)$) defines a function constant in γ_i for all admissible values of the other elements of γ. Otherwise, that element is a direct cause of α^* (resp. β^*). Here, each element of γ directly causes both α^* and β^*.

In this example, we have the settable system representation

$$\mathcal{X}_1(0) = r_1(\mathcal{X}_0(1)), \qquad \mathcal{X}_2(0) = r_2(\mathcal{X}_0(1)),$$

where $\mathcal{X}_0(0) := \mathcal{X}_0(1) := \gamma, \mathcal{X}_1(1) := \alpha, \mathcal{X}_2(1) := \beta$ as before, but now $\mathcal{X}_1(0) := \alpha^*$ and $\mathcal{X}_2(0) := \beta^*$.

Finally, we note that the system outputs of the clamped and unclamped systems are *mutually consistent*, in the sense that if we plug the responses of the unclamped system into the response functions $(\tilde{r}_1, \tilde{r}_2)$ of the clamped system as settings, we obtain clamped responses that replicate the responses of the unclamped system. That is, putting $\mathcal{X}_1^c(1) = \mathcal{X}_1^u(0)$ and $\mathcal{X}_2^c(1) = \mathcal{X}_2^u(0)$, where we now employ the superscripts *c* and *u* to clearly distinguish clamped and unclamped system settable variables, we have

$$\mathcal{X}_1^u(0) = \tilde{r}_1(\mathcal{X}_2^u(0), \mathcal{X}_0(1)), \qquad \mathcal{X}_2^u(0) = \tilde{r}_2(\mathcal{X}_1^u(0), \mathcal{X}_0(1)),$$

as some simple algebra will verify. This mutual consistency is ensured by the governing principle of optimization.

3.6 Partitioning in Settable Systems

In the PCM, the role of submodels (Pearl, 2000, Def. 7.1.2) is to specify which endogenous variables are subject to manipulation; the do operator specifies which values the manipulated variables take. In settable systems, submodels and the do operator are absent. Nevertheless, settable systems do have an analog of submodels, but instead of specifying which variables are to be manipulated, a settable system specifies which system variables are free to respond to the others. In our learning examples, settable systems specify which variables are unclamped. In our first example, both variables are unclamped. In the second example, the variables are considered one at a time, and each variable is unclamped in turn.

3.6.1 PARTITIONING

A formal mathematical implementation of these specifications is achieved by *partitioning*. Partitioning operates on an index set I whose elements are in one-to-one correspondence to the structurally endogenous (non-fundamental) settable variables. In our learning examples, there are two such variables, so the index set can be chosen to be $I = \{1,2\}$.

Let I be any set with a countable number of elements. A partition Π is a collection of subsets $\Pi_1, \Pi_2, ...$ of I that are mutually exclusive ($\Pi_a \cap \Pi_b = \varnothing, a \neq b$) and exhaustive ($\cup_b \Pi_b = I$). Examples are the *elementary partition*, $\Pi^e := \{\Pi^e_1, ..., \Pi^e_n\}$, where $\Pi^e_1 := \{1\}, \Pi^e_2 := \{2\}, ...$, and the *global partition* $\Pi^g := \{\Pi^g_1\}$, where $\Pi^g_1 := I$.

When $I = \{1,2\}$, these are the only two possible partitions: $\Pi^e = \{\Pi^e_1, \Pi^e_2\}$, where $\Pi^e_1 = \{1\}$ and $\Pi^e_2 = \{2\}$; and $\Pi^g = \{\Pi^g_1\}$, where $\Pi^g_1 = \{1,2\}$.

We interpret the partition elements as specifying which of the system variables are jointly free to respond to the remaining variables of the system, according to the governing principle of the system (e.g., optimization). In our machine learning examples with $I = \{1,2\}$, the element $\Pi^e_1 = \{1\}$ of the elementary partition Π^e specifies that variable 1 (i.e., $\tilde{\alpha}^*$) is free to respond to all other variables of the system (i.e., (β, γ)), whereas $\Pi^e_2 = \{2\}$ specifies that variable 2 (i.e., $\tilde{\beta}^*$) is free to respond to all other variables of the system (i.e., (α, γ)). The element $\Pi^g_1 = \{1,2\}$ of the global partition specifies that variables 1 and 2 (i.e., (α^*, β^*)) are jointly free to respond to all other variables of the system (i.e., γ).

In settable systems, response functions are partition specific. With Π^e, we have

$$\tilde{\alpha}^* = \tilde{r}_1(\beta, \gamma), \quad \tilde{\beta}^* = \tilde{r}_2(\alpha, \gamma);$$

with Π^g, we have

$$\alpha^* = r_1(\gamma), \quad \beta^* = r_2(\gamma)$$

for the response functions $(\tilde{r}_1, \tilde{r}_2)$ and (r_1, r_2) defined above. This implies that the settable variables and the resulting causal relations are *partition specific*.

We note that the distinction between the response functions $(\tilde{r}_1, \tilde{r}_2)$ and (r_1, r_2) is not due to additional constraints imposed on the optimization problem per se. Instead, the distinction follows from whether learning occurs with or without clamping and hence on whether or not alpha and beta respond jointly. Thus, different optimization problems yield different corresponding partitions and response functions.

These partitioning concepts and principles extend to systems with any number of structurally endogenous variables. We discuss further examples below.

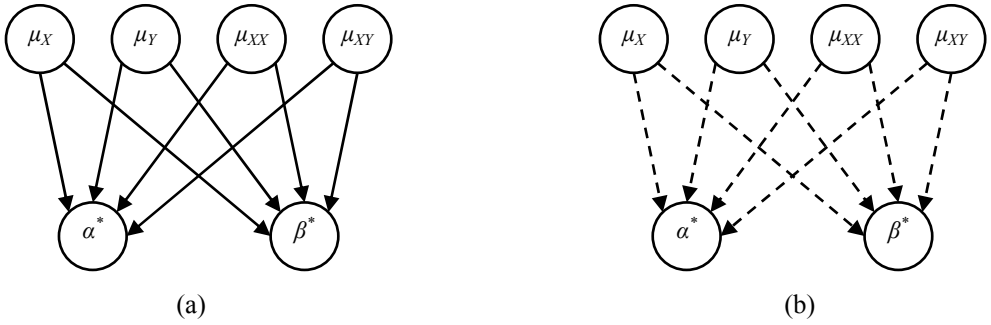

Figure 1: PCM Directed Acyclic Graphs

3.6.2 SETTABLE SYSTEM CAUSAL GRAPHS

Given the applicability of the PCM to the unclamped learning example, this system has an associated PCM directed acyclic graph (DAG). The particular representation of this graph depends on whether or not the background variables are observable or not. Figure 1(a) depicts the case of observable γ and Figure 1(b) that of unobservable γ. The solid arrows in Figure 1(a) indicate that the background variables are observable, whereas the dashed arrows in Figure 1(b) indicate that the background variables are not observable.

In interpreting these graphs, note that the arrows, whether solid or dashed, represent the functional relationships present. They do not, however, represent causal relations, as in the PCM these are defined to hold only between endogenous variables, and no arrows link the endogenous variables here. Pearl (2000) often uses the term "influence" to refer to situations involving functional dependence, but not causal dependence. In this sense, the arrows in these DAGs represent "influences."

In contrast, due to the lack of a fixed point, the PCM does not apply to the learning with clamping example. Necessarily, the PCM cannot supply a causal graph.

In settable systems, partitions play a key role in constructing causal graphs that represent direct causality relations. To see how, consider our clamped learning example. Here, μ_Y (i.e., $\mathcal{X}_{0,2}(1)$) does not directly cause $\tilde{\beta}^*$ ($\mathcal{X}_2(0)$), whereas $\mu_X, \mu_{XX}, \mu_{XY}$, and α ($\mathcal{X}_{0,1}(1), \mathcal{X}_{0,3}(1), \mathcal{X}_{0,4}(1)$, and $\mathcal{X}_1(1)$) are direct causes of $\tilde{\beta}^*$ ($\mathcal{X}_2(0)$). We can succinctly and unambiguously state these causal relations in terms of settable variables by saying that $\mathcal{X}_{0,2}$ does not directly cause $\mathcal{X}_2$, whereas $\mathcal{X}_{0,1}, \mathcal{X}_{0,3}, \mathcal{X}_{0,4}$, and $\mathcal{X}_1$ are direct causes of $\mathcal{X}_2$.

For each block Π_b of a partition $\Pi = \{\Pi_b\}$, we construct a settable system causal graph by letting nodes correspond to settable variables. If one settable variable directly causes another, we draw an arrow from the node representing the cause to that representing the responding settable variable. Note that in contrast to the DAGs for the PCM, we represent all direct causal links as solid arrows, letting dashed nodes represent unobservable settable variables. The motivation for this is that unobservability is a property of the settable variable (the node), not the links between nodes.

Figures 2(a) and 2(b) depict the causal graphs for our clamped learning example. There are two causal graphs, as the clamped learning example expresses the elementary partition $\Pi^e = \{\{1\}, \{2\}\}$. For purposes of illustration, we depict the case in which γ is unobserved.

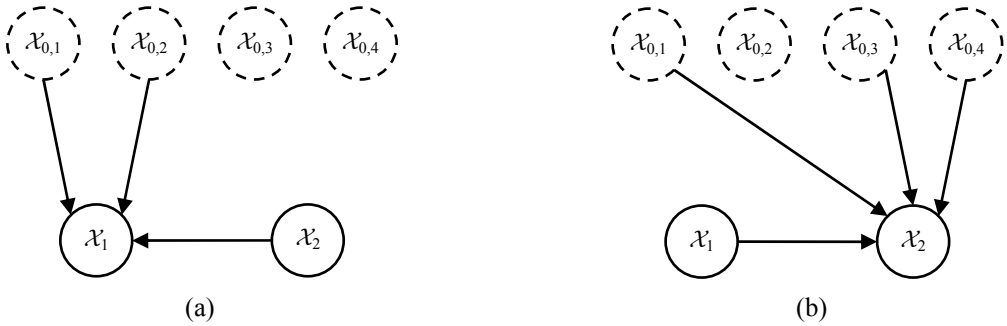

Figure 2: Block-specific Settable System Causal Graphs for the Elementary Partition

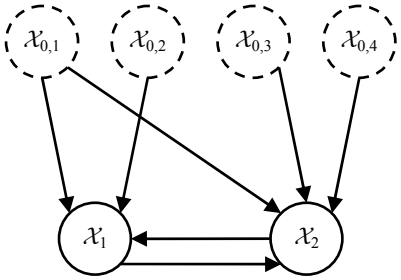

Figure 3: Settable System Superimposed Causal Graph for the Elementary Partition

For convenience, we may superimpose settable system causal graphs. Superimposing Figures 2(a) and 2(b) gives Figure 3. This is a cyclic graph. Nevertheless, this cyclicality does not represent true simultaneity; it is instead an artifact of the superimposition.

The settable system causal graph for the global partition $\Pi^g = \{\{1,2\}\}$ representing unclamped learning is depicted in Figure 4. Observe that this reproduces the connectivity of Figure 1. Note that in Figure 4, the nodes represent settable variables and the arrows represent direct causes. In Figure 1, the nodes represent background or endogenous variables and the arrows represent non-causal "influences."

We emphasize that the causal graphs associated with settable systems are not necessary to the analysis. Rather, they are sometimes helpful in succinctly representing and studying causal relations.

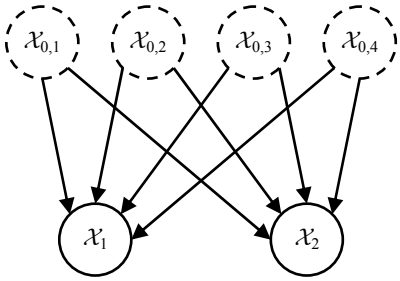

Figure 4: Settable System Causal Graph for the Global Partition

3.7 Further Examples Motivating Settable System Features

We now introduce two further features of settable systems, countable dimension and attributes, using examples involving machine learning algorithms and networks with hidden units. This provides further interesting contrasts between settable systems and the PCM.

3.7.1 A MACHINE LEARNING ALGORITHM AND COUNTABLE DIMENSIONALITY

So far, we have restricted attention to the optimal network weights for linear least-squares machine learning. Now consider the machine learning algorithm itself. For this, let

$$\hat{\mu}_{x,0} = \hat{\mu}_{y,0} = \hat{\mu}_{xx,0} = \hat{\mu}_{xy,0} = \hat{\alpha}_0 = \hat{\beta}_0 = 0,$$

and perform the recursion

$$\begin{aligned}
\hat{\mu}_{x,n} &= \hat{\mu}_{x,n-1} + n^{-1}(x_n - \hat{\mu}_{x,n-1}) \\
\hat{\mu}_{y,n} &= \hat{\mu}_{y,n-1} + n^{-1}(y_n - \hat{\mu}_{y,n-1}) \\
\hat{\mu}_{xx,n} &= \hat{\mu}_{xx,n-1} + n^{-1}(x_n^2 - \hat{\mu}_{xx,n-1}) \\
\hat{\mu}_{xy,n} &= \hat{\mu}_{xy,n-1} + n^{-1}(x_n y_n - \hat{\mu}_{xy,n-1}) \\
\hat{\beta}_n &= 1_{\{\hat{\mu}_{xx,n} - \hat{\mu}_{x,n}^2 > 0\}}[\hat{\mu}_{xx,n} - \hat{\mu}_{x,n}^2]^{-1}(\hat{\mu}_{xy,n} - \hat{\mu}_{x,n}\hat{\mu}_{y,n}) \\
\hat{\alpha}_n &= \hat{\mu}_{y,n} - \hat{\beta}_n \hat{\mu}_{x,n}, \quad n = 1, 2, ...
\end{aligned} \tag{9}$$

Variables determined outside the system are the observed data sequences $x := (x_1, x_2, ...)$ and $y := (y_1, y_2, ...)$. Variables determined within the system are $\hat{\mu}_x := (\hat{\mu}_{x,0}, \hat{\mu}_{x,1}, ...), \hat{\mu}_y := (\hat{\mu}_{y,0}, \hat{\mu}_{y,1}, ...), \hat{\mu}_{xx} := (\hat{\mu}_{xx,0}, \hat{\mu}_{xx,1}, ...), \hat{\mu}_{xy} := (\hat{\mu}_{xy,0}, \hat{\mu}_{xy,1}, ...), \hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, ...)$, and $\hat{\beta} := (\hat{\beta}_0, \hat{\beta}_1, ...)$. Under mild conditions, $\hat{\alpha}_n$ converges to α^* and $\hat{\beta}_n$ converges to β^*.

We now ask whether this system falls into the PCM. The answer is no, because the PCM requires the dimensions of the background and endogenous variables to be finite. Here these dimensions are countably infinite. The PCM does not apply. (As a referee notes, however, a countably infinite version of the PCM has recently been discussed by Eichler and Didelez, 2007).

In contrast, settable systems encompass this learning system by permitting the settable variables to be of countably infinite dimension. The definitions of direct causality and the notion of partitioning operate identically in either the finite or the countably infinite case. Settable systems generally accommodate any recursive learning algorithm involving data sequences of arbitrary length.

3.7.2 LEARNING WITH A HIDDEN UNIT NETWORK AND ATTRIBUTES

To motivate the next feature of settable systems, we return to considering the effect on an optimal network weight of interventions to distributional parameters, γ, and another network weight. Now, however, we modify the prediction function to be that defined by

$$f(X; \alpha, \beta) = \alpha\, \phi(\beta X).$$

This is a single hidden layer feedforward network with a single hidden unit having the activation function ϕ. For concreteness, let ϕ be the standard normal density. This activation function often appears in radial basis function networks. For clarity, we consider only a single input X, a single

input-to-hidden weight β, and a single hidden-to-output weight α. This elementary structure suffices to make our key points and keeps our notation simple.

Now the expected squared prediction error is

$$L(\alpha,\beta,\gamma;\phi) := E_\gamma([Y-\alpha\,\phi(\beta X)]^2).$$

Here, γ reverts to representing the general parameter indexing F_γ. The choice $\gamma := (\mu_X,\mu_Y,\mu_{XX},\mu_{XY})$ considered above is no longer appropriate, due to the nonlinearity in network output induced by ϕ. Further, note the presence of the hidden unit activation function ϕ in the argument list of L. We make this explicit, as ϕ certainly helps determine prediction performance, and it has a key role to play in our subsequent discussion.

Now consider the clamped optimization problems corresponding to the elementary partition Π^e. This yields solutions

$$\widetilde{\mathbb{A}}^*(\beta,\gamma;\phi) \;:=\; \arg\min_{\alpha\in\mathbb{A}} L(\alpha,\beta,\gamma;\phi),$$
$$\widetilde{\mathbb{B}}^*(\alpha,\gamma;\phi) \;:=\; \arg\min_{\beta\in\mathbb{B}} L(\alpha,\beta,\gamma;\phi).$$

We ensure the existence of compact-valued correspondences $\widetilde{\mathbb{A}}^*(\beta,\gamma;\phi)$ and $\widetilde{\mathbb{B}}^*(\alpha,\gamma;\phi)$ by (among other things) taking $\mathbb{A}$ and $\mathbb{B}$ to be compact subsets of $\mathbb{R}$. Elements $\tilde\alpha^*$ of $\widetilde{\mathbb{A}}^*(\beta,\gamma;\phi)$ and $\tilde\beta^*$ of $\widetilde{\mathbb{B}}^*(\alpha,\gamma;\phi)$ satisfy the necessary first order conditions

$$\begin{aligned} E_\gamma([\phi(\beta X)]^2)\tilde\alpha^* - E_\gamma(\phi(\beta X)Y) &= 0, \\ E_\gamma(D\phi(\tilde\beta^* X)Y) - \alpha E_\gamma[D\phi(\tilde\beta^* X)\phi(\tilde\beta^* X)] &= 0, \end{aligned}$$

where $D\phi$ denotes the first derivative of the ϕ function. We caution that although these relations necessarily hold for elements of $\widetilde{\mathbb{A}}^*(\beta,\gamma;\phi)$ and $\widetilde{\mathbb{B}}^*(\alpha,\gamma;\phi)$, not all (α,β) values jointly satisfying these implicit equations are members of $\widetilde{\mathbb{A}}^*(\beta,\gamma;\phi)$ and $\widetilde{\mathbb{B}}^*(\alpha,\gamma;\phi)$. Some solutions to these equations may be local minima, inflection points, or (local) maxima.

The PCM does not apply here, due (among other things) to the absence of a unique fixed point. Nevertheless, settable systems do apply, using a principled selection of elements from $\widetilde{\mathbb{A}}^*(\beta,\gamma;\phi)$ and $\widetilde{\mathbb{B}}^*(\alpha,\gamma;\phi)$, respectively. We write these selections

$$\tilde\alpha^* = \tilde r_1(\beta,\gamma;\phi), \quad \tilde\beta^* = \tilde r_2(\alpha,\gamma;\phi).$$

The feature distinguishing this example from our earlier examples is the appearance in the response functions of the hidden unit activation function ϕ. The key feature of ϕ is that it takes one and only one value: ϕ is the standard normal density. It is therefore not a variable. Consequently, it cannot be a setting, and it is distinct from any of the other objects we have previously examined. We define an *attribute* to be any object specified a priori that helps determine responses but is not variable. We associate attributes with the system units. Any attribute of the system itself is a *system attribute*; we formally associate system attributes to each system unit. Here, ϕ is a system attribute. Because a unit's associated attributes are constant, they are not subject to counterfactual variation. Nevertheless, attributes may differ across units.

One generally useful attribute is the attribute of *identity*. This is a label assigned to each unit of a given system that can take only the assigned value, and whose value is shared by no other unit

of the system. The identity attribute is required by settable systems, as the identity labels are those explicitly used in the partitioning operation. The identity attribute is also a feature of the PCM, as background and endogenous variables are distinct types of objects, and elements of each distinct type have identifying subscripts.

When attributes beyond identity are present, they need not be distinct across units. For example, the quantity n appearing in several of the response functions in the learning algorithm of Equation (9) is an attribute shared by those units.

We emphasize that attributes are relative to the particular structural system, not somehow absolute. Some objects may be taken as attributes solely for convenience. For example, one might consider several different possible activation functions and attempt to learn the best one for a given problem. In such systems, the hidden unit activation is no longer an attribute but is an endogenous variable. In other cases, it may be more convenient to treat the activation function as hard-wired, in which case the activation function is an attribute. Indeed, any hard-wired aspect of the system is properly an attribute. Convenience may even dictate treating as attributes objects that are in principle variable, but whose degree of variation is small relative to that of other system variables of interest.

Other system aspects are more inherently attributes. Because of their fundamental role and their invariance, such attributes are easily taken for granted and thus overlooked. Our least-squares learning example is a case in point. Specifically, the loss function itself is properly viewed as an attribute.

A useful way to appreciate this is to consider the loss functions

$$L_p(\alpha,\beta,\gamma) := \int |y - f(x;\alpha,\beta)|^p dF_\gamma(x,y), \quad p > 0.$$

In our examples so far, we always take $p = 2$, so $L = L_2$. Different choices are possible, yielding different loss functions. A leading example is the choice $p = 1$. Whereas $p = 2$ yields forecasts that approximate the conditional mean of Y given X, $p = 1$ yields forecasts that approximate the conditional median of Y given X.

Because p is a constant specified a priori and because p helps determine the optimal responses, p is an attribute. When the forecaster's goal is explicitly to provide a forecast based on the conditional mean, it makes little sense to consider values of p other than 2, because no other value of p is guaranteed to generally deliver an approximation to the conditional expectation. Put somewhat differently, it may not make much sense to attempt to endogenize p and choose an "optimal" value of p from some set of admissible values because the result of choosing different values for p is to modify the very goal of the learning exercise. Nor can one escape from attributes by endogenizing p; as long as there is some optimality criterion at work, this criterion is properly an attribute of the system.

Another important example of inherent attributes is provided by the sets $\mathbb{S}_i$ that specify the admissible values taken by the settings $\mathcal{X}_i(1)$ and responses $\mathcal{X}_i(0)$. These are properly specified a priori; they take one and only one value for each unit i; and they play a fundamental role in determining system responses.

Because attributes in settable systems are fixed a priori for a given unit, they take values in a (non-empty) degenerate set. Accordingly, attributes cannot be settings, and thus can never be potential causes, much less direct causes. This formal distinction between attributes and potential causes is unambiguous in settable systems.

3.7.3 ATTRIBUTES IN THE PCM

In contrast, a somewhat ambiguous situation prevails in the PCM. Viewing attributes as a subset of those objects having no causal status, Pearl (2000, p. 98) states that attributes can be treated as elements of u, the background variables.[3] This overlooks the key property we wish to assign to attributes: for a given unit, they are fixed, not variable. Such objects thus cannot belong to u if one takes the word "variable" at face value. In our view, assigning attributes to u misses the opportunity to make an important distinction between invariant aspects of the system units on the one hand and counterfactual variation admissible for the system unit values on the other. Among other things, assigning attributes to u interferes with assigning natural causal roles to structurally exogenous variables.

Further, just as for endogenous and exogenous variables, the PCM does not provide guidance about how to select attributes. In contrast, settable systems clearly identify attributes as invariant features of the system units that embody fundamental aspects of the decision problem represented by the settable system.

Below, we will further distinguish attributes from variables when we discuss stochastic settable systems.

3.8 A Comparative Review of PCM and Settable System Features

At this point, it is helpful to take stock of the features of settable systems that we have so far introduced and contrast these with corresponding features of the PCM.

(1) Settable systems explicitly represent the dual roles of the variables of structural systems using settable variables. These dual roles are present but implicit in the PCM. Settable variables can be responses, or they can be set to given values (settings). The explicit representation of these dual roles in settable systems makes possible implicitly defined structural relations that may not be representable in the PCM. Further, these implicit structural relations may involve correspondences rather than functions. Principled selections from these correspondences yield unique response functions in settable systems.

(2) In settable systems, all variables of the system, structurally exogenous or endogenous, have causal status, in that they can be potential causes or direct causes. Further, no assumptions are made as to the observability of system variables: structurally exogenous variables may be either observable or unobservable; the same is true for structurally endogenous variables. In particular, this permits settable systems to admit unobserved causes and results in causal relations that are not relative to an observer. In contrast, the PCM admits causal status only for endogenous variables. For the PCM, structurally exogenous unobserved causes are ruled out. Although the PCM does permit treating background variables as endogenous variables in alternative systems, it is silent as to how to distinguish between exogenous and endogenous variables. On the other hand, the governing principles in settable systems provide a formal and explicit means for distinguishing between endogenous and exogenous variables.

(3) Settable systems admit straightforward definitions of interventions and direct effects. These notions, while present, are less direct in the formal PCM.

(4) In settable systems, partitioning permits specification of different mutually consistent versions of a given structural system in which different groups of variables are jointly free to respond to the other variables of the system. In particular, system variables can respond either singly or jointly

3. This possibility is also suggested by two referees.

to the other variables of the system, as illustrated by our examples of learning with or without clamping. Similar exercises are possible in the PCM using submodels and the do operator, but the PCM requirement of a unique fixed point limits its applicability. Specifically, we saw that learning with clamping falls outside the PCM. Halpern's (2000) GPCM does apply to such systems, but causal discourse is problematic, due to the absence of the potential response function. In settable systems, fixed points are not required, and causal notions obtain without requiring the potential response function. This permits settable systems to provide causal meaning in our examples of learning with or without clamping.

(5) Settable systems can have a countable infinity of units, whereas the PCM requires a finite number of units.

(6) In settable systems, attributes are a priori constants associated with the units that help determine responses. In the PCM, attributes are not necessarily linked to the system units. Further, they are treated not as constants, but as background variables, resulting in potential ambiguity. The PCM is silent as to how to distinguish between attributes and variables.

Some features of settable systems, such as relaxing the assumption of unique fixed points (point 4) and accommodating an infinity of agents (point 5), are entirely unavailable in the PCM. The remaining settable systems features above rigorously formalize and extend or refine related PCM features and thus permit more explicit causal inference.

4. Stochastic Settable Systems: Heuristics

In the PCM, randomness does not formally appear until definition 7.1.6 (*probabilistic causal model*). Nevertheless, Pearl's (2000) definitions 7.1.1 through 7.1.5 (*causal model*, *submodel*, *effect of action*, *potential response*, and *counterfactual*) explicitly refer to "realizations" *pa* or *x* of endogenous variables *PA* or X. These references make sense only if *PA* and X are interpreted as random vectors. Although *u* is not explicitly called a realization, the language of definition 7.1.1 further suggests that *u* is viewed as a realization of random background variables, U. This becomes explicit in definition 7.1.6, where PCM background variables U become governed by a probability measure P. Randomness of endogenous variables is then induced by their dependence on U. In this sense, definitions 7.1.1 through 7.1.5 do not have fully defined content until definition 7.1.6 resolves the meaning of U, V, PA, and X. Nevertheless, definitions 7.1.1 through 7.1.5 are perfectly meaningful, simply viewing the referenced variables as real numbers.

The settable systems discussed so far are entirely non-stochastic: the settings and responses defined in Section 3 are real numbers, not random variables. Nevertheless, we can connect causal and stochastic structures in settable systems by viewing settings and responses as realizations of random variables, in much the same spirit as the PCM. In this section we discuss some specifics of this connection.

4.1 Introducing Randomness into Settable Systems

First, instead of only the background variables *u* representing realizations of random variables, in settable systems *all* settings represent realizations of random variables governed by an underlying probability measure. For example, in our hidden unit clamped learning example, (α, β, γ) are realizations of random variables (A, B, C) governed by a probability measure P on an underlying measurable space $(\Omega, \mathcal{F})$. The randomness of the responses is induced by their dependence on the

settings. Thus, in the hidden unit clamped learning example, we have random responses

$$\tilde{A}^* = \tilde{r}_1(B,C;\phi), \qquad \tilde{B}^* = \tilde{r}_2(A,C;\phi).$$

Second, the underlying probability measure for settable systems can depend on the attribute vector, call it **a**, of the system. Whereas in the PCM attributes are "lumped together" with other background variables, and may therefore be random, this is not permitted in settable systems. In settable systems, attributes are specified a priori and take one and only one value, **a**. Because of its a priori status, this value is non-random.

It follows that the probability measure governing the settable system can be indexed by **a**. This is not an empty possibility; it has clear practical value. One context in which this practical value arises is when attention focuses only on the units of some subsystem of a larger system. For example, consider the least squares machine learning algorithm of the equations in (9), and focus attention on the subsystem

$$\begin{aligned} \hat{B} &= 1_{\{\hat{M}_{xx}-\hat{M}_x^2>0\}}[\hat{M}_{xx}-\hat{M}_x^2]^{-1}(\hat{M}_{xy}-\hat{M}_x\hat{M}_y), \\ \hat{A} &= \hat{M}_y - \hat{B}\hat{M}_x. \end{aligned}$$

Note that we have modified the notation to reflect the fact that the settings $\hat{M}_x, \hat{M}_y, \hat{M}_{xx}$, and $\hat{M}_{xy}$ are now random variables. These generate realizations $\hat{\mu}_{x,n}, \hat{\mu}_{y,n}, \hat{\mu}_{xx,n}$, and $\hat{\mu}_{xy,n}$ under a probability measure P_n, which is that induced by the probability measure governing the random fundamental settings $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. Note the explicit dependence of the probability measure P_n on the attribute n. The fact that this probability measure can depend on attributes underscores their nature as a priori constants in settable systems.

4.2 Some Formal Properties of Stochastic Settable Systems

Given attributes **a**, we let $(\Omega, \mathcal{F}, P_{\mathbf{a}})$ denote the complete probability space on which the settings and responses are defined. Here, Ω is a set (the "universe") whose elements ω index possible outcomes ("possibilities"); $\mathcal{F}$ is a $\sigma-$field of measurable subsets of Ω whose elements represent events; and $P_{\mathbf{a}}$ is a probability measure (indexed by **a**) on the measurable space $(\Omega, \mathcal{F})$ that assigns a number $P_{\mathbf{a}}(F)$ to each event $F \in \mathcal{F}$. See, for example, White (2001, Chap. 3) for an introductory discussion of measurable spaces and probability spaces.

We decompose ω as $\omega := (\omega_r, \omega_s)$, with $\omega_r \in \Omega_r, \omega_s \in \Omega_s$, so that $\Omega = \Omega_r \times \Omega_s$. As we discuss next, this enables distinct components of ω to underlie responses (ω_r) and settings (ω_s). This facilitates straightforward and rigorous definitions of counterfactuals and interventions. These notions in turn support a definition of direct effect.

To motivate the foundations for defining counterfactuals, again consider the hidden unit clamped learning example. Formally, the random settings (A,B,C) are measurable functions $A : \Omega_s \to \mathbb{R}$, $B : \Omega_s \to \mathbb{R}$, and $C : \Omega_s \to \mathbb{R}^m$, $m \in \mathbb{N}$. Letting ω_s belong to Ω_s, we take the setting values to be the realizations

$$\begin{aligned} \alpha &= A(\omega_s) =: \mathcal{X}_1(\omega,1), \\ \beta &= B(\omega_s) =: \mathcal{X}_2(\omega,1), \\ \gamma &= C(\omega_s) =: \mathcal{X}_0(\omega,1). \end{aligned}$$

Observe that the settings depend only on the ω_s component of ω. We make this explicit in $A(\omega_s)$, $B(\omega_s)$, and $C(\omega_s)$, but leave this implicit in writing $\mathcal{X}_0(\omega,1)$, $\mathcal{X}_1(\omega,1)$, and $\mathcal{X}_2(\omega,1)$ for notational convenience.

The responses are determined similarly:

$$\begin{aligned}\tilde{A}^*(\omega) &= \tilde{r}_1(B(\omega_s),C(\omega_s),\omega_r;\phi) = \tilde{r}_1(\mathcal{X}_2(\omega,1),\mathcal{X}_0(\omega,1),\omega_r;\phi) =: \mathcal{X}_1(\omega,0),\\ \tilde{B}^*(\omega) &= \tilde{r}_2(A(\omega_s),C(\omega_s),\omega_r;\phi) = \tilde{r}_2(\mathcal{X}_1(\omega,1),\mathcal{X}_0(\omega,1),\omega_r;\phi) =: \mathcal{X}_2(\omega,0).\end{aligned}$$

Note that we now make explicit the possibility that the response functions may depend directly on ω_r. This dependence was absent in all our previous examples but is often useful in applications, as this dependence permits responses to embody an aspect of "pure" randomness. From now on, we will include ω_r as an explicit argument of the response functions.

In the deterministic systems previously considered, we viewed the fundamental setting $\mathcal{X}_0(1)$ as a primitive object and adopted the convention that $\mathcal{X}_0(0) := \mathcal{X}_0(1)$. Once settings and responses depend on ω, it becomes necessary to modify our conventions regarding the fundamental settable variables $\mathcal{X}_0$, as $\mathcal{X}_0$ is no longer determined outside the system. The role of the system primitive is now played by ω, the *primary setting*. We represent this as the settable variable defined by $\mathcal{X}_*(\omega,0) := \mathcal{X}_*(\omega,1) := \omega$. We now view $\mathcal{X}_0(\omega,0)$ as a response to ω_s and we take $\mathcal{X}_0(\cdot,1) := \mathcal{X}_0(\cdot,0)$.

In the current stochastic framework, the feature that distinguishes $\mathcal{X}_0$ from other settable variables is that the response $\mathcal{X}_0(\omega,0)$ depends only on ω_s, whereas responses of other settable variables can depend directly on other settings and on ω_r. Given the availability of $\mathcal{X}_*$, there is no guarantee or requirement that such a settable variable $\mathcal{X}_0$ exists. Nevertheless, such *fundamental stochastic settable variables* $\mathcal{X}_0$ are often an important and useful feature in applications, as our machine learning examples demonstrate.

The definition of direct causality in stochastic settable systems is closely parallel to that in the non-stochastic case. Specifically, consider the partition $\Pi = \{\Pi_b\}$, and suppose i belongs to the partition element Π_b. Let $\mathcal{X}_{(b)}(\omega,1)$ denote setting values for the settable variables whose indexes do not belong to Π_b, together with the settings $\mathcal{X}_0(\omega,1)$. Then the response $\mathcal{X}_i(\omega,0)$ is given by

$$\mathcal{X}_i(\omega,0) := r_i(\mathcal{X}_{(b)}(\omega,1),\omega_r;\mathbf{a}) = r_i(z_{(b)},\omega_r;\mathbf{a}),$$

where r_i is the associated response function, $\mathbf{a}$ is the attribute vector, and for convenience we write $z_{(b)} := \mathcal{X}_{(b)}(\omega_s,1)$. Then we say that $\mathcal{X}_j$ *does not directly cause* $\mathcal{X}_i$ if $r_i(z_{(b)},\omega_r;\mathbf{a})$ defines a function constant in the element z_j of $(z_{(b)},\omega_r)$ for all values of the other elements of $(z_{(b)},\omega_r)$. Otherwise, we say that $\mathcal{X}_j$ *directly causes* $\mathcal{X}_i$. Thus, $\mathcal{X}_*$ can directly cause $\mathcal{X}_i$; for this, take $z_j = \omega_r$. If $\mathcal{X}_0(\omega,0)$ does not define a constant function of ω_s, we also say that $\mathcal{X}_*$ directly causes $\mathcal{X}_0$. As always, direct causality is relative to the specified partition.

4.3 Counterfactuals, Interventions, and Direct Effects

We now have the foundation necessary to specify "counterfactuals." We begin by defining what is meant by "factual." Suppose for now that all setting and response values apart from ω_r are observable. Specifically, suppose we have realizations of setting values (β,γ) and response value $\tilde{\alpha}^* = \tilde{r}_1(\beta,\gamma,\omega_r;\phi)$, and that ω is such that $\beta = B(\omega_s)$, $\gamma = C(\omega_s)$, and $\tilde{\alpha}^* = \tilde{A}^*(\omega)$, where

$$\tilde{A}^*(\omega) = \tilde{r}_1(B(\omega_s),C(\omega_s),\omega_r;\phi).$$

Then we say that $(\tilde{\alpha}^*, \beta, \gamma)$ are *factual* and that $\omega = (\omega_r, \omega_s)$ is *factual*. Otherwise, we say that $(\tilde{\alpha}^*, \beta, \gamma)$ and ω are *counterfactual*. Specifically, if the realization $(\tilde{\alpha}^*, \beta, \gamma)$ does not obtain, then we say that $(\tilde{\alpha}^*, \beta, \gamma)$ is counterfactual, whereas if we have the realizations $(\tilde{\alpha}^*, \beta, \gamma)$, but ω is such that $\beta \neq B(\omega_s)$, $\gamma \neq C(\omega_s)$, or $\tilde{\alpha}^* \neq \tilde{A}^*(\omega)$ then we say that ω is counterfactual.

There need not be a unique factual ω since it is possible that multiple ω's yield the same realizations of random variables; this creates no logical or conceptual difficulties. Also, we need not observe all settings and responses; an observable subset of these may be factual or counterfactual. To the extent that a given ω generates realizations compatible with factual observables, it may also be viewed as factual to that degree. An ω generating realizations incompatible with factual observables is necessarily counterfactual.

In non-stochastic settable systems, we defined an intervention to a settable variable as a pair of distinct admissible setting values for that settable variable. For example, $\alpha_1 \to \alpha_2 := (\alpha_1, \alpha_2)$. In stochastic settable systems, we express interventions similarly. Specifically, again consider the partition $\Pi = \{\Pi_b\}$, and suppose i belongs to the partition element Π_b, so that

$$\mathcal{X}_i(\omega, 0) := r_i(\mathcal{X}_{(b)}(\omega, 1), \omega_r; \mathbf{a}) = r_i(z_{(b)}, \omega_r; \mathbf{a}).$$

Then an *intervention* $(z_{(b),1}, \omega_{r,1}) \to (z_{(b),2}, \omega_{r,2})$ is a pair $((z_{(b),1}, \omega_{r,1}), (z_{(b),2}, \omega_{r,2}))$ whose elements are admissible and distinct.

Interventions necessarily involve counterfactuals: at most, only one setting can be factual; and for an intervention to be well defined, the other setting value must be distinct. We note that the notion of counterfactuals is helpful mainly for describing interventions. Although our definitions of causality, interventions, or, as we see next, direct effects implicitly involve counterfactuals, they do not formally require this notion.

The *direct effect on $\mathcal{X}_i$ of the intervention* $(z_{(b),1}, \omega_{r,1}) \to (z_{(b),2}, \omega_{r,2})$ is the associated response difference

$$r_i(z_{(b),2}, \omega_{r,2}; \mathbf{a}) - r_i(z_{(b),1}, \omega_{r,1}; \mathbf{a}).$$

Our definitions of interventions and direct effects permit *ceteris paribus* interventions and direct effects. For these, only some finite number (e.g., one) of the settings differs between $(z_{(b),1}, \omega_{r,1})$ and $(z_{(b),2}, \omega_{r,2})$; the other elements are "held constant."

Under suitable conditions (specifically, that the settings $\mathcal{X}_{(b)}(\cdot, 1)$ are an "onto" function), the interventions $(z_{(b),1}, \omega_{r,1}) \to (z_{(b),2}, \omega_{r,2})$ can be equivalently represented as a *primary intervention*

$$\omega_1 \to \omega_2 := (\omega_1, \omega_2) = ((\omega_{r,1}, \omega_{s,1}), (\omega_{r,2}, \omega_{s,2})).$$

That is, primary interventions are pairs (ω_1, ω_2) of elements of Ω. This representation is ensured by specifying that $\omega = (\omega_r, \omega_s)$, permitting ω_r and ω_s to be variation free (i.e., ω_r can vary without inducing any necessary variation in ω_s, and vice versa).

Primary interventions yield a definition of *total effect* as a response difference. In our hidden unit clamped learning example, the total effect on $\mathcal{X}_1$ of $\omega_1 \to \omega_2$ is

$$\begin{aligned} \Delta\mathcal{X}_1(\omega_1, \omega_2, 0) \ &: = \mathcal{X}_1(\omega_2, 0) - \mathcal{X}_1(\omega_1, 0) \\ &= \ \tilde{r}_1(B(\omega_{s,2}), C(\omega_{s,2}), \omega_{r,2}; \phi) - \tilde{r}_1(B(\omega_{s,1}), C(\omega_{s,1}), \omega_{r,1}; \phi). \end{aligned}$$

This is also the direct effect on $\mathcal{X}_1$ of $(Z_{(1)}(\omega_{s,1}), \omega_{r,1}) \to (Z_{(1)}(\omega_{s,2}), \omega_{r,2})$. We emphasize that these effects are, as always, relative to the governing partition. The total effect above is relative to the elementary partition, corresponding to clamped learning.

4.4 Review of Stochastic Settable System Features

In stochastic settable systems, all settings are governed by the underlying probability measure, whereas in the PCM, only the background variables are subject to random variation. Because of their status as a priori constants, attributes can index the settable system probability measure. Stochastic settable systems distinguish between primary settings and, when they exist, fundamental settable variables. Responses may contain an element of pure randomness. The structure of stochastic settable systems also supports straightforward rigorous definitions of direct causes, counterfactuals, interventions, direct effects, and total effects.

5. Stochastic Settable Systems: A Formal Definition

In Sections 3 and 4, we motivated the features of settable systems using a series of closely related machine learning examples. Here we integrate these features to provide a rigorous formal definition of a stochastic settable system $\mathcal{S}^{\Pi} := \{(\mathbf{A},\mathbf{a}),(\Omega,\mathcal{F},P_{\mathbf{a}}),(\Pi,\mathcal{X}^{\Pi})\}$.

To give a concise definition, we first introduce some convenient notation. We write the positive integers as $\mathbb{N}^+$; we also write $\bar{\mathbb{N}}^+ = \mathbb{N}^+ \cup \{\infty\}$. When $n = \infty$, we interpret $i = 1,...,n$ as $i = 1,2,....$ We also write $\mathbb{N} := \{0\} \cup \mathbb{N}^+$, and $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$. When $m = 0$, we interpret $k = 1,...,m$ as being omitted; thus, when $m = 0$, terms like $\times_{k=1}^{m} A$ or $\times_{k=1}^{m} \mathbb{S}_{0,k}$ are ignored. The notation $\#\Pi$ denotes the number of elements (the cardinality) of the set Π.

Definition 1 (Stochastic Settable System) *Let $n \in \bar{\mathbb{N}}^+$, and let the* unit attribute space *A be a non-empty set. For each* unit *$i = 1,...,n$, let a* unit attribute *a_i be a fixed element of A, such that a_i includes a component of* admissible settings *$\mathbb{S}_i$, a multi-element Borel-measurable subset of $\mathbb{R}$.*

Let $m \in \bar{\mathbb{N}}$. For each $k = 1,...,m$, let a fundamental unit attribute *$a_{0,k}$ be a fixed element of A, such that $a_{0,k}$ includes a component of* admissible fundamental settings *$\mathbb{S}_{0,k}$, a multi-element Borel-measurable subset of $\mathbb{R}$. Write $a_0 := (a_{0,1},...,a_{0,m})$ and $\mathbf{a} := (a_0,a_1,...,a_n) \in \mathbf{A} := (\times_{k=1}^{m} A) \times (\times_{i=1}^{n} A)$, the* joint attribute space.

Let $(\Omega_r,\mathcal{F}_r)$ and $(\Omega_s,\mathcal{F}_s)$ be measurable spaces such that Ω_r and Ω_s each contain at least two elements, and let $(\Omega,\mathcal{F},P_{\mathbf{a}})$ be a complete probability space, where $\Omega := \Omega_r \times \Omega_s$, $\mathcal{F} := \mathcal{F}_r \otimes \mathcal{F}_s$, and $P_{\mathbf{a}}$ is a probability measure indexed by $\mathbf{a} \in \mathbf{A}$.

For each $k = 1,...,m$, let a fundamental response *$Y_{0,k} : \Omega_s \to \mathbb{S}_{0,k}$ be a measurable function and let the corresponding* fundamental setting *be $Z_{0,k} := Y_{0,k}$. Write fundamental settings and responses as $Z_0 := (Z_{0,1},...,Z_{0,m})$ and $Y_0 = Z_0$.*

Let $\Pi = \{\Pi_b\}$ be a partition of $\{1,...,n\}$, with $B := \#\Pi \in \bar{\mathbb{N}}^+$, let $\ell_b := \#\Pi_b$, and let $\mathbf{a}$ determine the multi-element Borel measurable set $\mathbb{S}^{\Pi}_{(b)}(\mathbf{a}) \subset \times_{j \notin \Pi_b} \mathbb{S}_j \times_{k=1}^{m} \mathbb{S}_{0,k}$, $b = 1,...,B$. Suppose there exist measurable functions called settings, *$Z_i^{\Pi} : \Omega_s \to \mathbb{S}_i, i = 1,\ldots,n$, measurable functions called* responses, *$Y_i^{\Pi} : \Omega \to \mathbb{S}_i, i = 1,\ldots,n$, and measurable functions called* joint response functions,

$$r^{\Pi}_{[b]}(\,\cdot\,;\mathbf{a}) : \times_{i \in \Pi_b} \mathbb{S}_i \times \mathbb{S}^{\Pi}_{(b)}(\mathbf{a}) \times \Omega_r \to \mathbb{R}^{\ell_b} \quad b = 1,...,B,$$

such that

$$r^{\Pi}_{[b]}(Y^{\Pi}_{[b]}(\omega), Z^{\Pi}_{(b)}(\omega_s), \omega_r; \mathbf{a}) = \mathbf{0}, \;\; b = 1,...,B,$$

for each $\omega := (\omega_r,\omega_s) \in \Omega_r \times \Omega^{\Pi}_{(b)}(\mathbf{a})$, $\Omega^{\Pi}_{(b)}(\mathbf{a}) := \{\omega_s : Z^{\Pi}_{(b)}(\omega_s) \in \mathbb{S}^{\Pi}_{(b)}(\mathbf{a})\}$, where $Z^{\Pi}_{(b)}$ is the vector containing Z^{Π}_j, $j \notin \Pi_b$ and $Y^{\Pi}_{[b]}$ is the vector containing $Y^{\Pi}_i, i \in \Pi_b$. Write

$$\mathcal{X}^{\Pi}_0(\omega;0) \;:= Y_0(\omega_s), \quad \mathcal{X}^{\Pi}_0(\omega;1) := Z_0(\omega_s),$$

$$\mathcal{X}_i^{\Pi}(\omega;0) \ := Y_i^{\Pi}(\omega), \quad \mathcal{X}_i^{\Pi}(\omega;1) := Z_i^{\Pi}(\omega_s),\ i=1,...,n,$$

so that the fundamental settable variables $\mathcal{X}_0^{\Pi}$ *and* settable variables $\mathcal{X}_i^{\Pi}$, $i=1,...,n$ *are mappings such that:*

$$\mathcal{X}_0^{\Pi}:\Omega\times\{0,1\}\to\times_{k=1}^{m}\mathbb{S}_{0,k} \text{ and } \mathcal{X}_i^{\Pi}:\Omega\times\{0,1\}\to\mathbb{S}_i,\ i=1,...,n.$$

Finally, write

$$\mathcal{X}^{\Pi}:=(\mathcal{X}_0^{\Pi},\mathcal{X}_1^{\Pi},,...,\mathcal{X}_n^{\Pi}).$$

Then $\mathcal{S}^{\Pi}:=\{(\mathbf{A},\mathbf{a}),(\Omega,\mathcal{F},P_{\mathbf{a}}),(\Pi,\mathcal{X}^{\Pi})\}$ *is a* stochastic settable system.

A stochastic settable system consists of *units* $i=1,...,n$ with *unit attributes* a_i belonging to *unit attribute space* A. When $m>0$, the system has optional *fundamental units* $k=1,...,m$ with fundamental unit attributes $a_{0,k}$ also belonging to A. The *joint attributes* $\mathbf{a}:=(a_{0,1},...,a_{0,m},a_1,...,a_n)$ belong to the *joint attribute space* $\mathbf{A}$. By construction, the unit attributes include the *admissible settings* $\mathbb{S}_i$ for each unit ($\mathbb{S}_{0,k}$ for any fundamental units). $\mathbb{S}_i$ must contain at least two values, necessary to ensuring that interventions are well defined.

The probability space $(\Omega,\mathcal{F},P_{\mathbf{a}})$ embodies the stochastic structure of the system. By representing the *primary settings* as elements $\omega:=(\omega_r,\omega_s)$ of $\Omega:=\Omega_r\times\Omega_s$, we provide explicit means for variation-free interventions to primary setting values ω_r and the remaining setting values (via ω_s). By "variation-free", we mean that we can consider interventions $(\omega_1,\omega_2)=((\omega_{r,1},\omega_s),(\omega_{r,2},\omega_s))$ in which only the ω_r component differs or interventions $(\omega_1,\omega_2)=((\omega_r,\omega_{s,1}),(\omega_r,\omega_{s,2}))$ in which only the ω_s component differs. Requiring that Ω_r and Ω_s each have at least two elements ensures that interventions to ω_r and ω_s are well defined.

The probability measure $P_{\mathbf{a}}$ is indexed by the attributes $\mathbf{a}$ and governs the joint distribution of the random settings and responses. $P_{\mathbf{a}}$ may be determined by nature, determined by a researcher, or determined in part by nature and in part by a researcher. $P_{\mathbf{a}}$ can embody any probabilistically meaningful dependence or independence for events involving ω_r and ω_s. Completeness of the probability space is a technical requirement ensuring that the collection of events $\mathcal{F}$ contains every subset of any event having $P_{\mathbf{a}}-$probability zero.

We call the random variables $Z_i^{\Pi}(\cdot)$ *settings* and realizations $Z_i^{\Pi}(\omega_s)$ *setting values*. By suitably choosing Ω_s and Z_i^{Π}, we also achieve variation-free interventions for the individual setting values. Specifically, let $\Omega_s:=(\times_{k=1}^{m}\mathbb{S}_{0,k})\times(\times_{i=1}^{n}\mathbb{S}_i)$, so that Ω_s has typical element $\omega_s:=(z_{0,1},...,z_{0,m},z_1,...,z_n)$. Further, let $Z_{0,k}(\omega_s)=z_{0,k}$ and $Z_i^{\Pi}(\omega_s)=z_i$ be the projection functions that select the specified component of ω_s. By construction, these functions are surjective (onto). That is, the range $Z_i^{\Pi}(\Omega_s)$ equals the co-domain $\mathbb{S}_i$, so that there is (at least) one ω_s corresponding to each admissible value in $\mathbb{S}_i$. With a suitable choice of $\mathcal{F}_s$ (e.g., that generated by the measurable finite dimensional product cylinders), these choices for $Z_{0,k}$ and Z_i^{Π} are also measurable, as required. (White (2001, Section 3.3) provides relevant background and discussion.) Thus, different values $\omega_{s,1}$ and $\omega_{s,2}$ can generate interventions referencing just a single settable variable, so that $Z_i^{\Pi}(\omega_{s,1})\neq Z_i^{\Pi}(\omega_{s,2})$, but $Z_j^{\Pi}(\omega_{s,1})=Z_j^{\Pi}(\omega_{s,2})$ for $j\neq i$. Further, when surjectivity holds, it ensures that the primary interventions $(\omega_{s,1},\omega_{s,2})$ can represent every admissible intervention to the setting values.

When the system has fundamental units, these units have *fundamental responses* $Y_{0,k}$; these are random variables whose values $Y_{0,k}(\omega_s)$ are determined solely by $\omega_s\in\Omega_s$. By convention, *fundamental settings* $Z_{0,k}$ are random variables identical to $Y_{0,k}$. When $Y_{0,k}$ is surjective, then so is $Z_{0,k}$.

Each element Π_b of the partition $\Pi = \{\Pi_b\}$ identifies a group of (non-fundamental) units. The *joint response function* $r^{\Pi}_{[b]}$ specifies how these identified units jointly and freely respond to given jointly admissible setting values of all units not belonging to Π_b.

The given values are setting values $Z^{\Pi}_j(\omega_s)$ for j not belonging to Π_b, including $Z_0(\omega_s)$ and ω_r, represented here by $(Z^{\Pi}_{(b)}(\omega_s), \omega_r)$. The values $Z^{\Pi}_{(b)}(\omega_s)$ belong to the set of *jointly admissible setting values* $\mathbb{S}^{\Pi}_{(b)}(\mathbf{a})$, a subset of $\times_{j\notin\Pi_b}\mathbb{S}_j \times^m_{k=1}\mathbb{S}_{0,k}$. In the absence of constraints, we have $\mathbb{S}^{\Pi}_{(b)}(\mathbf{a}) = \times_{j\notin\Pi_b}\mathbb{S}_j \times^m_{k=1}\mathbb{S}_{0,k}$. Often, however, applications place joint restrictions on the admissible setting values. For example, when the settings represent probabilities (as in the mixed strategy games considered shortly), the constraint that probabilities add to one jointly restricts admissible setting values. The constraints are encoded in $\mathbf{a}$, and implemented by $\mathbb{S}^{\Pi}_{(b)}(\mathbf{a})$.

The *response values* are $Y^{\Pi}_{[b]}(\omega)$, the vector containing unit i's response value $Y^{\Pi}_i(\omega)$ for each i in Π_b, satisfying

$$r^{\Pi}_{[b]}(Y^{\Pi}_{[b]}(\omega), Z^{\Pi}_{(b)}(\omega_s), \omega_r; \mathbf{a}) = \mathbf{0}. \tag{10}$$

Note that we do not explicitly require that $Y^{\Pi}_{[b]}(\omega)$ is the unique solution to the equations in (10). As discussed in our machine learning examples, the governing principles of the system (e.g., optimization and/or equilibrium) operate to deliver a selected system response satisfying these equations. By including the governing principles (including appropriate selection operators) among the attributes $\mathbf{a}$, as is fully rigorous and proper, the presence of $\mathbf{a}$ in the response function can ensure a unique response value. Note that the response function depends on the full system attribute vector $\mathbf{a}$, not just the attributes associated with the units of the given block b. This has been a common feature of our examples. We call the random variables $Y^{\Pi}_i(\cdot)$ *responses.*

Our expression for the responses is in implicit form, as is appropriate for solutions of optimization problems. Nevertheless, it is often convenient to abuse notation somewhat and write response values explicitly as

$$Y^{\Pi}_{[b]}(\omega) = r^{\Pi}_{[b]}(Z^{\Pi}_{(b)}(\omega_s), \omega_r; \mathbf{a}).$$

Because the partition is exhaustive, the collection of response functions $r^{\Pi} := (r^{\Pi}_{[1]}, ..., r^{\Pi}_{[B]})$ provides a description of how each unit in the system responds when it is free to do so in the company of other specified freely responding units. In given circumstances, it may be that only one of these sets of responses is factual; the others are then counterfactual.

Settable variables $\mathcal{X}^{\Pi}_i : \Omega \times \{0,1\} \to \mathbb{S}_i$ embody the dual aspects of settings and responses. Responses $\mathcal{X}^{\Pi}_i(\,\cdot\,,0) := Y^{\Pi}_i$ are random variables taking values in $\mathbb{S}_i$ in response to settings of other settable variables of the system outside the block to which i belongs, say Π_b. The settings $\mathcal{X}^{\Pi}_i(\,\cdot\,,1) := Z^{\Pi}_i$ are random variables taking values in $\mathbb{S}_i$ whose realized values determine the realized responses of other settable variables. The optional fundamental settable variables $\mathcal{X}^{\Pi}_0 : \Omega \times \{0,1\} \to \times^m_{k=1}\mathbb{S}_{0,k}$ yield identical random responses and settings whose values drive responses of other settable variables. We collect together all settable variables of the system by writing $\mathcal{X}^{\Pi} := (\mathcal{X}^{\Pi}_{0,1}, ..., \mathcal{X}^{\Pi}_{0,m}, \mathcal{X}^{\Pi}_1, ..., \mathcal{X}^{\Pi}_n)$. Observe that $\mathcal{X}^{\Pi}$ actually depends on $\mathbf{a}$ through the response functions r^{Π}, so it would be formally correct and more explicit to write $\mathcal{X}^{\Pi}_{\mathbf{a}}$ instead of $\mathcal{X}^{\Pi}$. We forego this for notational simplicity, but this dependence should not be overlooked.

Our notation for the stochastic settable system, $\mathcal{S}^{\Pi} := \{(\mathbf{A}, \mathbf{a}), (\Omega, \mathcal{F}, P_{\mathbf{a}}), (\Pi, \mathcal{X}^{\Pi})\}$, references each component of the system in a way that expresses the hierarchy of these components. At the lowest level is the *attribute structure*, $(\mathbf{A}, \mathbf{a})$; next comes the *stochastic structure*, $(\Omega, \mathcal{F}, P_{\mathbf{a}})$; resting on these is the *causal structure*, $(\Pi, \mathcal{X}^{\Pi})$.

6. Game Theory, Settable Systems, and the PCM

So far, our machine learning examples have shown how settable systems apply to decision problems where optimization operates as a governing principle. We now discuss examples showing how settable systems apply to groups of interacting and strategically competitive decision-making agents. In economics, these agents are usually viewed as consumers, firms, and/or government entities. Of direct relevance to machine learning is that agents may also be artificial intelligences, as in automated trading systems. In addition to their empirical relevance (e.g., the analysis of FCC spectrum auctions or U.S. Treasury Bill auctions), such environments present the opportunity for emergent and distributed computation of otherwise difficult to compute quantities, like prices.

Game theory, the study of multi-agent decision problems, provides a rich formal framework in which to understand and explain the behavior of interacting decision makers. Gibbons (1992) provides an excellent introduction. By showing how the structures of game theory map to settable systems, we establish the foundations for causal analysis of such systems. A central feature of such structures is that their outcomes are determined by suitable equilibrium mechanisms, specifically *Nash equilibrium* and its refinements. Among other things, these mechanisms play a key role in ensuring the mutual consistency of various partitions relevant to the analysis of a given game.

6.1 Pure-Strategy Games and Pure-Strategy Nash Equilibria

The simplest games are static games of complete information (Gibbons, 1992, Chap. 1). In these games, each of n players has: (i) a number of playable strategies (let player i have K_i playable strategies, $s_{i,1},...,s_{i,K_i}$); and (ii) a utility (or "payoff") function u_i that describes the payoff π_i to that player when each player plays one of their given strategies. That is, $\pi_i = u_i(s_1,...,s_n)$, where $s_j \in S_j := \{s_{j,1},...,s_{j,K_j}\}$, $j = 1,...,n$. The players simultaneously choose their strategies; then each receives the payoff specified by the collection of the jointly chosen strategies and the players' payoff functions. Such games are "static" because of the simultaneity of choice. They are "complete information" games because the players' possible strategies and payoff functions are known to all players. (Thus, each player can assess the game from every other player's viewpoint.) An n-player static game of complete information is formally represented in "normal form" as $\mathcal{G} = \{S_1,...,S_n; u_1,...,u_n\}$.

These games map directly and explicitly to the settable system framework. Specifically, the players correspond to units $i = 1,...,n$. The unit attributes a_i include the identity attribute i, the strategies $\mathbb{S}_i = S_i$ available to player i, and the player's utility function $u_i : S_1 \times ... \times S_n \to \mathbb{R}$. When a strategy for player i is set arbitrarily, we denote its value as $z_i \in S_i$; when player i chooses a strategy (a response) we represent its value as $y_i \in S_i$. For concreteness and without loss of generality, we take $S_i := \{1,...,K_i\}$. The players' utility functions implicitly account for the possibility that strategy 1 for player i may represent a different action than that of strategy 1 for player j.

Each player seeks to maximize their payoff given the strategies of the others, so that

$$y_i = r_i^e(z_{(i)};\mathbf{a}) = \arg\max_{z_i \in S_i} u_i(z_1,...,z_n).$$

In economics, this goal-seeking behavior is called "rationality;" an equally fitting (or perhaps superior) term is "intelligence." Thus, game theory analyzes rational or intelligent agents.

Here, we write the responses r_i^e using the superscript e to denote that these response are those for the elementary partition, $\Pi^e := \{\Pi_1^e,...,\Pi_n^e\}$ with $\Pi_i^e = \{i\}$, as each response takes the strategies of all other players as fixed.

For convenience, we assume that for each player there is a unique utility-maximizing response, but just as in our previous machine learning examples, we can generally make a principled selection when the optimal decision is a set. Below, we discuss this further.

In game theory, r_i^e is called a "best-response" function. In settable systems, we refer generically to functions like r_i^e as "response" functions, in part motivated by this usage. Because in game theory the specific game $\mathcal{G}$ under consideration is almost always clear, there is usually no need to explicitly reflect its elements in the players' best response functions. The explicit appearance of players' joint attributes $\mathbf{a}$ (which characterize the game) in the response functions $r_i^e(\,\cdot\,;\mathbf{a})$ emphasizes their role in determining player responses.

Now consider the PCM representation of this game. In the PCM, the attributes become background variables u. The attributes $a_i = (i, S_i, u_i)$ do not map directly to PCM background variables, as S_i is a set and u_i is a function; the PCM requires the background variables to be real numbers. Nevertheless, some simple modifications deliver a conforming representation: we can replace S_i with the integers 1 through K_i and u_i with the vector of values taken by $\pi_i = u_i(s_1,...,s_n)$ as the strategies range over all possible values. We collect these values together across all players and write them as u. Endogenous variables $v = \{s_1,...,s_n\}$ represent player strategies, and the structural functions $f = \{f_1,...,f_n\}$ represent the best response for agent i as $s_i = f_i(s_{(i)},u), \quad i = 1,...,n$.

The final condition required by the PCM is that there exists a unique fixed point, defined by functions g_i such that $s_i = s_i^* := g_i(u), i = 1,...,n$. When such a unique fixed point exists, it represents a *pure-strategy Nash equilibrium* (Nash, 1950). By definition this satisfies

$$u_i(s_1^*,...,s_i^*,...,s_n^*) \geq u_i(s_1^*,...,s_i,...,s_n^*) \;\; for\ all\ s_i \in S_i, i = 1,...,n.$$

Gibbons (1992, p. 8) provides further discussion of pure-strategy Nash equilibrium.

Just as we saw in Section 3, the PCM faces difficulties arising from its requirement of a unique fixed point. A first difficulty for the PCM is that there are important games for which a pure-strategy Nash equilibrium does not exist; the PCM therefore has nothing to say about such games. A leading example of such games is known as *matching pennies* (Gibbons, 1992, p. 29). In this game, each of two players has a penny that they can choose to display face up (heads) or face down (tails). If the pennies match, player 2 gets both; otherwise player 1 gets both. This game applies to any situation in which one player would like to outguess the other, as in *poker* (bluffing), *baseball* (pitcher vs. hitter), and *battle*.

Given the interest attaching to such games, one would like to have an applicable causal model. This need is met by the settable system framework. Because this framework imposes no fixed point requirement, it applies regardless of the existence of a unique pure-strategy Nash equilibrium. For games with no pure-strategy Nash equilibrium, the response functions $r_i^e(z_{(i)};\mathbf{a})$ of the elementary partition $\Pi^e := \{\{1\},...,\{n\}\}$ readily provide complete information about the best response for all counterfactual strategy combinations of the other players.

If a unique pure-strategy Nash equilibrium exists, it has settable system representation

$$s_i^* = r_i^g(\mathbf{a}), \;\; i = 1,...,n,$$

where r_i^g is the response function for the global partition, $\Pi^g := \{\{1,...,n\}\}$. An interesting feature of these response functions is that they depend only on the attributes $\mathbf{a}$; no fundamental or even primary settings appear. Observe also that the Nash equilibrium condition ensures the mutual consistency of the elementary and global partitions.

When there is no pure strategy Nash equilibrium, as in the matching pennies game, there need not exist a valid settable system for the global partition. This provides an interesting example in which we have a well-defined settable system for the elementary partition, but not for the global partition. In contrast, the PCM does not apply at all.

Another difficulty for the PCM is that the unique fixed point requirement prevents it from applying to games with multiple pure-strategy Nash equilibria. An example is the game known as *battle of the sexes* (Gibbons, 1992, p. 11). In this game, two players (Ralph and Alice) are trying to decide on what to do on their next night out: attend a boxing match or attend an opera. Each would rather spend the evening together than apart, but Ralph prefers boxing and Alice prefers the opera. With the payoffs suitably arranged (symmetric), there is a unique best response for each player, given the strategy of the other. Nevertheless, this game has two pure-strategy Nash equilibria: (i) both select boxing; (ii) both select the opera. Thus, the PCM does not apply.

In contrast, the settable system framework does apply, as it does not impose a unique fixed point requirement. The elementary partition describes each agent's unique best response to a given strategy of the other. Further, when multiple Nash equilibria exist, the global partition can yield a well-defined settable system by selecting one of the possible equilibria. As Gibbons (1992, p. 12) notes, "In some games with multiple Nash equilibria one equilibrium stands out as the compelling solution to the game," leading to the development of "conventions" that provide standard means for selecting a unique equilibrium from the available possibilities.

An example is the classic *coordination game*, in which there are two pure-strategy Nash equilibria, but one yields greater payoffs to both players. The convention is to select the higher payoff equilibrium. If such a convention exists, the global partition can specify the response functions r_i^g to deliver this. In such cases, the global partition responses satisfy not only a fixed-point property, but also embody equilibrium selection.

Interestingly, battle of the sexes is not a game with such a convention, as both equilibria seem equally compelling. A more elaborate version of this game, involving incomplete information, does possess a unique equilibrium, however (Gibbons, 1992, pp. 152-154).

6.2 Mixed-Strategy Games and Mixed-Strategy Nash Equilibria

As just suggested, one can modify the character of a game's equilibrium set by elaborating the game. Specifically, consider "mixed-strategy" static games of complete information. Instead of optimally choosing a pure strategy, each player h now chooses a vector of probabilities $p_h := (p_{h,1}, ..., p_{h,K_h})$ (a mixed strategy) over their available pure strategies, say $S_h := \{1, ..., K_h\}$, so that $p_{h,j}$ is the probability that player h plays strategy $j \in S_h$. For example, the probability vector $(1, 0..., 0)$ for player h represents playing the pure strategy $s_h = 1$.

Note that we have modified the notation for the player index from i to h. This enables us to continue to index units using i. Here, the units i correspond to *agent-decision pairs* (h, j). The values h and j become part of the unit attributes, a_i. When referencing i, we may for convenience reference the corresponding h, j, so, for example, we may write a_i or $a_{h,j}$, whichever is more convenient.

Each player h now behaves rationally or intelligently by choosing mixed-strategy probabilities to maximize their expected payoff given other players' strategies,

$$\bar{\pi}_h = \upsilon_h(p^n) := \sum_{s^n \in S^n} u_h(s^n) \Pr(s^n; p^n),$$

where for conciseness we now write $s^n := (s_1, ..., s_n)$, $S^n := S_1 \times ... \times S_n$, and $p^n := (p_1, ..., p_n)$. (Maximizing expected payoff is not the only possibility, but we focus on this case for concreteness.) The strategies are chosen independently, so that $\Pr(s^n; p^n)$, the probability that the agents jointly choose the configuration of strategies s^n, is given by $\Pr(s^n; p^n) = \prod_{h=1}^{n} p_{h,s_h}$.

It is a famous theorem of Nash (1950) that if n is finite and if K_h is finite, $h = 1, ..., n$, then there must exist at least one Nash equilibrium for $\mathcal{G}$, possibly involving mixed strategies (e.g., Gibbons, 1992, p. 45).

We map mixed-strategy games to settable systems as follows. As mentioned above, units i are agent-decision pairs (h, j), so that unit attributes a_i include the agent and decision designators, h and j. Because settings and responses are now probabilities, unit attributes also specify admissible settings $\mathbb{S}_{h,j}$ as a subset of $[0, 1]$. We further discuss $a_{h,j}$ below.

For each agent h, there is a $K_h \times 1$ vector of settings and responses. We denote the probabilities of the mixed strategy for agent h as $z_{h,j}, j = 1, ..., K_h$, when these are set, and as $y_{h,j}, j = 1, ..., K_h$, when these constitute the agent's best response. Let z_h be the $K_h \times 1$ vector with elements $z_{h,j}$, and let y_h be the $K_h \times 1$ vector with elements $y_{h,j}$. Given all other player's mixed strategies $z_{(h)}$, agent h's best response is

$$y_h = r_h^a(z_{(h)}; \mathbf{a}) = \sigma_h(\arg\max_{z_h \in \mathbf{S}_h} \upsilon_h(z_1, ..., z_n)),$$

where the maximization is taken over the simplex $\mathbf{S}_h := \{z \in [0, 1]^{K_h} : \sum_{j=1}^{K_h} z_j = 1\}$. The operator σ_h performs a measurable selection, discussed below.

Several aspects of this representation are notable. First, we write the response function r_h^a to denote that it is the response function for the *agent partition* $\Pi^a := \{\Pi_h^a, h = 1, ..., n\}$, where $\Pi_h^a = \{(h, 1), ..., (h, K_h)\}$. In contrast, the elementary partition is $\Pi^e := \{\Pi_{h,j}^e, j = 1, ..., K_h; h = 1, ..., n\}$, with $\Pi_{h,j}^e := \{(h, j)\}$. The response functions $r_{h,j}^e$ for the elementary partition describe the best response for agent h's strategy j given not only all other agents' strategies, but also all other strategies for agent h. The elementary partition is usually not of particular interest; the agent partition and the global partition are typically the main objects of interest in this context.

The superscript ^{a} in r_h^a and elsewhere to denote the agent partition creates no confusion with the joint attributes $\mathbf{a}$, as the former is always a superscript, and the latter never is.

Next, we note that the unit attributes $a_{h,j}$ contain admissible values $\mathbb{S}_{h,j} \subset [0, 1]$, so that $0 \leq z_{h,j} \leq 1$. This is not enough to fully specify the admissible values for the vector z_h, however, as the probabilities must add up to 1. This means that z_h must belong to the simplex $\mathbf{S}_h$. We enforce this constraint by making $\mathbf{S}_h$ a component of each unit attribute $a_{h,j}, j = 1, ..., K_h$. Just as an attribute common to all system units is a system attribute, any attribute common to a given subset of units is an attribute of that subset. Thus, $\mathbf{S}_h$ is an attribute of agent h; agent h is that subset of the units with agent designator equal to h.

An interesting feature of mixed-strategy games is that the set $\arg\max_{z_h \in \mathbf{S}_h} \upsilon_h(z_1, ..., z_n)$ can easily fail to have a unique element. This set thus defines the player's best-response correspondence, rather than simply giving a best-response function. We obtain a best-response function by applying a measurable selection operator σ_h to the set of maximizers. The operator σ_h is an attribute, specifically of agent h; thus, we include it as a component of the unit attributes $a_{h,j}, j = 1, ..., K_h$.

By definition, the agent is indifferent between elements of the arg-max set; the choice of selection operator is not crucial. In fact, the selection may be random, implemented by letting σ_h depend

on $\omega_r \in \Omega_r$, so that one has

$$y_h = r_h^a(z_{(h)}, \omega_r; \mathbf{a}) = \sigma_h(\arg\max_{z_h \in \mathbf{S}_h} \upsilon_h(z_1, ..., z_n), \omega_r).$$

Now consider how this game maps to the PCM. Again, the attributes map to the background variables u, although now the attributes are, among other things, sets with a continuum of values and correspondences. Mapping these to a vector of real numbers is problematic, so we simply view u as a general vector whose elements may be numbers, sets, functions, or correspondences. The endogenous variables are most appropriately represented as $K_h \times 1$ vectors p_h such that $v = \{p_1, ..., p_n\}$. The elements of $f := \{f_1, ..., f_n\}$ are correspondingly vector-valued. These must satisfy $p_h = f_h(p_{(h)}, u) := \sigma_h(\arg\max_{p_h \in \mathbf{S}_h} \upsilon_h(p_1, ..., p_n))$.

In order to apply the PCM, we require a unique fixed point. Even when a unique Nash equilibrium exists, to obtain this as the fixed point requires choosing the selection operators σ_h so that they specifically produce the Nash equilibrium response. In the usual situation, the properties of f determine whether or not a fixed point exists. Here, however, knowledge of the unique fixed point is required to properly specify σ_h, hence f_h, an awkward reversal signaling that the PCM is not well-suited to this application. Indeed, the selection cannot be random, a plausible response when the player is indifferent between different strategies.

An interesting feature of this example is that when the PCM applies, it does so with vector-valued units rather than the scalar-valued units formally treated by Pearl (2000) or Halpern (2000). The PCM is thus necessarily silent about what happens when components of an agent's strategy are arbitrarily set. In contrast, settable systems apply to partitions both finer and coarser than the agent partition. (The elements (sets of unit indexes) of a "coarser" partition are unions of the elements of a "finer" partition. Thus, the agent partition is coarser than the elementary partition and finer than the global partition.)

Unlike the case of pure-strategy games, there must always be at least one mixed-strategy Nash equilibrium, so the PCM does not run into the difficulty that there may be no equilibrium. Nevertheless, mixed-strategy games can also have multiple Nash equilibria, so the PCM does not apply there. For a given game, the GPCM does apply to the agent partition, but it does not incorporate equilibrium selection mechanisms. In contrast, the settable system framework permits causal analysis at the level of the agent partition (as well as coarser or finer partitions); represents the unique Nash equilibrium at the level of the global partition without requiring a selection operator when a unique equilibrium exists; and otherwise represents the desired responses when a unique mixed-strategy Nash equilibrium does not exist but conventions or other plausible selection mechanisms apply.

Static games of complete information are the beginning of a sequence of increasingly richer games, including dynamic games of complete information, static games of incomplete information, and dynamic games of incomplete information. Each of these games employs progressively stronger equilibrium concepts that rule out implausible equilibria that would survive under equilibrium concepts suitable for simpler games (Gibbons, 1992, p. 173). These implausible equilibria all satisfy fixed-point (simple Nash equilibrium) requirements.

The unique fixed point requirement of the PCM thus acts to severely limit its applicability in game theory, due to the many opportunities for multiple Nash equilibria. Although GPCMs formally apply, they cannot support discourse about causal relations between endogenous variables, due to the lack of an analog of the potential response function. In contrast, by exploiting attributes and partitioning, settable systems permit implementation of whichever stronger and/or more re-

fined equilibria criteria are natural for a given game, together with any natural equilibrium selection mechanism.

6.3 Infinitely Repeated Dynamic Games

Dynamic games are played sequentially. For example, two players can repeatedly play *prisoner's dilemma*. In infinitely repeated games, play proceeds indefinitely. Clearly, infinite repetition cannot be handled in a finite system, so the PCM cannot apply.

In infinitely repeated dynamic games of complete and perfect information (see Gibbons, 1992, Section 2.3.B), players play a given static game in "stages" or periods $t = 1, 2, ...$ The period t payoff to player h is $\pi_{h,t} = u_{h,t}(\alpha_{1,t}, ..., \alpha_{n,t})$, where $u_{h,t}$ is player h's payoff function for period t, whose arguments are the "actions" $\alpha_{j,t}$ at time t of all n players. (The strategies of static games correspond to the actions of dynamic games.) Information is "complete," as each player knows the others' possible actions and payoff functions. Information is "perfect," as at every t, each player knows the entire history of play up to that period.

Rational players act to maximize their average present discounted value of payoff,

$$\bar{\pi}_h = \bar{u}_h(\alpha_1, ..., \alpha_n) := (1-\delta)\sum_{t=1}^{\infty} \delta^{t-1} u_{h,t}(\alpha_{1,t}, ..., \alpha_{n,t}),$$

where $\alpha_j := \{\alpha_{j,t}\}$ denotes player j's countable sequence of actions, and $0 < \delta < 1$ is a "discount rate" (common across players, for simplicity) that converts payoffs $\pi_{h,t}$ in period t to a value in period 1 as $\delta^{t-1}\pi_{h,t}$. A player's best response to any collective sequence of actions by the others is a solution to the problem

$$\max_{s_h \in S_h} \bar{u}_h(\alpha_1, ..., \alpha_n) \;\; subject\ to \;\; \alpha_{h,t} = s_{h,t}(\alpha_1^{t-1}, ..., \alpha_n^{t-1}), \;\; t = 1, 2, ...,$$

where S_h is player h's set of all admissible sequences $s_h := \{s_{h,t}\}$ of "strategy functions" $s_{h,t}$. These represent player h's action in period t as a function only of the prior histories of player actions, $\alpha_1^{t-1}, ..., \alpha_n^{t-1}$. (For $t = 1$, $s_{h,t}$ is a constant function.) Player h's best responses are $\alpha_{h,t}^* = s_{h,t}^*(\alpha_1^{t-1}, ..., \alpha_h^{*t-1}, ..., \alpha_n^{t-1})$, $t = 1, 2, ...$, where $s_h^* := \{s_{h,t}^*\}$ is a sequence of best response strategy functions. (These need not be unique, so $s_{h,t}^*$ may be a correspondence. The player is indifferent among the different possibilities.)

Such games generally have multiple Nash equilibria. Many of these are implausible, however, as they involve non-credible threats; and credibility is central to all dynamic games (see Gibbons, 1992, p. 55). Non-credible equilibria can be eliminated by retaining only "subgame perfect" Nash equilibria. These are Nash equilibria that solve not only the game beginning at time 1, but also the same game beginning at any time $t > 1$ (Gibbons, 1992, pp. 94-95). A celebrated result by James Friedman (1971) ensures the existence of one or more subgame perfect Nash equilibria, provided δ is close enough to one (see, e.g.. Gibbons, 1992, pp. 97-102). Significantly, such equilibria permit *tacit cooperation*, yielding outcomes superior to what players can achieve in the static game played at each stage.

We now map this game to settable systems. The units i now correspond to *agent-time pairs* (h,t). As $t = 1, 2, ...$, there is a countable infinity of units. Agent attributes include their admissible actions and their payoff functions for each period. That is, a_i (equivalently $a_{h,t}$) includes the admissible sequence of functions S_h, the utility function $u_{h,t}$, and the discount factor δ. When player actions

$\alpha_{h,t}$ are set arbitrarily, the settable system represents them as $z_{h,t}$. When players intelligently choose their actions, they are denoted $y_{h,t}$.

The agent partition $\Pi^a := \{\Pi^a_h, h = 1,...,n\}$, where $\Pi^a_h := \{(h,1),(h,2),...\}$, represents agents' best responses recursively as

$$y_{h,t} = \sigma_{h,t}(s^*_{h,t}(z_1^{t-1},...,y_h^{t-1},...,z_n^{t-1}),\omega_r), \;\; t = 1,2,...;h = 1,...,n,$$

where $\sigma_{h,t}$ is a measurable selection operator; $s^*_{h,t}$ is the agent's best response correspondence, which depends on the action histories of other agents, $z^{t-1}_{(h)}$, and agent h's history of prior best responses, y_h^{t-1}; and the realization ω_r determines random selections from the best response correspondence for agent h in period t. Recursive substitution for the elements of the history y_h^{t-1} yields a representation in terms of an agent-partition response function, $r^a_{h,t}$, namely

$$y_{h,t} = r^a_{h,t}(z^{t-1}_{(h)},\omega_r;\mathbf{a}), \;\; t = 1,2,...;h = 1,...,n.$$

The global partition represents whatever selection of the collection of subgame perfect Nash equilibria is natural or compelling. Equilibrium agent responses are given by the global-partition response functions $r^g_{h,t}$ as

$$y_{h,t} = r^g_{h,t}(\omega_r;\mathbf{a}) \quad t = 1,2,...;h = 1,...,n.$$

Notably, this example exploits each feature of stochastic settable systems, including countable dimension, attributes, partitioning, and pure randomness.

6.4 Settable Systems and Multi-Agent Influence Diagrams

Causal models other than the PCM are available in the machine learning literature. We focus here on the PCM because of its prevalence and to maintain a sharp focus for this paper.

A framework particularly related to the preceding discussion is that of Koller and Milch (2003) (KM), who introduce multi-agent influence diagrams (MAIDs) to represent noncooperative games. In particular, KM provide a graphical criterion for determining a notion of "strategic relevance." KM's "relevance graphs" are related to causal graphs. By casting games in the settable system framework, we can immediately construct causal graphs for games by applying the conventions of Section 3.6.2.

The most immediate similarity between settable systems and MAIDs is that they are both capable of modeling environments in which multiple agents interact. In contrast, "influence diagrams [...] have been investigated almost entirely in a single-agent setting" (KM, 2003, p. 189-190). Nevertheless, several features of settable systems distinguish them from MAIDs:

(i) A settable system is an explicit causal framework in which notions of partitioning, settings, interventions, responses, and causality are formally defined. Furthermore, the interrelations between these causal notions on the one hand and the notions of optimization, equilibrium, and learning on the other are made precise. In contrast, there is no formal mention of causality in KM.

(ii) MAIDs build on the "chain rule for Bayesian Networks" (KM, definition 2.2, p. 186). This is equivalent to assuming a set of (conditional) independence relations involving chance and decision variables and is necessary for the applicability of the "s-reachability" graphical criterion. On the other hand, settable systems permit but do not require any assumptions on the joint distribution of

settings and responses. In particular, responses may admit an aspect of "pure randomness" due to their direct dependence on the primary variable.

(*iii*) In KM, an agent's utility is additive (KM, p. 189-190). Settable systems do not impose this requirement.

(*iv*) The KM algorithm for finding Nash equilibria outputs one Nash equilibrium. It selects an equilibrium arbitrarily if multiple equilibria are found. Further, the algorithm cannot produce certain equilibria, such as a nonsubgame-perfect equilibrium (KM, p. 216). The settable system framework can represent principled selections from all relevant Nash equilibria.

We emphasize that the results in KM are very helpful for representing and studying games. In particular, under the MAID assumptions, the KM results permit an explicit representation of games and can lead to computational savings.

7. Machine Learning, Optimization, and Equilibrium

A general learning algorithm introduced by Kushner and Clark (1978) (KC) has the form

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \lambda_t M_t(\hat{\xi}_t, \hat{\theta}_t, \zeta_{t+1}), \tag{11}$$

$$\hat{\xi}_{t+1} = R_t(\hat{\xi}^t, \hat{\theta}^{t+1}, \zeta_{t+1}), \quad t = 0,1,2,..., \tag{12}$$

where $\hat{\theta}_t$ and $\hat{\xi}_t$ are random vectors, λ_t is a random scalar, M_t and R_t are known vector functions, $\hat{\xi}^t := (\hat{\xi}_0, ..., \hat{\xi}_t)$, $\hat{\theta}^{t+1} := (\hat{\theta}_0, ..., \hat{\theta}_{t+1})$, and ζ_t is an observable random vector. Initial values $\hat{\xi}_0$ and $\hat{\theta}_0$ are random vectors independent of $\{\zeta_t\}$. KC call this a Robbins and Monro (1951) (RM) algorithm with feedback (RMF). Equation (11) is an RM procedure; Equation (12) supplies the feedback. A main focus of interest is the convergence behavior of $\hat{\theta}_t$ as $t \to \infty$.

Chen and White (1998) analyze a version of RMF where each vector takes values in a real separable infinite-dimensional Hilbert space. We call this an HRMF algorithm. Because of the flexibility this affords, the HRMF supports nonparametric learning.

The RM procedure emerges when $\hat{\xi}_t$ has dimension zero, so $\hat{\theta}_{t+1} = \hat{\theta}_t + \lambda_t M_t(\hat{\theta}_t, \zeta_{t+1})$, $t = 0,1,2,....$ This contains recursive least squares (e.g., back-propagation), recursive maximum likelihood, and recursive method of moments procedures (e.g., Ljung and Soderstrom, 1983). The estimated weights are $\hat{\theta}_t$; $\{\zeta_t\}$ is the data sequence; λ_t is the "learning rate," for example, $\lambda_t = 1/t$; and M_t determines the learning method (least squares, maximum likelihood, etc.). By permitting feedback, the RMF accommodates the evolution of internal, possibly hidden states $\hat{\xi}_t$; thus, Kalman filter methods (Kalman, 1960) are a special case.

The RMF also describes learning in recurrent artificial neural networks (ANNs) (e.g., Elman, 1990; Jordan, 1992; Kuan, Hornik, and White, 1994). Here, the input sequence is $\{\zeta_t\}$; after t input observations, network weights are $\hat{\theta}_t$, and hidden unit activations are $\hat{\xi}_t$. The learning function is M_t, the learning rate is λ_t, and R_t determines hidden unit activations. The allowed randomness of λ_t accommodates simulated annealing.

The RMF and HRMF contain systems with learning by one or more optimizing agents. When there are multiple agents, the system can embody convergence to equilibrium. Specifically, Chen and White (1998) provide conditions ensuring system convergence as $t \to \infty$ to Nash equilibria or to "rational expectations" equilibria. As examples, Chen and White (1998) consider, among others, a learning agent solving a *stochastic dynamic programming* problem and the game of *fictitious play with continuum strategies* (an infinitely repeated dynamic game of incomplete information). The applications of the (H)RMF are thus broad; further, the settable systems framework contains both.

The units i are *time-agent-generalized decision triples*, (t,h,j). Specifically, at time t, agent h has generalized decisions indexed by j. Generalized decisions are *knowledge* (or in Bayesian frameworks, *beliefs*) denoted $\hat{\theta}_{t,h,k}$, $k = 1,...,k_h$, or *generalized actions*, denoted $\hat{\xi}_{t,h,\ell}$, $\ell = 1,...,\ell_h$. Generalized actions may be *actions* (as in Section 6.3) or *states*, as in the Kalman filtering and recurrent ANN examples. We write $\hat{\theta}_t := (\hat{\theta}'_{t,1},...,\hat{\theta}'_{t,n})'$, where $\hat{\theta}_{t,h} := (\hat{\theta}_{t,h,1},...,\hat{\theta}_{t,h,k_h})'$ takes values in Θ_h, a subset of $\mathbb{R}^{k_h}$, and $\hat{\xi}_t := (\hat{\xi}'_{t,1},...,\hat{\xi}'_{t,n})'$, where $\hat{\xi}_{t,h} := (\hat{\xi}_{t,h,1},...,\hat{\xi}_{t,h,\ell_h})'$ takes values in Ξ_h, a subset of $\mathbb{R}^{\ell_h}$, $h = 1,...,n$.

In addition to the time-agent-generalized decision indicators (t,h,j), attributes $a_{t,h,j}$ include as components the spaces Θ_h and Ξ_h and the function $\lambda_t : \Omega_r \to \mathbb{R}$. They can also include the functions $M_{t,h,k}$ or $R_{t,h,\ell}$, as appropriate. We write $M_t := (M'_{t,1},...,M'_{t,n})'$, with $M_{t,h} := (M_{t,h,1},...,M_{t,h,k_h})'$, and $R_t := (R'_{t,1},...,R'_{t,n})'$, with $R_{t,h} := (R_{t,h,1},...,R_{t,h,\ell_h})'$. The functions $M_{t,h,k}$ may be a consequence of an underlying optimization principle, as in our machine learning examples of Section 3. The same may be true of the functions $R_{t,h,\ell}$.

For the RMF, in Equations (11) and (12), n is finite, as are k_h and ℓ_h. Because t takes a countable infinity of values, we require a countably infinite settable system. For the HRMF, n may be countably infinite; similarly, k_h and/or ℓ_h may be countably infinite.

Equations (11) and (12) form a recursive or acyclic system. In such systems, there is a natural hierarchy of units, in which *predecessor* units drive *successor* units. The system evolves naturally (i.e., without intervention) when the response values at a given level of the hierarchy act as setting values for successors. Stochastic settable systems are sufficiently flexible to permit this. That is, given recursivity, for every ω_r in Ω_r, there exists ω_s in Ω_s such that $Z_i(\omega_s) = Y_i(\omega_r,\omega_s)$ for all i. When $\omega = (\omega_r,\omega_s)$ has this property, we call ω *canonical*, and we let $\Omega_c \subset \Omega$ denote the set of canonical primary settings ω. Response and setting values for a given unit thus coincide on Ω_c, implementing the natural evolution. In the present example, $Y_{t,h,j}$ and $Z_{t,h,j}$ correspond to an element of either $\hat{\theta}_t$ or $\hat{\xi}_t$. Fundamental settings are $\hat{\xi}_0, \hat{\theta}_0$, and $\{\zeta_t\}$, corresponding to elements of $\mathcal{X}_0(\,\cdot\,,1) := \mathcal{X}_0(\,\cdot\,,0)$.

Substituting Equation (11) into Equation (12) yields response functions for the *time partition*, $\Pi^t := \{\Pi^t_1,\Pi^t_2,...\}$, where $\Pi^t_b := \{(b,h,k), k = 1,...,k_h; (b,h,\ell), \ell = 1,...,\ell_h; h = 1,...,n\}$.

In the HRMF, agents' generalized decisions take values in real separable infinite-dimensional Hilbert spaces, so generalized decisions are not just vector-valued; their values may be suitably well-behaved functions. First, consider how a countably dimensioned settable system accommodates such objects when there is a single agent with a single action, a function, and a single knowledge element, also a function. We represent such functions by a countable vector whose elements are coefficients of terms in a suitable series representation, for example, a Fourier series. Further, this same approach applies without exhausting the dimensionality of the settable system, even when there is a countable infinity of agents, each having a countable infinity of knowledge elements and actions, which are themselves elements of real separable infinite-dimensional Hilbert spaces.

8. Summary and Concluding Remarks

This paper introduces settable systems, an extension of Pearl's (2000) causal model. Settable systems and the PCM share many common features. For example, in both frameworks the variables of the system have a dual role (set or free), there are mechanisms for specifying which variables are set or free (submodels and the do operator in the PCM, partitioning in settable systems), and attributes

may be accommodated (as background variables in the PCM and as a priori constants in settable systems).

The key difference between the PCM and settable systems is the way these common features interrelate to one another. Although we point out a number of limitations of the PCM in motivating settable systems, settable systems strictly build on the percepts of the PCM. Our intent is to show how modest reconfiguration and refinement of the elements of the PCM considerably enhance its explanatory power.

As we demonstrate, the PCM encounters obstacles when we attempt to apply it to certain machine learning examples. These limitations motivate particular features of settable systems. For example, the unique fixed point requirement of the PCM is a significant limitation. Like Halpern's (2000) GPCM, settable systems do not require the existence of a unique fixed point. The structure of settable systems nevertheless leads to natural notions of counterfactuals, interventions, direct causes, direct effects, and total effects. In contrast, the absence of the potential response function in the GPCM precludes causal discourse.

Another appealing feature of settable systems relative to the PCM is its ability to provide a causal role for structurally exogenous variables. This capability arises because settable systems distinguish between attributes and fundamental settings. In contrast, the PCM lumps together attributes and background variables, so neither can play a causal role. The PCM is silent on whether to treat variables as exogenous or endogenous and on how to specify attributes. In settable systems, the governing principles (e.g., optimization and equilibrium) provide explicit guidance for distinguishing exogenous variables and endogenous variables. Attributes are unambiguously defined as constants (numbers, functions, sets, etc.) associated with system units that define fundamental aspects of the decision problem represented by the settable system.

Our examples in game theory (Section 6) and machine learning with feedback (Section 7) further show that settable systems apply directly to systems where learning and/or optimizing agents interact in a process where outcomes satisfy or converge to appropriate equilibria. Settable systems thus provide rigorous foundations for causal analysis in these empirically relevant and computationally important systems.

These foundations are only a first step in analyzing phenomena conforming to settable systems. A particularly important research area is the study of general primitive conditions ensuring the identification of specific causal effects of interest under varying assumptions about the observability of causes of interest and other ancillary causes and under particular patterns of causal relation. In this context, identification means the equality of causally meaningful objects (e.g., expected effects) with corresponding stochastically meaningful objects, that is, quantities expressible solely as a functional of the joint distribution of observable random variables. When identification holds, it becomes possible to estimate various causal effects from data. Recent work of White (2006), White and Chalak (2007), Schennach et al. (2008), Chalak and White (2008a), and White and Kennedy (2009) provides results for identification and estimation of causal effects under varying assumptions.

Key to ensuring identification of effects of interest in all of these studies are specific independence or conditional independence conditions, for example, the conditional independence of causes of interest from unobservable ancillary causes given other observable variables (covariates). Chalak and White (2008b) provide primitive conditions on recursive settable system structures (in particular the response functions) that either ensure or rule out such independence or conditional independence relations. In pursuing this goal, notions of indirect and total effects of non-primary causes

emerge naturally and play key roles. These results also have direct implications for *d*-separation and *D*-separation (e.g., Geiger, Verma, and Pearl, 1990; Pearl, 2000, pp. 16-17).

These studies by no means exhaust the opportunities for deeper understanding and application of settable systems. For example, all of the studies of identification and estimation just mentioned are for recursive structures. Obtaining analogous results for non-recursive structures is of particular interest.

At the outset, we offered this paper as part of a cross-disciplinary dialog between the economics/econometrics community and the machine learning community, with the hope that both communities might gain thereby. For economists, the benefits are clear and precise notions of causal effects that apply broadly to economic structures and, in particular, to the powerful structures of game theory. These causal notions draw heavily on concepts at the heart of the PCM, but surmount a number of limitations that may have held back economists' acceptance of the PCM. For those in the machine learning community, one benefit is the extension of causal notions to systems fundamentally involving optimization, equilibrium, and learning, features common to a broad range of application domains relevant to machine learning. We also hope that the machine learning community, which has so far paid only limited attention to game theory, may begin to consider the possibilities it offers for understanding empirical phenomena and for distributed and emergent computation.

Acknowledgments

The authors are grateful for discussion and comments on previous work by Judea Pearl that stimulated this work, and for the comments and suggestions of James Heckman, Philip Neary, Douglas R. White, Scott White, the editor, and five referees. Any errors are solely the author's responsibility. The present paper is an expanded version of the foundational and definitional material contained in our earlier unpublished paper "A Unified Framework for Defining and Identifying Causal Effects."

Appendix A.

For completeness, we provide a formal definition of a non-stochastic settable system.

Definition 2 (Nonstochastic Settable System) *Let $n \in \bar{\mathbb{N}}^+$, and let A be a non-empty set. For each $i = 1,...,n$, let a_i be a fixed element of A, such that a_i includes a component $\mathbb{S}_i$, a multi-element Borel-measurable subset of $\mathbb{R}$.*

Let $m \in \bar{\mathbb{N}}$. For each $k = 1,...,m$, let $a_{0,k}$ be a fixed element of A, such that $a_{0,k}$ includes a component $\mathbb{S}_{0,k}$, a multi-element Borel-measurable subset of $\mathbb{R}$. Write $a_0 := (a_{0,1},...,a_{0,m})$ and $\mathbf{a} := (a_0, a_1, ..., a_n) \in \mathbf{A} := (\times_{k=1}^{m} A) \times (\times_{i=1}^{n} A)$.

For each $k = 1,...,m$, let $z_{0,k} \in \mathbb{S}_{0,k}$, and put $y_{0,k} := z_{0,k}$. Write $z_0 := (z_{0,1},...,z_{0,m})$ and $y_0 := z_0$.

Let $\Pi = \{\Pi_b\}$ be a partition of $\{1,...,n\}$, with $B := \#\Pi \in \bar{\mathbb{N}}^+$, let $\ell_b := \#\Pi_b$, and let $\mathbf{a}$ determine the multi-element Borel measurable set $\mathbb{S}^{\Pi}_{(b)}(\mathbf{a}) \subset \times_{j \notin \Pi_b} \mathbb{S}_j \times_{k=1}^{m} \mathbb{S}_{0,k}$, $b = 1,...,B$. Suppose there exist measurable functions

$$r^{\Pi}_{[b]}(\,\cdot\,;\mathbf{a}) : \times_{i \in \Pi_b} \mathbb{S}_i \times \mathbb{S}^{\Pi}_{(b)}(\mathbf{a}) \to \mathbb{R}^{\ell_b} \quad b = 1,...,B,$$

and real vectors $y^{\Pi}_{[b]} \in \times_{i\in\Pi_b}\mathbb{S}_i$ *such that for each* $z^{\Pi}_{(b)} \in \mathbb{S}^{\Pi}_{(b)}(\mathbf{a})$,

$$r^{\Pi}_{[b]}(y^{\Pi}_{[b]}, z^{\Pi}_{(b)}; \mathbf{a}) = \mathbf{0}, \;\; b = 1, ..., B.$$

Write $\mathcal{X}^{\Pi}_0(0) := y_0$, $\mathcal{X}^{\Pi}_0(1) := z_0$, $\mathcal{X}^{\Pi}_i(0) := y^{\Pi}_i$, $\mathcal{X}^{\Pi}_i(1) := z^{\Pi}_i$, $i = 1, ..., n$, *so that* $\mathcal{X}^{\Pi}_0 : \{0,1\} \to \times^m_{k=1}\mathbb{S}_{0,k}$ *and* $\mathcal{X}^{\Pi}_i : \{0,1\} \to \mathbb{S}_i, i = 1, ..., n$. *Finally, write*

$$\mathcal{X}^{\Pi} := (\mathcal{X}^{\Pi}_0, \mathcal{X}^{\Pi}_1, , ..., \mathcal{X}^{\Pi}_n).$$

Then $\mathcal{S}^{\Pi} := \{(\mathbf{A}, \mathbf{a}), (\Pi, \mathcal{X}^{\Pi})\}$ *is a* nonstochastic settable system.

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

C. Berge. *Espaces Topologiques*. Paris: Dunod (translation by E.M. Patterson, Topological Spaces. Edinburgh: Oliver and Boyd), 1963.

K. Chalak and H. White. An extended class of instrumental variables for the estimation of causal effects. Technical report, Department of Economics, University of California, San Diego, 2008a.

K. Chalak and H. White. Independence and conditional independence in causal systems. Technical report, Department of Economics, University of California, San Diego, 2008b.

X. Chen and H. White. Nonparametric adaptive learning with feedback. *Journal of Economic Theory*, 82:190–222, 1998.

A.P. Dawid. Influence diagrams for causal modeling and inference. *International Statistical Review*, 70:161–189, 2002.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

M. Eichler and V. Didelez. Causal reasoning in graphical time-series models. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence*, 2007.

J. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

F. Fisher. A correspondence principle for simultaneous equations. *Econometrica*, 38:73–92, 1970.

J. Friedman. A noncooperative equilibrium for supergames. *Review of Economic Studies*, 38:1–12, 1971.

D. Geiger, T. S. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

R. Gibbons. *Game Theory for Applied Economists*. Princeton: Princeton University Press, 1992.

J. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 282–317. Cambridge: MIT Press, 1986.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81: 945–970, 1986.

R.A. Howard and J. E. Matheson. Influence diagrams. In R.A. Howard and J.E. Matheson, editors, *Readings in the Principles and Applications of Decision Analysis*. Menlo Park, CA: Strategic Decisions Group, 1984.

M. Jordan. Constrained supervised learning. *Journal of Mathematical Psychology*, 36:396–425, 1992.

R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Series D, Journal of Basic Engineering*, 82:35 – 45, 1960.

D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45:181–221, 2003.

C.-M. Kuan, K. Hornik, and H. White. A convergence result for learning in recurrent neural networks. *Neural Networks*, 6:420–440, 1994.

H. Kushner and D. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Berlin: Springer-Verlag, 1978.

L. Ljung and T. Soderstrom. *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

J. Marsden and A. Tromba. *Vector Calculus*. New York: W.H. Freeman, 5th edition, 2003.

J. Nash. Equilibrium points in *n*-person games. In *Proceedings of the National Academy of Sciences*, volume 36, pages 48–49, 1950.

L.G. Neuberg. Review of *Causality: Models, Reasoning, and Inference*. *Econometric Theory*, 19: 675–685, 2003.

J. Pearl. *Causality*. New York: Cambridge University Press, 2000.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

S. Schennach, H. White, and K. Chalak. Estimating average marginal effects in nonseparable structural systems. Technical report, Department of Economics, University of California, San Diego, 2008.

Y. Sergeyev and V. Grishagin. Parallel asynchronous global search and the nested optimization scheme. *Journal of Computational Analysis and Applications*, 3:123–145, 2001.

B. Shipley. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference.* Cambridge University Press, 2000.

R. Strotz and H. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

S. van Huffel and P. Lemmerling. *Total Least Squares and Errors-In-Variables Modeling: Analysis, Algorithms and Applications*. Berlin: Springer-Verlag, 2002.

H. White. *Asymptotic Theory for Econometricians*. New York: Academic Press, 2001.

H. White. Time-series estimation of the effects of natural experiments. *Journal of Econometrics*, 135:527–566, 2006.

H. White and K. Chalak. Identifying effects of endogenous causes in nonseparable systems using covariates. Technical report, Department of Economics, University of California, San Diego, 2007.

H. White and P. Kennedy. Retrospective estimation of causal effects through time. In J. Castle and N. Shephard, editors, *The Methodology and Practice of Econometrics: A Festschrift in Honour of David Hendry*. Oxford: Oxford University Press, 2009.