Optimal Causal Imputation for Control

Roy Dong, Eric Mazumdar, and S. Shankar Sastry

Abstract— The widespread applicability of analytics in cyber-physical systems has motivated research into causal inference methods. Predictive estimators are not sufficient when analytics are used for decision making; rather, the flow of causal effects must be determined. Generally speaking, these methods focus on estimation of a causal structure from experimental data. In this paper, we consider the dual problem: we fix the causal structure and optimize over causal imputations to achieve desirable system behaviors for a minimal imputation cost. First, we present the optimal causal imputation problem, and then we analyze the problem in two special cases: 1) when the causal imputations can only impute to a fixed value, 2) when the causal structure has linear dynamics with additive Gaussian noise. This optimal causal imputation framework serves to bridge the gap between causal structures and control.

I. INTRODUCTION

Recently, data analytics have achieved amazing levels of success. As analytics penetrate more and more industrial applications, they are increasingly used for decision-making and planning. In these applications, it is important to use estimators that are not only predictive, but estimate the causal structure of the underlying processes.

Correlation is not the same as causation. However, in practice, it is not always easy to apply this principle. In many real-life applications, machine learning is used to determine the relationship between two variables. This analysis is often used as the basis for determining which actions to take. However, an algorithm with low test error does not necessarily mean that the causal effect has been estimated.

For example, one may train a classifier to estimate the energy consumption of a household given the presence and absence of eco-friendly devices, and this may provide guidelines for which devices should be discounted through rebate programs. Unless the causal structures are explicitly accounted for, there could easily be confounding variables or incorrect causal relationships that change the behavior of the system under consideration.

This has motivated new interest in causal inference techniques. Generally speaking, these techniques take experimental data and attempt to uncover the causal structure. (We defer a literature review of these methods to Section III, when a more formal model of causality has been developed.) In this paper, we consider the dual problem: we fix the causal structure and attempt to determine what causal actions will lead to system behaviors we desire at a minimal cost.

R. Dong, E. Mazumdar, and S. S. Sastry are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, 94707, USA {roydong,emazumdar,sastry}@eecs.berkeley.edu

A. Outline

The rest of the paper is organized as follows. We discuss the main paradigms for modeling causality in Section II. In Section III, we outline the mathematical formulation of a causal structure, discuss relevant literature in causal estimation, and define the problem of optimal causal imputation. In Section IV, we provide theoretical analysis of two special cases of the optimal causal imputation problem: the case where imputation can only be done to a single value, and the case where the dynamics are linear and the noise is Gaussian. Finally, we present closing remarks in Section V.

II. BACKGROUND

There are three main paradigms for the mathematical modeling of causality:

1) Rubin causality
2) Granger causality
3) Pearl's structural equation modeling (SEM)

Each of these paradigms has a vast literature in its own right; we will try to present a few representative samples from each field here. Note that each paradigm uses its own notation, so we will change notation as we switch from approach to approach.

It should be noted that these paradigms are not mutually exclusive: for example, a problem that is modeled using Granger causality can be put into Pearl's SEM if the underlying processes operate in discrete time. Rubin causality can often be phrased as an SEM problem, but in applications this will require more structural assumptions to learn the causal structure. A full exposition of the intersections and non-intersections of these three paradigms is outside the scope of this paper, but we note that these paradigms can often model the same phenomena and shed different insights on the causal behaviors observed.

Rubin causality was first introduced in [1]. In the basic formulation of Rubin causality, we are given some control variable X taking values in $\{0, 1\}$. There are also two distinct random variables Y_0 and Y_1. If $X = 0$, then we observe Y_0 and not Y_1. If $X = 1$, then we only observe Y_1, and not Y_0. Another way to write this notationally is that we observe Y_X but do not observe Y_{1-X}, which is often called the counterfactual. The fact that we can only observe one or the other, but not both, is the fundamental misery of causality.

One of the key results that the Rubin causality paradigm provides is that if X is independent of Y_0 and Y_1, then randomly assigning $X \in \{0, 1\}$ yields a dataset that can provide valid estimates of the counterfactuals; thus, Rubin causality provides the theoretical foundation for randomized

control trials. This paradigm has also been extended to consider many covariates [2], handle confounding variables and incorporate instrumental variables [2], and incorporate some machine learning approaches [3]. Sample applications include estimating the causal effect of residential demand response in the Western United States [4] or the causal effects of providing money, healthcare and education to the very poor in Ethiopia, Ghana, Honduras, India, Pakistan, and Peru [5].

Granger causality was first introduced in [6]. In this paradigm, we are given data from two stationary random processes X and Y, both indexed by time. First, let U_t denote all the information available in the universe at time t, and let $(U - X)_t$ denote all the information available at time t except for X. Then, let $\sigma^2(Y|U)$ denote the error variance of the unbiased, least-squares estimator of Y_t using U_t, and similarly let $\sigma^2(Y|U - X)$ denote the error variance of the unbiased, least-squares estimator of Y_t using $(U - X)_t$. Then, X Granger-causes (or G-causes, for short) Y if $\sigma^2(Y|U) < \sigma^2(Y|U-X)$, i.e. the estimator that utilizes X has lower variance on its error than the one that cannot. In other words, X has explanatory power for Y.

Granger causality essentially relies on the relationship between causal effects and the arrow of time to distinguish it from general correlations. Although this framework does not address many of the more pernicious philosophical aspects of causality, oftentimes prior knowledge allows us to make the inductive leap from time-lagged correlations to causality. This paradigm is particularly appealing because it is easy to calculate in practice. Sample applications include determining which neuron assemblies Granger-cause other neuron assemblies to fire synapses [7] or finding that exchange rates Granger-cause stock market prices in Asia [8].

Pearl's SEM approach to causality models the statistical relationship between random elements with a Bayesian network [9]. Bayesian networks are directed acyclic graphs, such that the distribution of a random element at node i only depends on the values taken at the parent nodes. This is meant to model causal relationships between nodes in the graph. Pearl defines the imputation operator as follows: if one imputes at a node i, one disconnects i from all its parents and deterministically sets its value to some fixed, predetermined constant. We will be building on this approach in this paper, so we will defer the formal development of Pearl's SEM until Section III.

At a high level, the imputation operator captures a lot of our intuitions about how the subjunctive conditional should function. When one says *If it had rained today, I would have brought my umbrella*, what does one mean? Intuitively, one often means: 'If everything else were the same, only it is the case that it is raining today instead of sunny, these are the actions I would have taken.' One does not mean that the world is structured in a way such that the necessary processes to induce rain today were instead the case. In other words: causal imputation does not travel upstream, e.g. backwards through time. This is captured in Pearl's SEM.

More practically, consider the question: *What are the*

causal effects of this medication? If we wish to estimate this, we should 'set' medication taken to TRUE, and see the consequences of this imputation. If we do not explicitly 'set' this value, then the decision to take medication is a consequence of preceding factors. This makes it difficult to determine if the observed effects are a result of the medication or some other confounding variables[1]. Again, this will be more formally discussed in Section III.

Thus, we can think of these paradigms in terms of the central phenomenon it is designed to model. In summary:

1) Rubin causality is focused on the estimation of the *counterfactual*.
2) Granger causality is focused on the *explanatory power* one process provides over another process.
3) Pearl's SEM is focused on the causal effects of the *imputation* operator.

Throughout this paper, we use Pearl's SEM. However, we note again that oftentimes problems framed in the Rubin causality or Granger causality paradigm often can be translated to an equivalent formulation in SEM.

A. Notation

For any set A, we denote the powerset of A as 2^A, which can also be thought of as the set of functions mapping $A \to \{0, 1\}$. For a collection of sets $\{A_i\}_{i \in I}$, we denote the Cartesian product as $\prod_{i \in I} A_i$.

Also, I will denote the identity matrix, where context will often be sufficient to determine its dimensions.

We let $U[a, b]$ denote the uniform distribution on the interval $[a, b]$ and $N(\mu, \Sigma)$ to denote the multivariate Gaussian distribution with mean μ and covariance matrix Σ.

III. CAUSAL FRAMEWORK

In this section, we introduce our framework for modeling causal effects, and then define the problem of optimal causal imputation.

A. Causal structure

We build on the structural equation modeling framework presented in [9]. First, we will introduce Bayesian networks.

Definition 1. *A* directed graph *$G = (V, E)$ is a set of nodes V and a set of edges $E \subset V \times V$. Throughout this paper we will assume V is at most countably infinite.*

A path *from $v_0 \in V$ to $v_N \in V$ is a finite sequence of edges $(v_0, v_1), (v_1, v_2), \ldots, (v_{N-1}, v_N) \in E$.*

We define the parents *of node i as $\mathrm{pa}(i) = \{j : (j, i) \in E\}$.*

We can iterate this relationship to define the ancestor relationship: let $\mathrm{pa}^n(i) = \{j : k \in \mathrm{pa}^{n-1}(i), (j, k) \in E\}$, where $\mathrm{pa}^1(i) = \mathrm{pa}(i)$ defined above. Then, the ancestors *of a node i are given by $\mathrm{anc}(i) = \cup_{n=1}^{\infty} \mathrm{pa}^n(i)$.*

We say j is a descendant *of i if $i \in \mathrm{anc}(j)$.*

A directed graph is acyclic *if $i \notin \mathrm{anc}(i)$ for every $i \in V$. We will refer to such graphs as* directed acyclic graphs *(DAGs).*

---

[1]We note that similar reasoning can be done in the Rubin causality formulation as well.

Definition 2. *A* random process X *indexed by a set* V *is a collection of random elements* $(X_i)_{i \in V}$. *We will let* $\mathcal{X}_i$ *denote the possible values of* X_i, *and* $\mathcal{X} = \prod_{i \in V} \mathcal{X}_i$.

When there is an associated graph $G = (V, E)$, *we will use the notation* $\mathrm{pa}(X_i)$ *to denote the tuple* $(X_j)_{j \in \mathrm{pa}(i)}$.

Definition 3. *A* random process X *indexed by* V *is* Markov *relative to a DAG* $G = (V, E)$ *if its distribution factorizes:*

$$P(X) = \prod_{i \in V} P(X_i | \mathrm{pa}(X_i))$$

We can also say that X *and* G *are* compatible, *or* G *represents* X.

This formalization will serve as our model for causality. The interpretation is that if there is an edge going from i to j, then X_i causes X_j.

Throughout this paper, we will treat the causal structure $G = (V, E)$ as given. Estimation of this causal structure is a non-trivial task, and an active topic of research. Some approaches to the task of causal inference include: using metrics like directed information to estimate the causal strength between random variables [10], [11], graphical-model based methods for estimating structure between random variables [12], [13], [14], [15], and regression based approaches [16], [17], [18]. Again, this list is far from exhaustive as an extensive literature review of this general field is outside the scope of this paper. For a broader overview of various approaches to the problem of causal inference, see [12], [9].

Although the estimation of causal structures is never a simple task, the growing field of research promises more and more applications in which accurate estimation of causal structures is feasible.

Previous work has focused on the *estimation* of causal structures. In contrast, our contribution is to consider the problem of *control* of causal structures. In other words, once we are given a causal structure, how can we impute causal effects to drive the overall system into a desirable state?

For example, once we can estimate the causal effects of issuing rebates for energy-efficiency appliances, how do we best distribute these rebates to induce more energy-efficient consumption patterns? To the best of our knowledge, this is the first paper to consider the problem of when and where to impute on a causal structure.

There is an equivalent formulation of the condition in Definition 3 which utilizes *disintegration* results in probability theory. This is referred to as the structural equation modeling framework in [9].

Proposition 1. *[19], [9] A* random process X *indexed by* V *is Markov relative to* $G = (V, E)$ *if and only if there exists a collection of functions* $(f_i)_{i \in V}$ *and independent random elements* $(\xi_i)_{i \in V}$ *such that:*

$$X_i = f_i(\mathrm{pa}(X_i), \xi_i) \tag{1}$$

Furthermore, if $\mathcal{X}_i$ *are Borel spaces[2], then* ξ_i *can be taken*

---

[2] A measurable space S is Borel if there exists a measurable function $S \to [0, 1]$ with a measurable inverse.

---

to be $U[0, 1]$.

We note that Borel spaces are a very general category of measurable spaces: they include Polish spaces equipped with the Borel σ-algebra[3]. This includes finite sets, $\mathbb{R}$, $\mathbb{R}^n$, $L^p(\mathbb{R}^n)$, the set of p-integrable functions defined on $\mathbb{R}^n$. Additionally, the space of probability distributions on any Borel space is also a Borel space.

Assumption 1. *Throughout the rest of this paper, we will always use* X *to denote a random process indexed by* V *that is Markov relative to a DAG* $G = (V, E)$, *where* X_i *takes values in* $\mathcal{X}_i$. *Similarly,* f_i *shall denote the functions as specified in Equation 1, and similarly* ξ_i.

B. Causal imputation

In this section, we will formally define the *causal imputation* operation. Intuitively, imputation of X produces a new random process Y. This random process Y is equal to X prior to the causal imputation, is forced to some value at the node of imputation, and experiences causal effects after the node of imputation. This is formally defined below.

Definition 4. *[9] A* random process Y *indexed by* V *is the* imputation *of* X *at* $i \in V$ *to a constant* $x_i \in \mathcal{X}_i$ *if:*

- $Y_i = x_i$.
- *For any* j *that is not a descendant of* i, $Y_j = X_j$.
- *For any* j *that is a descendant of* i, $Y_j = f_j(\mathrm{pa}(Y_j), \xi_j)$.

If this is the case, we will write $Y = \mathrm{do}(X; i, x_i)$.

The imputation operator produces a copy of the original process that is exactly equal at all nodes that do not causally depend on the node of imputation X_i. At the point of imputation, the node is disconnected from its parents and forced a constant value x_i. The nodes X_j that causally depend on X_i are replaced with new values that depend on the causal effects of X_i, keeping the innovation terms ξ constant throughout.

Referring back to the discussions in Section II, this can be thought of as manually setting the value of X_i to x_i. This should be something that is done exogenously, as a control variable, rather than as a consequence of endogenous factors: this is why Y_i is disconnected from $\mathrm{pa}(Y_i)$.

From this definition, it immediately follows that the imputation operator commutes.

Proposition 2. *Let* $i, j \in V$ *such that* $i \neq j$ *and* $x_i \in \mathcal{X}_i$ *and* $x_j \in \mathcal{X}_j$. *Then* $\mathrm{do}(\mathrm{do}(X; i, x_i); j, x_j) = \mathrm{do}(\mathrm{do}(X; j, x_j); i, x_i)$ *almost surely.*

This allows us to define imputation on any set of nodes, rather than just at a single node.

Definition 5. *For any* $I \subset V$ *and* $x_I \in \prod_{i \in I} \mathcal{X}_i$, *we define the imputation* $Y = \mathrm{do}(X; I, x_I)$ *as the sequential application of element-wise* do *operations. This is almost surely unique by Proposition 2.*

---

[3] A topological space T is Polish if it is separable and completely metrizable. The Borel σ-algebra of a topological space is the smallest σ-algebra containing all the open sets.

C. Optimal causal imputation

In the previous section, we defined the causal imputation operator. We can think of our system designer as having the capacity of issuing control commands that have causal effects on the system downstream. When we can define the cost of imputation as well as a control objective, we can formulate the optimal causal imputation problem.

We suppose we are given a collection of functions $(c_I)_{I \subset V}$ where each $c_I : \prod_{i \in I} \mathcal{X}_i \to \mathbb{R}$. These functions can be interpreted as the cost of imputation at a set of nodes $I \subset V$. Drawing on our running example, c represents the cost of issuing rebates for eco-friendly refrigerators at a set of households.

Furthermore, we suppose we are given an operational objective in the form of a cost function $g : \mathcal{X} \to \mathbb{R}$. For example, g can be a penalty on energy-wasting consumption patterns.

Definition 6. *The problem of* optimal causal imputation *is given by:*

$$\min_{I \subset V} \min_{x_I \in \prod_{i \in I} \mathcal{X}_i} c_I(x_I) + \mathbb{E}_Y[g(Y)] \tag{2}$$

$$\text{subject to} \quad Y = \mathrm{do}(X; I, x_I) \tag{3}$$

IV. APPLICATIONS

In Section III, we defined the optimal causal imputation problem in its full generality. In this section, we shall provide methods to solve the optimal causal imputation problem in special cases. In particular, we consider two contexts: 1) situations where imputation is only allowed to a single value, 2) situations where the dynamics are linear-Gaussian. In both instances, we shall assume $\mathcal{X}_i = \mathbb{R}^{n_i}$ for some n_i.

A. Single-value case

In many applications where we can causally impute values, we can only impute to one particular value. For example, when issuing incentives, we may be able to only offer one form of rebate to consumers. Motivated by this context, we consider situations where the optimal causal imputation problem can be reduced to one of submodular optimization.

Assumption 2. *In this section, we assume V is a finite set and that for each $I \subset V$, there exists an x_I such that $c_I(x_I) < \infty$ and $c_I(x'_I) = \infty$ for any $x'_I \neq x_I$. We shall refer to this as the single-value case.*

In the single-value case, we use the shorthand $F(I) = c(I) + \mathbb{E}[g(\mathrm{do}(X; I))]$, where we drop dependencies on x as it can only take a single value.

1) Submodular minimization:

Definition 7. *The set mapping $F : 2^V \to \mathbb{R}$ is* submodular *if for any $I_1 \subset I_2 \subset V$ and $i \in G \setminus I_2$, we have:*

$$F(I_1 \cup \{i\}) - F(I_1) \geq F(I_2 \cup \{i\}) - F(I_2) \tag{4}$$

Intuitively, this definition is motivated by economies of scale. We often expect economies of scale from these imputations, e.g. the per-customer cost of a rebate is non-increasing

as the number of customers increases, due to bulk-purchase discounts. In our running example, the additional cost of issuing a rebate to customer i is higher when you have issued few rebates than when you have issued a lot of rebates. (When $I_1 \subset I_2$, then I_2 corresponds to the situation where you have issued more rebates than I_1.)

From a combinatorial optimization perspective, submodularity is a very well-behaved property that makes optimization, or approximate optimization, very tractable. We shall quickly outline the details now, but we refer the interested reader to [20] for more details.

First, note that there is a very direct correspondence between a subset $I \subset V$ and a tuple in $\{0, 1\}^V$. For example, if $V = \{0, 1, 2\}$, then $(0, 1, 1)$ corresponds to the subset $\{1, 2\}$. Thus, we can think of $F : \{0, 1\}^V \to \mathbb{R}$. Now, we define the Lovász extension [21].

Definition 8. *Let $\lambda \sim U[0, 1]$. Then, for any set mapping $F : \{0, 1\}^V \to \mathbb{R}$, we define the* Lovász extension $f : [0, 1]^V \to \mathbb{R}$ *as:*

$$f(z) = \mathbb{E}_\lambda[F(\{i : z_i > \lambda\})]$$

For the rest of this section, an unindexed f will denote the Lovász extension of F.

We note two nice properties of the Lovász extension immediately.

Proposition 3. *[21] For any $z \in \{0, 1\}^V$, we have $f(z) = F(z)$.*

Proposition 4. *[21] F is submodular if and only if f is convex.*

Note that the optimal causal imputation problem can be written as:

$$\min_{z \in \{0, 1\}^V} F(z)$$

The Lovász extension provides us with an easy solution to the problem.

Proposition 5. *[21] If F is submodular, then the following is a convex optimization program.*

$$\min_{z \in [0, 1]^V} f(z) \tag{5}$$

Furthermore, there exist minimizers of (5) in $\{0, 1\}^V$.

In other words, the combinatorial optimization problem can be solved tractably with convex optimization if F is submodular. Thus, we are motivated in searching for conditions under which $F(I) = c(I) + \mathbb{E}[g(\mathrm{do}(X; I))]$ is submodular. We provide a common sufficient condition for submodularity of F in the following theorem:

Theorem 1. *If:*
- *$g(Y) = \|Y_i - \mathbb{E}Y_i\|_2^2$ for some $i \in V$.*
- *There exists functions f_j^ξ such that, if X_j has no parents, $X_j = f_j^\xi(\xi_j)$ and otherwise $X_j = \mathrm{pa}(X_j) + f_j^\xi(\xi_j)$.*
- *For each $j \in \mathrm{anc}(i)$, there exists one unique path from j to i.*

- $c(I)$ *is submodular.*

Then $F(I) = c(I) + \mathbb{E}[g(\mathrm{do}(X; I))]$ *is submodular.*

Note here that we treat $\mathrm{pa}(X_i)$ as a vector in $\mathbb{R}^{n_i}$, where n_i is the appropriate dimension. These assumptions encompass many graphical models where a node's parents set a location parameter, and the control objective is the second moment of some feature.

Proof. Note that the desired result will follow if we show that the set mapping $G : I \mapsto \mathbb{E}[g(\mathrm{do}(X; I))]$ is submodular, since the sum of submodular functions is submodular. Throughout this proof, we use i to refer to the index i pulled out by the function g.

We can see that $G(\emptyset) = \mathbb{E}[g(X)]$. By the independence of the $(\xi_i)_{i \in V}$ and the form of the $(X_i)_{i \in V}$ relationships, we can write this as $\mathbb{E}[g(X)] = \sum_{j \in \mathrm{anc}(i)} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$. (Note that the unique path assumption ensures that each variance is only counted once in this sum.)

More generally, we can write an expression for $G(I)$. Note that if we impute at a node j, all the uncertainty due to node j, and the ancestors of j, is zeroed out. Thus, we can write $G(I) = \mathbb{E}[g(X)] - \sum_{j \in (I \cup \mathrm{anc}(I))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$, where we define $\mathrm{anc}(I) = \cup_{j \in I}\mathrm{anc}(j)$.

Now, we can verify the submodularity condition on G. Pick $I_1 \subset I_2$ and $i' \in V \setminus I_2$. Then:

$$G(I_1 \cup \{i'\}) - G(I_1) =$$

$$\sum_{j \in (I_1 \cup \mathrm{anc}(I_1))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2 -$$

$$\sum_{j \in (I_1 \cup \{i'\} \cup \mathrm{anc}(I_1 \cup \{i'\}))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2 =$$

$$- \sum_{j \in \{i'\} \cup (\mathrm{anc}(i') \setminus \mathrm{anc}(I_1))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$$

In words, the change in G due to adding i' to I_1 is the variances due to the terms related to i' and the ancestors of i' that have not already been zeroed out due to imputation, i.e. the ancestors of i' that are not already ancestors of I_1. A similar derivation can be done for I_2.

Thus, we can verify that $G(I_1 \cup \{i'\}) - G(I_1) \geq G(I_2 \cup \{i'\}) - G(I_2)$ by noting that $\mathrm{anc}(i') \setminus \mathrm{anc}(I_2) \subset \mathrm{anc}(i') \setminus \mathrm{anc}(I_1)$, so the right-hand side of the inequality adds more negative terms. This concludes our proof. $\square$

2) Submodular maximization: Alternatively, suppose we are attempting to maximize a submodular function subject to a constraint, i.e. $F(I) = c(I) + \mathbb{E}[g(\mathrm{do}(X; I))]$ subject to a constraint that $I \in S \subset 2^V$ and our objective is to solve $\max_{I \in S} F(I)$.[4]

First, consider the greedy method for submodular maximization. This is presented as Algorithm 1. At each iteration, it simply adds an element to I which maximizes $F(I \cup \{i\})$, if one exists. If one does not exist, it terminates and returns

---

[4]Strictly speaking, to remain consistent with the problem in Section III, we should be solving $\min_{I \in S} -F(I)$, but we express it as a maximization for clarity of presentation.

Algorithm 1 The greedy approach for combinatorial maximization.

$I \leftarrow \emptyset$
while $\max_{i:I \cup \{i\} \in S} F(I \cup \{i\}) - F(I) \geq 0$ **do**
  Pick $i^* \in \arg\max_{i:I \cup \{i\} \in S} F(I \cup \{i\})$
  $I \leftarrow I \cup \{i^*\}$
end while
return I

I. Under certain structural conditions, this algorithm yields approximate optimizers.

Definition 9. *A set mapping $F : 2^V \to \mathbb{R}$ is nondecreasing if $F(S) \leq F(T)$ whenever $S \subset T$.*

The monotonicity condition effectively prevents the algorithm from straying too far from the optimum when taking the greedy approach, as shown in [22]. Note that if F is nondecreasing, then the condition $\max_{i:I \cup \{i\} \in S} F(I \cup \{i\}) - F(I) \geq 0$ is equivalent to the existence of $i \in V$ such that $I \cup \{i\} \in S$.

Proposition 6. *[22] If F is nondecreasing and submodular, then the greedy method presented in Algorithm 1 will return $I^* \in S$ such that $F(I^*) \geq \left(\frac{e-1}{e}\right) \max_{I \in S} F(I)$.*

We now present a quick corollary of Theorem 1, which provides conditions under which we can leverage the existing results for maximization of nondecreasing submodular functions.

Corollary 1. *If:*
- $g'(Y) = -\|Y_i - \mathbb{E}Y_i\|_2^2$ *for some $i \in V$.*
- *There exists functions f_j^ξ such that, if X_j has no parents, $X_j = f_j^\xi(\xi_j)$ and otherwise $X_j = \mathrm{pa}(X_j) + f_j^\xi(\xi_j)$.*
- *For each $j \in \mathrm{anc}(i)$, there exists one unique path from j to i.*
- $c(I)$ *is nondecreasing and submodular.*

Then $F(I) = c(I) + \mathbb{E}[g'(\mathrm{do}(X; I))]$ is nondecreasing and submodular.

Proof. This follows from Theorem 1 if we can show that $G' : I \mapsto \mathbb{E}[g(\mathrm{do}(X; I))]$ is nondecreasing. Let $Y = \mathrm{do}(X; I)$, and note that adding elements to I can only decrease the variance of Y_i. This can be formalized by noting, similar to the arguments in the proof of Theorem 1, $G'(I) = \mathbb{E}[g'(X)] + \sum_{j \in (I \cup \mathrm{anc}(I))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$. Thus, G', the additive inverse of the variance of Y_i, is nondecreasing. $\square$

Note the minus sign in g' in Corollary 1: in most instances where you are maximizing a submodular cost, you would still wish to reduce uncertainty, i.e. have a lower variance.

B. Linear-Gaussian case

In this section, we consider causal imputation on a discrete-time linear dynamical system with Gaussian noise. That is, we analyze the special case of a random process with the form:

$$X_{t+1} = AX_t + \epsilon_t$$

Where $X_t \in \mathbb{R}^n$, $\epsilon_t \sim N(0, \sigma^2 I)$ independently for $t = 0, ..., T$, and $A \in \mathbb{R}^{n \times n}$ is a matrix representing the dependencies.

This process can be represented as a causal graph in the form of a trellis, where the random variables are all Gaussian. More specifically, each node has its expected value equal to a linear combination of their parents, as described by a matrix A, and additive noise of the distribution $N(0, \sigma^2)$.

To analyze our optimal casual imputation problem, we first redefine the indices for this problem. Since our causal graph represents a process over time, we index into the process by state k, for $k = 1, ...n$ as well as a time t for $t = 0, ..., T$. Thus X_{kt} indicates the value of state k at time t, and our graph has vertices $V = \{1, ..., n\} \times \{0, ..., T\}$. As before, X_t represents the value of the vector of all the states of X at time t, and we can think of X as a vector in $\mathbb{R}^{nT}$. We assume that the cost of imputation $c_I(x_I)$ has the following form for some parameters $\delta_i, q_i \geq 0$:

$$c_I(x_I) = \sum_{i \in I} \delta_i + q_i x_i^2$$

Further, we look at the case where the system cost of interest is minimizing the expected distance of the the random process from some target trajectory $\bar{y}$. Thus $g(Y) = \|Y - \bar{y}\|_2^2$.

Our optimal causal imputation problem in this case is thus:

$$\min_{S \subset V} \min_{x_S \in \mathbb{R}^S} \sum_{i \in S} \left(\delta_i + q_i x_i^2 \right) + \mathbb{E}\left[\|Y - \bar{y}\|_2^2 \right] \tag{6}$$
$$\text{subject to } Y = \text{do}(X; S, x_S)$$

The summation term can be thought of as a cost of issuing control commands and the expectation term can be thought of as a trajectory tracking objective.

Given our structure on the random process, we can rewrite this optimization problem more concretely.

We first define $Q \in \mathbb{R}^{nT \times nT}$ to be diagonal matrix with the q_i's on the diagonal. We define $\delta \in \mathbb{R}^{nT}$ to be the vector of δ_i's. Further, let $\mathbf{1}_{nT}$ denote the column vector of all ones in $\mathbb{R}^{nT}$. Lastly, we define $\text{diag}(S)$ to be the square matrix with the elements of S on the diagonal, and zeros everywhere else.

The optimization in (6) now becomes:

$$\min_{\substack{S \in \{0,1\}^{nT} \\ \bar{x} \in \mathbb{R}^{nT}}} \bar{x}^\top (Q + D)\bar{x} + \sigma^2 \text{trace}(DI_S) + \delta^\top S - 2\bar{y}^\top \bar{x}$$
$$\text{subject to} \quad (S_i - 1)\bar{x}_i = 0 \text{ for all } i = 1, ..., nT$$
$$P = (1 - \tilde{A})^{-1}$$
$$D = P^\top P$$
$$I_S = I - \text{diag}(S)$$
$$\tilde{A} = I_S \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ A & 0 & 0 & \dots & 0 & 0 \\ 0 & A & 0 & \dots & 0 & 0 \\ 0 & 0 & A & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A & 0 \end{bmatrix}$$

We note that for any matrix A and any S, the matrix $I - \tilde{A}$, with $\tilde{A}$ as defined above, is invertible, so P will always be well-defined.

Additionally, for a fixed S, the optimization across $\bar{x}$ is easy to solve. That is, D is entirely determined by S. If we let $(Q + D)_S$ denote the submatrix of $(Q + D)$ indexed by the non-zero elements of S, and similarly $\bar{x}_S$ and $\bar{y}_S$, then the optimizer is given by $\bar{x}_S^* = (Q + D)_S^{-1} \bar{y}_S$, with the other entries of $\bar{x}^*$ equal to 0.

Thus, we can easily calculate a set mapping $F(S)$ such that optimal causal imputation in the linear-Gaussian case is simply $\min_{S \subset V} F(S)$. We can solve this when nT is relatively small, and are currently investigating properties of $F(S)$ which would allow us to apply combinatorial optimization techniques [20].

V. CONCLUSION AND FUTURE WORK

The previous literature on mathematical formulations of causality has been focused on the *estimation* of causal structures. In this paper, we presented the problem of *control* of causal structures. We formally defined the problem of optimal causal imputation, and formulate solutions for it in two cases: where imputation is allowed to only a single value, and the case where the dynamics are linear and the noise is Gaussian.

In future work, we hope to apply this framework to real situations which allow both the estimation of causal structures, as well as verification of the consequences and costs of imputation. Additionally, we hope to generalize our results to consider dynamical systems whose behavior are influenced by different features. For example, we can consider the dynamics of the power grid, but also account for frequently used machine learning features as well, such as the zip code of different energy consumers and the age of deployed assets.

We believe that considering the control aspects of causality is increasingly more relevant. In many smart infrastructure applications, we no longer have control commands that directly affect the dynamics, but rather our control actions act more like causal imputations. The optimal causal imputation framework is a promising direction to model these interactions between machine learning and control, and provides a model for closing the loop on analytics in cyber-physical systems.

REFERENCES

[1] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66(5), pp. 688–701, 1974.
[2] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.
[3] S. Athey and G. W. Imbens, "Machine learning methods for estimating heterogeneous causal effects," *arXiv*, 2015.
[4] D. Zhou, M. Balandat, and C. J. Tomlin, "Residential demand response targeting using machine learning with observational data," in *55th IEEE Conference on Decision and Control (CDC)*, 2016.

[5] A. Banerjee, E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry, "A multifaceted program causes lasting progress for the very poor: Evidence from six countries," *Science*, vol. 348, no. 6236, 2015. [Online]. Available: http://science.sciencemag.org/content/348/6236/1260799

[6] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: http://www.jstor.org/stable/1912791

[7] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, "Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9849–9854, 06 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC470781/

[8] C. W. Granger, B.-N. Huang, and C.-W. Yang, "A bivariate causality between stock prices and exchange rates: Evidence from recent Asian flu," *The Quarterly Review of Economics and Finance*, vol. 40, no. 3, pp. 337 – 354, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1062976900000429

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

[10] C. Gourieroux, A. Monfort, and E. Renault, "Kullback causality measures," *Annales d'conomie et de Statistique*, no. 6/7, pp. 369–410, 1987. [Online]. Available: http://www.jstor.org/stable/20075662

[11] P. Amblard and O. J. J. Michel, "The relation between granger causality and directed information theory: a review," *CoRR*, vol. abs/1211.3169, 2012. [Online]. Available: http://arxiv.org/abs/1211.3169

[12] J. Pearl, *Graphical Models for Probabilistic and Causal Reasoning*. Dordrecht: Springer Netherlands, 1998, pp. 367–389.

[13] S. L. Lauritzen, "Causal inference from graphical models," 2001.

[14] J. Li, S. Ma, T. D. Le, L. Liu, and J. Liu, "Causal decision trees," *CoRR*, vol. abs/1508.03812, 2015. [Online]. Available: http://arxiv.org/abs/1508.03812

[15] S. Athey and G. Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *ArXiv e-prints*, Apr. 2015.

[16] A. P. Dawid, "Causal inference without counterfactuals," *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 407–424, 2000. [Online]. Available: http://www.jstor.org/stable/2669377

[17] D. Heckerman, C. Meek, and G. Cooper, *A Bayesian Approach to Causal Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–28.

[18] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and P. B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 689–696. [Online]. Available: http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf

[19] O. Kallenberg, *Foundations of Modern Probability*. Springer, 2002.

[20] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*. Springer-Verlag Berlin Heidelberg, 2003.

[21] L. Lovász, *Submodular functions and convexity*. Springer, 1983, pp. 235–257.

[22] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.