Causal Inference out of Control: The Steerability of Consumption

Anonymous Author(s)
Affiliation
Address
email

Abstract

Regulators and academics are increasingly interested in the causal effect that algorithmic actions of a digital platform have on consumption. We introduce a general causal inference problem we call the steerability of consumption that abstracts many settings of interest. Focusing on observational designs, we exhibit a set of assumptions for identifiability that significantly weakens the often unrealistic coverage assumptions of standard designs. They key insight behind our assumptions is to model the dynamics of consumption, viewing the platform as a controller acting on a dynamical system. From this dynamical systems perspective, we are able to show that exogenous variation in consumption and appropriately responsive control actions are sufficient for indentifying steerability of consumption. Our results illustrate the fruitful interplay of control theory and causal inference, which we illustrate with examples from econometrics, macroeconomics, and machine learning.

1 Introduction

How much does an increase in advertisement decrease screen time? Do algorithmic recommendations increase consumption of inflammatory content? Does exposure to diverse news sources mitigate political polarization? These are but a few of the questions that firms, researchers, and regulators alike ask about digital platforms. The answer to any such question requires non-trivial causal inference, since past consumption influences both future consumption as well as the algorithmic selection of content. Resolving confounding through randomization in the form of A/B tests is standard in the industry. However, randomization is often not possible. Experiments may be ethically fraught as past experience shows [Kramer et al., 2014], technically challenging to implement, or prohibitively expensive. Moreover, outside investigators may simply not have the power to experimentally intervene in the practices of a platform. Observational causal inference presents an intriguing alternative. Unfortunately, standard observational causal designs would require that the data satisfy a coverage assumption with respect to past consumption. Since user data on digital platforms is often high-dimensional, coverage is unlikely to hold [D'Amour et al., 2017] resulting in invalid inferences.

In this work, we study a general causal inference problem, called *steerability of consumption*, that unifies numerous questions about the causal effects of algorithmic actions on digital platforms. In our problem formulation, consumption evolves over time according to an unknown dynamical system. A platform aims to influence consumption through actions performed at each point in time. Our goal is to identify the effect that the action has on the state of the dynamical system. Taking this control theoretic perspective of the problem, we're able to exhibit a significantly weaker set of assumptions sufficient for identifiability and estimation. Informally speaking, it suffices that consumption has exogenous variation and that the control action of the platform is non-degenerate.

Submitted to *A causal view on dynamical systems* workshop at NeurIPS 2022. Do not distribute.

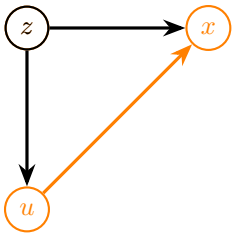

Figure 1: Standard model of the problem

Our results illustrate the benefits of taking a control-theoretic perspective on causal inference. Rather than omitting the role that time plays in causal inference, we explicitly keep track of a time index. This allows us for a more precise problem formulation that admits weaker identifiability assumptions. We illustrate the merits of our problem formulation with examples from econometrics, macroeconomics, and machine learning.

1.1 Our work

At the outset, we're interested in the causal effect of an action u on a variable x subject to observed confounding by a variable z. We interpret the action u as an algorithmic intervention of a firm in a digital platform, while the variable x captures relevant user features, such as what content the user consumed. Figure 1 presents the standard causal model for the scenario. The confounding variable captures all available information that influences both u and x. To apply observational causal inference machinery in the standard model, the data generating distribution must assign positive weight to all strata defined by the confounding variable z. In the common case where z is high-dimensional, this assumption is unlikely to hold [D'Amour et al., 2017].

The starting point of our work is therefore a departure from the standard model. Instead, we model a dynamical system where user features x evolves over time. Furthermore, we assume the control action u is not chosen once, but updated repeatedly based on the most recent observations of x. For example, video recommendations are updated based on recent views, and ads are chosen based on past clicks.

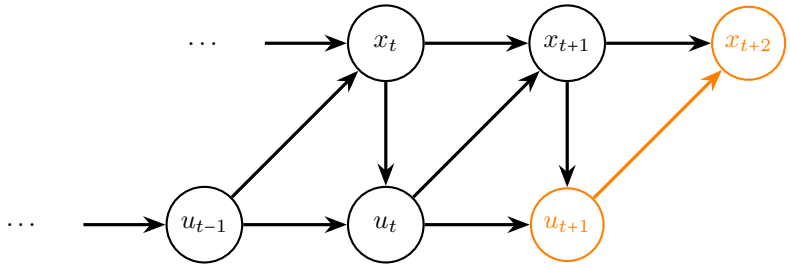

Figure 2: Autoregressive confounding structure

These assumptions result in the dynamical system illustrated in Figure 2. Technically speaking, the graph in Figure 2 could be expressed as an instance of the standard model in Figure 1 where the confounder z corresponds to the set of all past state and action variables. While this general model encapsulates our problem without loss of generality, it ignores the salient structure of the problem, resulting in overly strong assumptions. Our main contribution is to instead exploit this structure to make estimating the steerability of consumption from observational data possible under more plausible assumptions. Our technical contributions can be summarized as follows:

1. In Section 2 we discuss our refined structural model for studying the steerability of consumption in detail. The main feature of our model is that it explicitly models temporal dynamics and posits a structure on how previous states and actions interact. We provide several concrete examples of problems that can be approximated by our model.
2. In Section 3, we study identifiability of steerability of consumption within the context of our model. We demonstrate that exogenous variation in consumption and sufficient expressivity

in the control response enable causal identification, circumventing direct interventions on the control action. Usually, observational designs require strong coverage assumptions, especially when covariates are high-dimensional [D'Amour et al., 2017]; as a consequence of our structural assumptions, these coverage assumptions can greatly be weakened. In addition, we study a linear version of our causal model. We show that the auditor can benefit from the linear structure and thus requires even less exogenous consumption variation to identify the steerability of consumption.

3. In Section 4, we use a linear dynamics with Gaussian noise example to show the connection between the rank condition underlying our theory and identifiability. We find that the conditioning of our observations improves over time as more variations are accumulated, making the causal inference problem 'easier'.
4. To complement our study, in Appendix A, we provide a basic treatment of how practitioners could use our proposed observational model. In particular, we outline a non-parametric estimator of the steerability of consumption, and provide an associated convergence result for the general observational setting. For the linear setting we propose a Double-ML-type estimator [Chernozhukov et al., 2017] that can take advantage of the separable structure of the model and analyze it in finite samples.

1.2 Backgound

The fact that digital platforms, their predictions and actions non-trivially impact the individuals that interact with the platform has widely been recognized in diverse applications ranging from traffic forecasting, content recommendation to social media [c.f., Shmueli and Tafti, 2020, Thai et al., 2016, Fleder et al., 2010, Adomavicius et al., 2013].

Recently Hardt et al. [2022] have drawn a formal connection between the extent to which a platform can steer user behavior and the economic concept of power. Their proposal of *performative power* measures the strength of this effect and crucially relies on being able to estimate the steerability of consumption. From this lens, our work provides a roadmap for sufficient conditions for how performative power can be assessed from observational data. More broadly the implications of predictions impacting populations in the context of machine learning have been studied under the umbrella of performative prediction Perdomo et al. [2020]. Related to our work Mendler-Dünner et al. [2022] focus on estimating the causal effect of predictions on eventual outcomes from observational data. However their problem statement and the static model does not account for time varying effects and repeated interactions between the predictor and the population.

Our work draws a connection between the causal question of estimating steerability of consumption and dynamical system theory. The causal model we propose in this paper is reminiscent of dynamical systems in control theory. Abbasi-Yadkori et al. [2011], Abbasi-Yadkori and Szepesvari [2011] propose methods of controlling linear quadratic control systems with unknown dynamics. Their methods crucially rely on performing system identification—estimating the unknown parameters which describe the dynamics. Their observation model differs from the one we propose, as they focus only on the linear setting, and they opt to analyze one longer rollout of the dynamical system, instead of shorter, iid rollouts. While we give [Abbasi-Yadkori et al., 2011, Abbasi-Yadkori and Szepesvari, 2011] special attention because of the similarity of their model to ours, system identification, described more generally by Ljung [2010] as the "art and science of building mathematical models of dynamic systems from observed input-output data" is ubiquitous in control theory.

2 Model

As we highlighted in the introduction a feature of our model is that it unrolls time and makes the temporal dynamics of the problem explicit. Therefore, let $x_t \in \mathbb{R}^d$ and $u_t \in \mathbb{R}^p$ denote the demand and control action at time step t respectively. We assume for all $t \geq 0$ the dynamics of the system follow

$$\begin{aligned} x_t &= H_x(f(x_{t-1}), g(u_{t-1})) + \xi_t \\ u_t &= H_u(h(x_t), r(u_{t-1})) \end{aligned} \tag{1}$$

with $(u_{-1}, x_{-1}) \sim P_{-1}$ for a given distribution P_{-1}, describing the joint distribution over state and action pairs at time step $t = -1$. The state x_t is some measure of consumption, and $\xi_t \in \mathbb{R}^d$ will

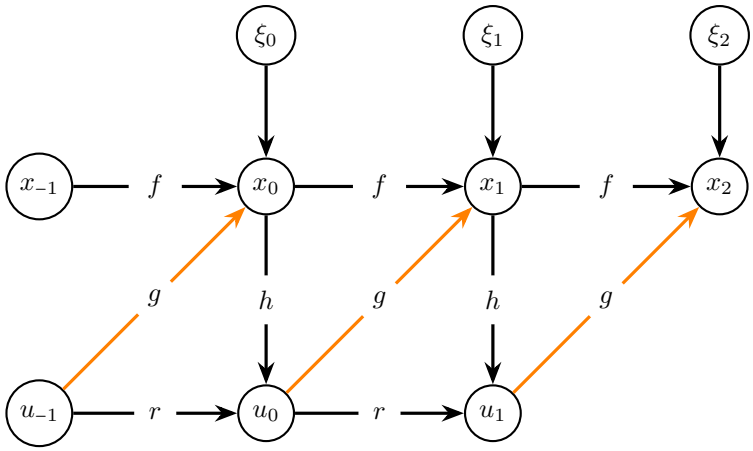

Figure 3: Model of dynamics

model the exogenous variation in x_t. We do not make any assumption on P_{-1} and ξ_t for now. We focus on time steps $t \geq 0$ without loss of generality. We will use the graph in Figure 3 as a visual representation of our dynamics.

The functions $f : \mathbb{R}^d \to \mathbb{R}^d$, $g : \mathbb{R}^p \to \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^p$, and $r : \mathbb{R}^p \to \mathbb{R}^p$ map the state and control variables to the reals, the aggregation functions $H_x : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, $H_u : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$ aggregate these quantities. The primary purpose of using H_x and H_u is to generalize the model beyond addition. In particular, we will consider aggregation functions which satisfy the following assumption (including e.g., addition or sigmoid composed element-wise with addition):

Assumption 1 (Reasonable Aggregation Function)**.** *$H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is a symmetric mapping (i.e., $H(a, b) = H(b, a)$ for all $a, b \in \mathbb{R}^d$ such that $H^a(b) := H(a, b)$ is a surjective, invertible mapping with an invertible Jacobian J^a.*

We assume the auditor observes x_t, u_t across multiple time steps. To describe the data available to the auditor we introduce the notation P_T denote the joint distribution over the state and preceeding control action $\{x_t, u_{t'} \mid t = 0, \ldots, T \quad t' = 0, \ldots, T-1\}$ for $t = 0, ..., T$. These variables are generated according to (1).

As we mentioned earlier, the auditor is interested in estimating the steerability of consumption from observational data. In this context of the model we outlined, we define this quantity of interest as follows:

Definition 2.1 (Steerability of Consumption)**.** *For the dynamical system specified by Equation* (1) *and Figure 3, the steerability of consumption at time t with respect to a set of control actions U and a norm $\|\cdot\|$ is defined as follows:*

$$\mathcal{S}_t(U) := \sup_{u, u' \in U} \|\mathbb{E}[x_t \mid \mathrm{do}(u_{t-1} := u)] - \mathbb{E}[x_t \mid \mathrm{do}(u_{t-1} := u')]\| .$$

This quantity captures the amount platforms can alter user behavior. We note that because we have a repeating structure in our graph (e.g., the relationship between x_t and u_{t-1} is the same as x_k and u_{k-1}), the steerability of consumption at time t is the same as that at time k.

2.1 Running example

Consider an auditor who interested in estimating the impact of YouTube's video recommendation on the consumption patterns of its users. Let $y_t \in \mathbb{R}^p$ be some measure of content consumption (number of hours streamed) for $p < d$ video categories of interest on YouTube during week t for a given user, and $z_t \in \mathbb{R}^{d-p}$ comprised of other measures of performance such as revenue per category, click-through rate per category, unique weekly users, unique advertisers per category, competitors' performance, etc. which could be confounders. We can think of the joint vector $[z_t^T, y_t^T]^T \in \mathbb{R}^d$ as the user attribute variable x_t. $u_t \in \mathbb{R}^p$ is a measure of how many videos from the d categories of interest are recommended to a given user during week t. The controller (YouTube) interfaces using

u_t with the goal of maximizing total profits, which is some deterministic function of x_t. The auditor observes is interested in estimating how the control action u_{t-1} impacts the average watch habits y_t of users , i.e., the steerability of consumption.

We make the natural assumption that user consumption changes over time based on the recommendations by the algorithm, as well as external factors e.g., new trends. Formally, taking inspiration from Jambor et al. [2012], we can model the dynamics of the system as

$$z_t = f_1(z_{t-1}, y_{t-1}) + g_1(u_{t-1}) + \xi_t^{(1)}$$
$$y_t = f_2(z_{t-1}, y_{t-1}) + g_2(u_{t-1}) + \xi_t^{(2)}.$$

f_1 models how the performance metrics chosen as a target by the firm evolve over time, and g_1 models YouTube's ability to control this metric over time. f_2 models how much interest users retain in each video category from week to week, as well as the effect of confounders on viewership (e.g., how many hours of viewing time can a competitors poach). g_2, the quantity the auditor wants to estimate, models how much consumption increases as more recommendations get served. $\xi_t^{(1)}, \xi_t^{(2)}$ is the natural variation in user preferences. For example, the Bitcoin price may increase due to changes in economic conditions, leading to many more users watching cryptocurrency videos on YouTube; this change in behavior is independent of past consumption and YouTube's recommendations. We can model the controller similarly as

$$u_t = h(z_t, y_t) + r(u_{t-1}).$$

h models the feedback loop of how viewer statistics and other metrics affect YouTube's recommendations in the future. r models how YouTube regularizes its recommendations to avoid overfitting to recent activity. For example, a sudden spike in viewership in the "music video" category could be the result of transitive internet trend; overfitting to this viewership spike may be suboptimal for the future. This example serves to illustrate how our model of user dynamics (1) maps to a concrete use-case.

Beyond recommender systems. The steerability of consumption is not a term specific to recommender systems; rather, it is a general term referring to the relationship control actions have on consumption. There are many other settings our model is applicable. Microeconomists are often interested in estimating the effect product prices have on demand, termed the *price elasticity of demand*. If we model product demand using x_t and model product prices using u_t, then this quantity is precisely the steerability of consumption. Confounders like product quality can be accounted for in the state variable x_t. A classical problem in macroeconomics is estimating the effect the Federal Interest Rate has on inflation and unemployment. We can use our framework to estimate this effect by modeling the Federal Interest Rate as the control action u_t and inflation and unemployment rates as the state x_t. In this example, GDP and other measures of the global economy could be possible confounders to account for. Many digital advertising platforms (and third-party auditors) are interested in whether personalized advertising increases platform activity. On one hand, advertisements clutter user interfaces, making the user experience less streamlined, but on the other hand, personalized advertisements provide users with more opportunities to engage, giving the platform more influence over user lives. Again, we can use our model to estimate the effect advertisements have on engagement; this time letting x_t be some measure of engagement (e.g., clicks, time online) and u_t be some measure of the type and quantity of ads served, while accounting for confounders like other platform performance measures such as monthly active users.

Remark 2.1. *In many settings, including the above examples, our model provides a reasonable first-order approximation to the problem of interest. Certainly, there might aspects of the system, such as unobserved state variables, or more complex dependencies across time that our model is not able to capture. However, we are interested in understanding how powerful it is to explicitly model time and interactions among confounding variables for tackling causal questions. Thereby, we aim to provide a mostly unexplored direction for designing assumptions for causal inference in data scarce settings and illustrate how to leverage data across multiple time steps for causal estimation.*

3 Identifiability from exogenous state variations

Let us start from the general causal graph in Figure 1 and recall classical results from causal identifiability in the presence of an observed confounding variable z [Pearl, 2009]. They tell us that a sufficient condition for identifiability of the causal effect of u on x is *admissability* and *coverage*.

Definition 3.1 (Admissibility [Pearl, 2009])**.** *We say that a continuous random variable Z with density p_Z is admissible with respect to treatment U and observation X if it satisfies the adjustment formula:*

$$\mathbb{E}[X \mid \mathrm{do}(U := u)] = \int_z \mathbb{E}[X \mid U = u, Z = z] p_Z(z)\, dz. \tag{2}$$

Definition 3.2 (Full Coverage)**.** *A continuous, $\mathbb{R}^d$-valued random variable W with density p has full coverage if $p(w) > 0$ for all $w \in \mathbb{R}^d$.*

Admissability is a property of a set of random variables Z that ensures the adjustment formula (2) is a valid tool for adjusting for confounding. It states that the causal quantity on the left-hand-side can be expressed as a function of pre-interventional data, and thus can be estimated from such data without bias. Admissibility coupled together with full coverage over treatment, confounder pairs—ensuring that $p(u, z) > 0$ for any pair u, z of interest—guarantees that $\mathbb{E}[X|U = u, Z = z]$ is well defined for any u, z and thereby allows the auditor to estimate steerability of consumption in a non-parametric way by adjusting for the confounder z using the adjustment formula (2).

To apply the adjustment formula in the context of our refined causal model (Figure 3) we aim to find the smallest set of admissible random variables for adjustment and estimation of the steerability of consumption. Without loss of generality we focus on identifying the steerability of consumption $\mathcal{S}_t$ for $t = 2$. Thus we are interested in the quantity $\mathbb{E}[x_2 \mid \mathrm{do}(u_1 := u)]$ from which we can be derived S_2. The proof of the following result is found in Appendix B.1.

Proposition 1 (Admissibility of our model)**.** *Given the causal model in Figure 3. The state variable x_1 is admissible with respect to treatment u_1 and observation x_2.*

Given Proposition 1 and the adjustment formula (2) we know that $\mathbb{E}[x_2 \mid \mathrm{do}(u_1 := u)] = \int_z \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz$. Hence, being able to compute the right hand side for any u from observational data is a sufficient condition for identifiability. This is equivalent to requiring the joint density of (x_1, u_1) being positive everywhere; the question of identifiability reduces to a question of coverage.

3.1 Key assumptions

We highlights the two assumptions on the dynamical system in (1) that allow us to establish coverage over (x_1, u_1). The first assumption posits the presence of natural perturbations in the state variable x. In particular, it requires that there is exogenous noise in the system that leads to sufficient variation in the user attribute variables across time.

Assumption 2 (Exogenous Noise)**.** *$(\xi_0, \xi_1)|\{x_{-1} = a, u_{-1} = b\}$ has density and has joint full coverage for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^p$.*

We note that the noise variables ξ_0, ξ_1 do not need to be independent of each other or of the previous states, all we need is full coverage. The second assumption concerns the controller. It needs to be sufficiently sensitive to the variations in the user attribute variable x, so that the variations provided by Assumption 2 are guaranteed to be propagated into the control action u.

Assumption 3 (Responsive Control Action)**.** *Let $q_c : \mathbb{R}^d \to \mathbb{R}^p$ defined as $q_c(y) := r(H_u(h(y), c))$ describe how the current state y affects the next control action given that the previous control action was c. Assume q_c is a surjective, invertible map with an invertible Jacobian for all $c \in \mathbb{R}^p$.*

To illustrate why one might expect the above two assumptions to hold, consider the YouTube recommender system example from Section 2.1. Exogenous perturbations ξ_t can correspond to factors like new trends or economic conditions which often have a very large effect on viewership statistics and other metrics captured in the user attribute variable x_t, giving us reason to believe that Assumption 2 is satisified in that setting. To argue for surjectivity, recall that h describes the effect previous control actions have on future recommendations, and r captures the effect the state has on the control action. First, we don't expect the control action to vary wildly between time steps, so we may expect the controller to adopt a simple policy, like discounting the previous action $h(u) = \alpha u$. This action happens to be surjective. Furthermore, in general, the dimensionality of the state is larger than the dimensionality of the control system (the controller has limited actions that it can take to influence the system and there are an arbitrarily large number of metrics and confounders we can add to the state variable). Thus, $d \geq p$, and it seems reasonable to expect that r is surjective. Thus, q_c, which

is the composition of r and h is also surjective. The invertible Jacobian condition imposes a form of "monotonicity" on q_c which in the YouTube setting corresponds to: more views in category i should lead to more recommendations in category i.

3.2 General identifiability result

Building on the two assumptions discussed in the previous section, we can now provide our general identifiability results. Proofs can be found in Appendix C.

Theorem 1 (Identifiability from two exogenous state perturbations)**.** *Consider the dynamical system in* (1) *with any arbitrary P_{-1} that has a density and with H_u satisfying Assumption 1. Let the auditor receive iid samples from P_2. Then, if Assumption 2 and Assumption 3 hold, (x_1, u_1) has full coverage i.e., $p(x_1 = a, u_2 = b) > 0$ for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^p$, and the steerabiltiy of consumption is identifiable for any U.*

In words, this result states that independent perturbations on two of the state variables are sufficient for the auditor to identify the steerability of consumption from observations. Interestingly, observing the system after two independent perturbations to the state is crucial for identifiability. Observing only one perturbation is not sufficient. We show this necessity by presenting Theorem 2, which shows that identifiability is not guaranteed by only getting samples from P_1.

Theorem 2 (Unidentifiabilty of one exogenous state perturbation)**.** *Consider the dynamical system in* (1) *with any arbitrary P_{-1} that has a density and with $H_x(a, b) \coloneqq a + b$. Let the auditor receive samples from P_1. Then, for any functions $f : \mathbb{R}^d \to \mathbb{R}^d, g : \mathbb{R}^p \to \mathbb{R}^d, h : \mathbb{R}^d \to \mathbb{R}^p$, and independent noise variables ξ_0 and ξ_1. If $x_{-1} = u_{-1} = 0$ steerability of of consumption is unidentifiable; i.e., there exists $(\hat{f}, \hat{g}, \hat{h}) \neq (f, g, h)$ which induces the same distribution P_1.*

Intuitively, a single perturbation is not enough for identifiability because the control action is a deterministic function of the user attribute variable. Thus, the auditor is only able to see one value of u_0 for any value of x_0. The second noise spike impacts the state and control variables differently, providing enough variation for full coverage making the steerability of consumption identifiable.

If the auditor believes there is sufficient variation in at least two time steps of the user attributes and the control response is appropriately sensitive to variations in the state, then our results state that the auditor can identify the steerability of consumption. Unlike observational designs which purely rely on coverage assumptions, one advantage of our assumptions is that Assumption 3 is an assumption on the design of the controller (which is typically a deterministic mechanism) and this assumption can be verified with enough knowledge of the control system, while coverage assumptions are not, as they model the behaviors of users, which could vary wildly for any number of reasons.

3.3 Identifiability in the linear model

In this section we outline how additional structural assumptions can be exploited for identification. In particular, we study a linear instantiation of the general model we proposed and show that one noise spike, instead of two, is sufficient for identifiability as long as the auditor observes the system for enough times steps. We provide some intuition why this is the case. Because the system is linear, explicit full coverage of (x_t, u_t) is not needed; instead the auditor just needs coverage over enough directions such that their span covers all values that (x_t, u_t) can take on. One spike and sufficient expressivity in the linear model ensures that observations of (x_0, u_0), and (x_1, u_1) provide "one direction" each so that the auditor can use linear combinations of these observations to recover the steerability of consumption.

In Section 3.3, we will assume the aggregation function is addition, $H_x(a, b) \coloneqq a + b$, and the functions f, g, h, r are linear and defined as

$$\begin{aligned} f(x) &\coloneqq Ax \qquad & g(x) &\coloneqq Bx \\ h(x) &\coloneqq Cx \qquad & r(x) &\coloneqq Dx, \end{aligned} \tag{3}$$

where $A \in \mathbb{R}^{d,d}, B \in \mathbb{R}^{d,p}, C \in \mathbb{R}^{p,d}, D \in \mathbb{R}^{p,p}$. In this model, for identifying steerability of consumption it is sufficient for the auditor to identify the matrix B. Once the auditor identifies B, they can readily compute the steerability of consumption in the linear model. There are many settings where the dynamical sytem we want to study is linear, or modeled as linear. For example, a linear

controller, such as a PI controller, controlling a physical system, such as a quadrotor, can be modeled with linear dynamics [Bouabdallah et al., 2004].

Before we present our results, we begin with an alternative definition of coverage that we will use in this section, termed "full-span". Intuitively, because we are dealing with linear systems, the auditor just needs to be able to see or reconstruct all the directions the data could span, as opposed to observing the full space.

Definition 3.3 (Full-Span)**.** *A random vector $x \in \mathbb{R}^d$ has full span if for all vectors $a \in \mathbb{R}^d$ such that $a \neq 0$, $a^T x$ is almost surely not a constant.*

With these supporting results, we present our first result in Theorem 3; the proof of Theorem 3 is found in Appendix C.3. This result shows that observing five nodes x_0, x_1, x_2 and u_0, u_1 is sufficient to identify the steerability of consumption.

Theorem 3 (Simple identifiable DAG)**.** *Consider the dynamical system in* (1) *with any arbitrary P_{-1}, with $H_x(a,b) \coloneqq a + b$, and with functions f, g, h, r defined in* (3)*. Let the auditor receive samples from P_K for $K \geq 2$. Let $\{\xi_t\}_{t\geq 0}$ be a set of mutually independent random vectors, and let ξ_0 have full span. If x_{-1} and u_{-1} are mutually independent of ξ_0, then B is uniquely identifiable and steerability of consumption can be estimated for any set U if $[DC, \ldots, D^{K-1}C]$ is full row rank. If further $x_{-1} = 0$, $u_{-1} = 0$, and $\xi_t = 0$ for $t \geq 1$, then $[DC, \ldots, D^{K-1}C]$ being full row rank is also necessary for the identifiabilty of B.*

We note that because Theorem 2 still holds even in the linear setting, Theorem 2 and Theorem 3 together fully characterize the tradeoff between identifiability, number of time steps observed, and rank conditions on the controller dynamics matrices. Moreover, in linear control problems like the aforementioned quadrotor example, the controller dynamics are often known, and thus, the rank conditions sufficient for identifiability are frequently easy to check.

4 Simulations

For our empirical investigation we analyze the dynamical system (1) with linear dynamics (3) with independent Gaussian noise on the states to show how the noise leads to identifiability. In particular we show how the conditioning of the problem evolves across time steps and how the presence of noise makes the problem easier.

We instantiate the model as follows: We let $\xi_t \overset{d}{=} \mathsf{N}(0, I)$ for all t, starting from $x_{-1} = u_{-1} = 0$. We consider the symmetric case where $d = p$. To generate B, we sample a random matrix W in $\mathbb{R}^{d,n}$ for $n \gg d$ with independent standard Gaussians as its entries, and we set $B = WW^T/n$. We repeat this process to generated A and D. This way of generating our dynamics matrices ensures the matrices are well conditioned. We generate C the same except by instead setting $W \in \mathbb{R}^{d,r}$ for $r < d$, making C rank r instead of rank d. We set $d = 100$, $n = 2000$, and $r = 80$.

For this system, we can explicitly write down how the covariance matrix of (x_t, u_t), denoted Σ_t, evolves. Namely, from the dymanics

$$\begin{bmatrix} x_t \\ u_t \end{bmatrix} = J \begin{bmatrix} x_{t-1} \\ u_{t-1} \end{bmatrix} + M\varepsilon_t \quad \text{where} \quad J \coloneqq \begin{bmatrix} A & B \\ CA & CB + D \end{bmatrix} \qquad M \coloneqq \begin{bmatrix} I \\ C \end{bmatrix}.$$

we can decipher that

$$\Sigma_t = J\Sigma_{t-1}J^T + MM^T = \sum_{k=0}^{t-1} (J^k) MM^T (J^k)^T.$$

We now plot the histogram of the eigenvalues of Σ_t for a random system that we generated. The important quantity to look out for is whether Σ_t is full rank. Directions in the null space of Σ_t (i.e., with zero eigenvalue) are directions that are unobserved by the auditor. Thus, when Σ_t becomes full rank, the system becomes identifiable because (x_t, u_t) becomes fully-spanning.

We note that in this system, $\operatorname{rank} DC = 80 = r < d$ and $\operatorname{rank} C = 80 = r < d$. Even thought the setup of this system is not the same as Theorem 3, linearity gives us a correspondence between the settings. The theorem tells us that observing (x_1, u_1) is not sufficient for identifiability. This is consistent with Figure 4 as there are still 0 eigenvalues. However, since in this system $\operatorname{rank}[DC, D^2C] =$

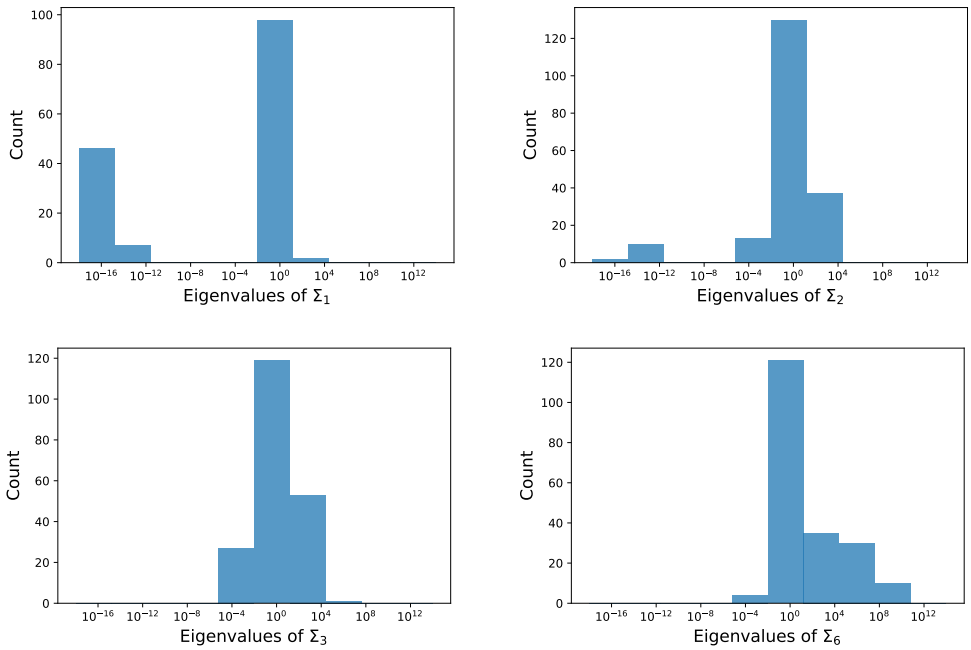

Figure 4: Histograms of eigenvalues of Σ_t defined in Section 4

$100 = d$, Theorem 3 tells us observing (x_2, u_2) (i.e., observing three noise spikes) is sufficient for identifiability. This is also consistent with Figure 4 as all eigenvalues are bounded away from 0 now. Moreover, we see that the eigenvalues of get larger as more time passes: the eigenvalue mass of Σ_6 is further to the right of the eigenvalue mass of Σ_3 in Figure 4.

Our experiments corroborate our theoretical results which say that three noise spikes for this low rank system is necessary and sufficient for identifiability. Furthermore, our experiment also suggests that more noise spikes over more time steps make the observations better conditioned, likely making estimating the steerability of consumption easier for the auditor in practice.

For future work we plan to investigate steerability of consumption in a practical application based on real data. We will compare the estimates obtained by relying on our modeling assumptions with an estimate obtained from a more naive approach based on the general model from Figure 1 that does not posit a Markovian assumption on how confounding variables interact. Such a study would help us demonstrate the utility of our model for reducing reliance on data coverage for estimating steerability of consumption in practice.

References

Yasin Abbasi-Yadkori and Csaba Szepesvari. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013.

S. Bouabdallah, A. Noth, and R. Siegwart. Pid vs lq control techniques applied to an indoor micro quadrotor. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, pages 2451–2456 vol.3, 2004. doi: 10.1109/IROS.2004.1389776.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James M. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics: Econometric & Statistical Methods - Special Topics eJournal*, 2017.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet S. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 2017.

Daniel Fleder, Kartik Hosanagar, and andreas buja. Recommender systems and their effects on consumers: the fragmentation debate. 06 2010.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *ArXiv*, abs/2203.17232, 2022.

Tamas Jambor, Jun Wang, and Neal Lathia. Using control theory for stable and efficient recommender systems. *Proceedings of the 21st international conference on World Wide Web*, 2012.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *ArXiv*, abs/1902.03736, 2019.

Adam D. I. Kramer, Jamie Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111 24:8788–90, 2014.

Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, 2010. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2009.12.001. URL `https://www.sciencedirect.com/science/article/pii/S1367578810000027`.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Predicting from predictions. *ArXiv*, abs/2208.07331, 2022. doi: 10.48550/ARXIV.2208.07331. URL `https://arxiv.org/abs/2208.07331`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7599–7609. PMLR, 2020.

Galit Shmueli and Ali Tafti. "Improving" prediction of human behavior using behavior modification. *Arxiv:2008.12138*, 2020.

Jérôme Thai, Nicolas Laurent-Brouty, and Alexandre M. Bayen. Negative externalities of gps-enabled routing applications: A game theoretical approach. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 595–601, 2016. doi: 10.1109/ITSC.2016.7795614.

A Estimators

To complement our identifiability results from the previous section, we provide a basic treatment of how practitioners could estimate steerability of consumption from ovbservational data under our proposed model. First, we outline a non-parametric estimator based on the adjustment formula and provide an associated convergence result for the non-parametric setting. Then, we propose a Double-ML-type estimator [Chernozhukov et al., 2017] that can exploit the separable structure of the model and analyze it in the linear model.

A.1 Adjustment formula estimator

Admissibility of the dynamical system we are studying (Proposition 1) makes estimating the adjustment formula (Definition 3.1) sufficient for estimating the steerability of consumption. To estimate this quantity by first selecting a compact subset of the user state and control spaces $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{U} \subset \mathbb{R}^p$ respectively and then discretizing the space using a finite cover, defined as follows.

Definition A.1 (Cover)**.** *For a space $\mathcal{X}$, let $\mathcal{N} = \{U_\alpha \subset \mathcal{X} : \alpha \in A\}$ is an indexed family of sets U_α. $\mathcal{N}$ is a cover of $\mathcal{X}$ if $\mathcal{X} \subset \cup_{U \in \mathcal{N}} U$. $\mathcal{N}$ is an ε-cover with respect to $\|\cdot\|_2$ if for all $U \in \mathcal{N}$ for all $x, y \in U$, $\|x - y\|_2 \leq \varepsilon$. $\mathcal{N}$ is a disjoint cover if for all $U, V \in \mathcal{N}$, $U \cap V = \varnothing$.*

Definition A.2 (Cover Representatives)**.** *For a space $\mathcal{X}$ and a disjoint cover $\mathcal{N}$, let $\mathcal{R} \subset \mathcal{X}$ such that for all $A \in \mathcal{N}$, there exists $x \in \mathcal{R}$ such that $x \in A$ and for all $y \in \mathcal{R}$ where $y \neq x$, $y \notin A$.*

With this finite discretization, we can estimate the integral from Definition 3.1 using a finite sum. The specifics of how the estimator is constructed can be found in Algorithm 1.

Algorithm 1 Adjustment formula estimator

Require: Cover granularity $\varepsilon > 0$, Margins of error $\delta \in (0,1)$, $\gamma > 0$, probability of failure $\rho \in (0,1)$, compact subset of control space $\mathcal{U} \subset \mathbb{R}^p$, compact subset of the state space $\mathcal{X} \subset \mathbb{R}^d$, control action $u \in \mathcal{U}$ of interest.

Step 1. Form a disjoint $\varepsilon/2$-cover of $\mathcal{U}$ such that every element has Lebesgue measure greater than 0, and label it $\mathcal{N}_\mathcal{U}$; let $\mathcal{R}_\mathcal{U}$ denote its respective cover representative. Form a disjoint $\varepsilon/2$-cover of $\mathcal{X}$ such that every element has Lebesgue measure greater than 0, and label it $\mathcal{N}_\mathcal{X}$; let $\mathcal{R}_\mathcal{X}$ denote its respective cover representative.

Step 2. Collect enough samples from P_2 such that $n_{u,x} \geq \frac{2d\sigma^2}{\gamma^2} \log(4|\mathcal{R}_\mathcal{X}|/\rho)$ for all $x \in \mathcal{X}$ and the total number of samples n satisfies $n \geq \max_{x \in \mathcal{R}_\mathcal{X}} \frac{1}{2\delta^2 P(x_1 \in \mathcal{N}_\mathcal{X}(x))^2} \log(4|\mathcal{R}_\mathcal{X}|/\rho)$. Take data to form dataset $\mathcal{D}^n = \{(x_2^{(k)}, u_1^{(k)}, x_1^{(k)})\}_{k=1}^n$, $u \in \mathbb{R}^p$.

Step 3. Let

$$Y^n(u,x) \coloneqq \frac{\sum_{k \in [n]} x_2^{(k)} \mathbf{1}\left\{u_1^{(k)} \in \mathcal{N}_\mathcal{U}(u), x_1^{(k)} \in \mathcal{N}_\mathcal{X}(x)\right\}}{\sum_{k \in [n]} \mathbf{1}\left\{u_1^{(k)} \in \mathcal{N}_\mathcal{U}(u), x_1^{(k)} \in \mathcal{N}_\mathcal{X}(x)\right\}}$$

$$Z^n(x) \coloneqq \frac{1}{n} \sum_{k=1}^n \mathbf{1}\left\{x_1^{(k)} \in \mathcal{N}_\mathcal{X}(x)\right\}.$$

return the following estimator:

$$\hat{x}_2 \coloneqq \sum_{x \in \mathcal{R}_\mathcal{X}} Y^n(u,x) Z^n(x).$$

We will need some mild assumptions to prove a guarantee on the estimator. Our first assumption controls how much previous user state and control actions affect future state actions. The magnitude of the effect must be bounded in proportion to the inputs.

Assumption 4. *The relationship between x_2 and x_1, u_1 is L-Lipschitz continuous in the sense that for any $z, z' \in \mathcal{X}$ and $u, u' \in \mathcal{U}$ it holds that*

$$\|\mathbb{E}[x_2 | u_1 = u, x_1 = z] - \mathbb{E}[x_2 | u_1 = u', x_1 = z']\| \leq L \left\| \begin{bmatrix} u \\ z \end{bmatrix} - \begin{bmatrix} u' \\ z' \end{bmatrix} \right\|.$$

We also need to control how far the empirical estimate $\mathbb{E}[Y^n(u,x)]$ deviates from $\mathbb{E}[x_2|u_1 = u, x_1 = x]$. To do this, we impose a regularity condition on the conditional distribution.

Assumption 5. *Let p denote the density of P_2. For all $w, w' \in \mathcal{X}$ and $u, u' \in \mathcal{U}$, for which*

$$\left\| \begin{bmatrix} u \\ w \end{bmatrix} - \begin{bmatrix} u' \\ w' \end{bmatrix} \right\| \leq \varepsilon,$$

and for any $x \in \mathcal{X}$ it holds that $p(u_1 = u, x_1 = w|x_2 = x) \leq (1 + \eta(\varepsilon))p(u_1 = u', x_1 = w'|x_2 = x)$ and $p(u_1 = u, x_1 = w|x_2 = x) \geq (1 - \eta(\varepsilon))p(u_1 = u', x_1 = w'|x_2 = x)$ for some $\eta(\varepsilon) \in (0, 1)$. Further, $\lim_{\varepsilon \to 0} \eta(\varepsilon) = 0$.

Finally, we will also assume that the aggregation function we are interested in satisfies a form of triangle inequality. The assumption holds for addition, for example, with $c_0 = 1$.

Assumption 6. *For all $a, b \in \mathbb{R}^d$, there exists $c_0 > 0$ such that*

$$\|H_x(a, b)\| \leq c_0 \|a\| + c_0 \|b\|.$$

We will also need to introduce some notation for our result. We let $y^{(k)}$ denote the kth sample of the random variable y; note that $x_t^{(k)}$ and $u_t^{(k)}$ are assumed to satisfy the relationship specified by the structural equation (1). We define $\mathbb{B}(u, \varepsilon)$ to be the $\|\cdot\|$-ball centered at u with radius ε. For a cover $\mathcal{N}$, we let $\mathcal{N}(u) \in \mathcal{N}$ denote the element such that $u \in \mathcal{N}(u)$. We define $n_{u,x} := \sum_{k \in [n]} \mathbf{1}\left\{u_1^{(k)} \in \mathcal{N}_{\mathcal{U}}(u), x_1^{(k)} \in \mathcal{N}_{\mathcal{X}}(x)\right\}$. We present our convergence result now in Theorem 1.

Theorem 4. *Consider the dynamical system* (1) *with any arbitrary P_{-1}. Let the auditor receive iid samples from P_2. Suppose x_2 is σ^2-subgaussian conditioned on u_1 and x_1. Let $\mathbb{E}[\xi_2|x_1 = x, u_1 = w] = 0$, $\mathbb{E}[\|\xi_2\| \mid x_1 = x, u_1 = w] \leq c_1$ for all $x \in \mathbb{R}^d$ and $w \in \mathbb{R}^p$. Let f and g be continuous functions, and define $D \in \mathbb{R}$ such that $\sup_{x \in \mathcal{X}, w \in \mathbb{B}(u,\varepsilon)} \max(\|f(x)\|, \|g(w)\|) \leq D$; $D < \infty$ because of compactness. Let the conditions of Theorem 1 hold, Assumption 4 hold with L, Assumption 5 hold with η, and Assumption 6 hold with c_0. For any compact subset $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{U} \subset \mathbb{R}^p$, for any specified $u \in \mathcal{U}$, for n large enough, Algorithm 1 executed with parameters $\varepsilon, \delta, \gamma, \rho, \mathcal{U}, \mathcal{X}, u$ returns $\hat{x}_2$ which has the following guarantee with probability at least $1 - \rho$:*

$$\begin{aligned} \|\hat{x}_2 - \mathbb{E}[x_2|do(u_1 := u)]\| \leq \delta\gamma + 2\delta c_0 D + \gamma + \frac{2\eta(\varepsilon)}{1 - \eta(\varepsilon)} (c_0 D + c_0 \|g(u)\| + c_1) \\ + L\varepsilon + c_0 \mathbb{E}[\|f(x_1)\| \mathbf{1}\{x_1 \notin \mathcal{X}\}] + c_0(1 - P_{x_1}(\mathcal{X})) \|g(u)\|. \end{aligned}$$

The proof of Theorem 4 can be found in Appendix D.1. We quickly go through all the terms in the bound, and verify that they can all be made arbitrarily small (with sufficient samples). δ and γ can be made smaller, so long as the auditor receives proportionally enough samples. The auditor can create a finer discretization to make ε smaller and therefore η smaller as well. If we assume that $\mathbb{E}[\|f(x_1)\|] \leq \infty$, then the last two terms tend to zero as the auditor's approximation of $\mathbb{R}^d$, i.e., $\mathcal{X}$, becomes larger.

A.2 Double-ML-type estimator for separable model

Now we present an estimator which leverages the structure of the data generation procedure and performs a regression between residuals, reminiscent of Double Machine Learning Chernozhukov et al. [2017]. The method assumes the aggregation functions H_x and H_u are the addition operation. The procedure requires hypotheses (i.e., sets of functions $\mathcal{F}^{d \to p}$, $\mathcal{F}^{d \to d}$, and $\mathcal{F}^{p \to d}$) of what the relationships between control and state are. The details of the estimator are presented in Algorithm 2.

We now present some intuition about the estimator by analyzing it in the linear setting from Section 3.3. Observe that the following relationship holds between control and state variables:

$$x_2 - (A + BC)x_1 = B(u_1 - Cx_1).$$

Thus, if we had good estimates of $H := (A + BC)$ and C denoted by $\hat{H}$ and $\hat{C}$ respectively, we could regress $x_2 - \hat{H}x_1$ against $u_1 - \hat{C}x_1$ to get an estimate of the steerability of consumption B. In fact, this is exactly what Algorithm 2 is doing in the linear setting. Let $x_t^{(k)}, u_t^{(k)}, \xi_t^{(k)}$ denote the

Algorithm 2 Double ML estimator

Require: Loss functions ℓ_1, ℓ_2, ℓ_3, Hypothesis classes $\mathcal{F}^{d\to p} \subset \{f | f : \mathbb{R}^d \to \mathbb{R}^p\}$, $\mathcal{F}^{d\to d} \subset \{f | f : \mathbb{R}^d \to \mathbb{R}^d\}$, and $\mathcal{F}^{p\to d} \subset \{f | f : \mathbb{R}^p \to \mathbb{R}^d\}$, Dataset $\mathcal{D}^n = \{(x_2^{(k)}, u_1^{(k)}, x_1^{(k)})\}_{k=1}^n$

Step 1.

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{F}^{d\to p}} \frac{1}{n} \sum_{k=1}^{n} \ell_1(u_0^{(k)}, h(x_0^{(k)}))$$

Step 2.

$$\hat{s} = \operatorname*{argmin}_{s \in \mathcal{F}^{d\to d}} \frac{1}{n} \sum_{k=1}^{n} \ell_2(x_1^{(k)}, s(x_0^{(k)}))$$

Step 3.

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{F}^{p\to d}} \frac{1}{n} \sum_{k=1}^{n} \ell_3(x_2^{(k)} - \hat{s}(x_1^{(k)}), g(u_1^{(k)} - \hat{h}(x_1^{(k)})))$$

return $\hat{g} : \mathbb{R}^p \to \mathbb{R}^d$

kth observations of x_t, u_t, and ξ_t respectively. Let $X_t \in \mathbb{R}^{d,n}$, $U_t \in \mathbb{R}^{p,n}$, and $E_t \in \mathbb{R}^{d,n}$ be matrices that comprise n samples of x_t, u_t, and ξ_t respectively. Algorithm 2 in the linear setting corresponds to first estimating H and C first with the values $\hat{H}$ and $\hat{C}$ respectively before computing an estimate of the steerability of consumption $\hat{B}$ as:

$$\begin{aligned}\widehat{B} &= \operatorname*{argmin}_{B \in \mathbb{R}^{d,p}} \frac{1}{2} \left\| X_2 - \widehat{H} X_1 - B(U_1 - \widehat{C} X_1) \right\|_{\mathrm{Fr}}^2 \\ &= (X_2 - \widehat{H} X_1)(U_1 - \widehat{C} X_1)^T ((U_1 - \widehat{C} X_1)(U_1 - \widehat{C} X_1)^T)^{-1}.\end{aligned}$$

We now present a convergence result about the Double-ML estimator in a restricted setting where we assume the relationships between controls and states are linear and that $x_{-1} = u_{-1} = 0$. Our results can be generalized further, but we feel these generalizations only obfuscate the main message; thus, we focus on this simplified setting. Our result will need the following assumption to hold.

Assumption 7 (ρ-Bounded System Dynamics)**.** *The linear dynamical system specified by* (1) *and* (3) *has ρ-Bounded System Dynamics if*

$$\frac{\|A + BC\|_{\mathrm{op}}}{\sigma_{\min}(DC)} \le \rho$$

Let's investigate the consequences of this assumption. First, consider the quantity $\|x_1 - \xi_1\|_2 / \|u_1\|_2$; this is the ratio between the magnitude of the state and control action after one time step of dynamics ignoring noise and assuming the system starts from equilibrium $x_{-1} = u_{-1} = 0$. Because $\|x_1 - \xi_1\|_2 / \|u_1\|_2 \le \frac{\|A+BC\|_{\mathrm{op}} \|x_0\|_2}{\lambda_{\min}(DC) \|x_0\|_2}$, having Bounded System Dynamics ensures that we have a bound on the relative magnitude between state and control action. We will also use the notation $\kappa_A \coloneqq \sigma_{\max}(A)/\sigma_{\min}(A)$ to denote the condition number of a matrix A and $\hat{\Sigma}_0 \coloneqq \frac{1}{n} \sum_{k=1}^n \xi_0^{(k)} (\xi_0^{(k)})^T$ to denote the sample covariance of the first noise spike. We now provide our convergence result for the Double ML-type estimator in Theorem 5; the proof can be found in Appendix D.2.

Theorem 5. *Consider the dynamical system in* (1) *with $x_{-1} = u_{-1} = 0$, with $H_x(a, b) \coloneqq a + b$, and with functions f, g, h, r defined in* (3)*. Suppose n samples are drawn iid from P_2, and suppose DC is full column rank. Suppose that $\mathbb{E}\|\xi_1\|_2^2 = \sigma_1^2 d$, and the data generation model has ρ-Bounded System Dynamics. Let $\mathcal{G}$ denote the event where $X_0 X_0^T$ is invertable. If $\mathbb{E}\left[\frac{\kappa_{\hat{\Sigma}_0}^2}{\lambda_{\min}(\hat{\Sigma}_0)}\right] \le \tau_0$, then*

$$\frac{1}{pd} \mathbb{E}\left[\left\| \widehat{B} - B \right\|_{\mathrm{Fr}}^2 \mid \mathcal{G} \right] \le \frac{\sigma_1^2 \rho^2 \kappa_{DC}^2 \tau}{n}.$$

In the case where $p = d$, if $\left\|\mathbb{E}\left[\hat{\Sigma}_0^{-1}\right]\right\|_{\mathrm{op}} \leq \tau_1$ we have that

$$\frac{1}{d^2}\mathbb{E}\left[\left\|\widehat{B} - B\right\|_{\mathrm{Fr}}^2 \mid \mathcal{G}\right] \leq \frac{\sigma_1^2\rho^2\tau_1}{n}.$$

We note that rank condition on DC in this result is consistent with the rank condition from the identifiability result in the linear setting (Theorem 3); both results require DC to be full rank for identifiability. The τ_0, τ_1 conditions are a bit technical, but they essentially just require ξ_0 to be well behaved. We provide a simple Gaussian noise example to better illustrate.

Example 1 (Gaussian noise)**:** Suppose ξ_0 and ξ_1 are drawn iid from $\mathsf{N}(0, \sigma_0^2 I_d)$ and $p = d$. We have $\mathbb{E}\left\|\xi_1\right\|_2^2 = \sigma_1^2 d$. For $n \geq d$, $\hat{\Sigma}_0$ is invertible a.s. $(X_0 X_0^T/\sigma_0^2)^{-1}$ has an inverse Wishart distribution and thus, $\mathbb{E}[\hat{\Sigma}_0^{-1}] = \frac{n}{(n-d-1)\sigma_0^2} I_d$ for $n > d+1$. Theorem 5 gives us

$$\frac{1}{d^2}\mathbb{E}\left\|\widehat{B} - B\right\|_{\mathrm{Fr}}^2 \leq \frac{\sigma_1^2\rho^2}{(n-d-1)\sigma_0^2}.$$

$\diamond$

B Proofs of Section 2

B.1 Proof of Proposition 1

Recall that the do action alters the data generation model by deleting incoming edges into u_1.

$$\begin{aligned}
\mathbb{E}[x_2|do(u_1 := u)] &= \mathbb{E}[H_x(f(x_1, w_1), g(u, v_1)) + \xi_t] \\
&= \int_z \mathbb{E}[H_x(f(z, w_1), g(u, v_1)) + \xi_t] p_{x_1}(z) dz \\
&= \int_z \mathbb{E}[H_x(f(x_1, w_1), g(u_1, v_1)) + \xi_t | u_1 = u, x_1 = z] p_{x_1}(z) dz \\
&= \int_z \mathbb{E}[x_2 | u_1 = u, x_1 = z] p_{x_1}(z) dz
\end{aligned}$$

C Proofs of Section 3

Lemma C.1 (Multivariate change of variables)**.** *Let X be a random variable with density p_X and let $Y = g(X)$ where g is an invertible mapping with Jacobian J_g, then $p_Y(a) = p_X(g^{-1}(a))|J_g(a)|^{-1}$.*

Proof

$$P(Y \in A) = P(X \in g^{-1}(A)) = \int_{g^{-1}(A)} p_X(x)dx = \int_A p_X(g^{-1}(x))|J_{g^{-1}}(x)|dx = \int_A p_X(g^{-1}(x))|J_g(x)|^{-1}dx$$

The definition of density gives the result. $\square$

C.1 Proof of Theorem 1

Coverage Let $z_{-1} := (u_{-1}, x_{-1})$. Because

$$p_{u_1,x_1}(u, x) = \int p_{u_1,x_1|z_{-1}=z}(u, x) p_{z_{-1}}(z) dz,$$

it suffices to show that $p_{u_1,x_1|z_{-1}=z}$ has full coverage for any $z \in \mathbb{R}^{p+d}$. For this reason, in this proof, we fix z_{-1} and for notional simplicity omit explicitly conditioning on the event $z_{-1} = z$. Let $c := r(u_{-1})$ and $d := g(u_{-1}) + f(x_{-1})$. Moreover, we let $\xi_0' := \xi_0 + d$; note that (ξ_0', ξ_1) still has joint full coverage. Using this modified notation, we have

$$\begin{aligned}
x_0 &= \xi_0' \\
u_0 &= H_u(h(\xi_0'), c) \\
x_1 &= H_x(f(\xi_0'), g(H_u(h(\xi_0'), c))) + \xi_1 \\
u_1 &= H_u(h(x_1), r(H_u(h(\xi_0'), c)))
\end{aligned}$$

We now show that (x_1, u_1) has joint full coverage (implicitly conditioning on z_{-1}). We first show that x_1 has full coverage. Recall ξ_1 has positive density over $\mathbb{R}^d$. Note that the addition operation satisfies Assumption 1. Lemma C.1 implies that $H_x(f(\xi_0'), g(H_u(h(\xi_0'), c))) + \xi_1 | \xi_0'$ has positive density over $\mathbb{R}^d$. Since ξ_0' also has positive density over $\mathbb{R}^d$, integration tells us that $H_x(f(\xi_0'), g(H_u(h(\xi_0'), c))) + \xi_1$ has positive density over $\mathbb{R}^d$.

Because $p_{x_1,u_1} = p_{u_1|x_1} p_{x_1}$ and x_1 has full coverage, it suffices to show that $u_1|x_1$ has full coverage over $\mathbb{R}^p \times \mathbb{R}^d$. Because H_u satisfies Assumption 1, it is sufficient to show that $p_{q_c(\xi_0')|x_1}$ is positive everywhere. Indeed, recall that $u_1 = H_u(h(x_1), q_c(\xi_0'))$; furthermore, $u_1|x_1 = H_u^{h(x_1)}(q_c(\xi_0')|x_1$. If indeed $q_c(\xi_0')|x_1$ had positive density everywhere, then by Assumption 1 and Lemma C.1, $u_1|x_1$ would have positive density everywhere. Moreover, because q_c satisfies the conditions of Lemma C.1, it suffices to show $\xi_0'|x_1$ has positive density everywhere.

$$p_{\xi_0'|x_1}(a, b) = \frac{p_{x_1|\xi_0'}(b, a) p_{\xi_0'}(a)}{p_{x_1}(b)}$$

Since x_1 has full coverage, the denominator is positive. Since ξ_0' has full coverage, $p_{\xi_0'}(a) > 0$ as well. We had already shown earlier in the proof that $H_x(f(\xi_0'), g(H_u(h(\xi_0'), c))) + \xi_1 | \xi_0'$ has positive density everywhere.

Identifiability Because p_{u_1,x_1} is positive everywhere. $\mathbb{E}[x_2 \mid u_1 = u, x_1 = z]$ is well defined. Moreover because x_1 has density, $\int_z \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz$ is well defined as well. Finally because Proposition 1 says that our model is admissible, $\mathbb{E}[x_2 \mid do(u_1 \coloneqq u)] = \int_z \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz$. The right hand side of this relationship can be computed from knowledge of P_2, thus, $\mathbb{E}[x_2 \mid do(u_1 \coloneqq u)]$ can be computed from the distribution of observations P_2. With this quantity identifiable, the steerability of consumption is also identifiable.

C.2 Proof of Theorem 2

Proof For a random variable ξ_0, let x_1, x_0, u_0 be generated from (f, g, h), and let $\hat{x}_1, \hat{x}_0, \hat{u}_0$ be generated from $(\hat{f}, \hat{g}, \hat{h})$. Fix any measurable function $\Delta : \mathbb{R}^p \to \mathbb{R}^d$ such that $\Delta \neq 0$, define, $\hat{f}(a) \coloneqq f(a) + \Delta(h(a))$, $\hat{g}(b) \coloneqq g(b) - \Delta(b)$, and $\hat{h}(c) = h(c)$. We see that

$$\begin{aligned}
x_0 &\overset{d}{=} \xi_0 \overset{d}{=} \hat{x}_0 \\
u_0 &\overset{d}{=} h(\xi_0) \overset{d}{=} \hat{h}(\xi_0) = \hat{u}_0 \\
x_1 &\overset{d}{=} f(\xi_0) + g(h(\xi_0)) + \xi_1 \\
&\overset{d}{=} f(\xi_0) + \Delta(h(\xi_0)) + g(h(\xi_0)) - \Delta(h(\xi_0)) + \xi_1 \\
&\overset{d}{=} \hat{f}(\xi_0) + \hat{g}(\xi_0) + \xi_1 = \hat{x}_1.
\end{aligned}$$

□

C.3 Proof of Theorem 3

We first outline a series of helpful supporting lemmas to prove our desried result. We present them here

Lemma C.2. *Let $\{\xi_i\}_{i=1}^n$ be a set of mutually independent random vectors in $\mathbb{R}^d$ with full span. Let $\{A_i\}_{i=1}^n$ be a set of deterministic matrices in $\mathbb{R}^{d,d}$. Let $v \in \mathbb{R}^d$ be a random vector in $\mathbb{R}^d$ mutually independent of $\{\xi_i\}_{i=1}^n$. $A_i = 0$ for all $i \in [n]$ and $v \overset{a.s.}{=} 0$ if and only if $v + \sum_{i=1}^n A_i \xi_i \overset{a.s.}{=} 0$.*

Proof The left to right direction is obvious. We now prove the right to left direction. Suppose v is almost surely a constant vector. Suppose that only one $j \in [n]$ such that $A_j \neq 0$, then its not possible that $A_j \xi_j \overset{a.s.}{=} -v$ by definition of full span. Suppose there exists $j, k \in [n]$ such that $A_j \neq 0$ and $A_k \neq 0$. This means that $A_j \xi_j$ is almost surely not a constant. We also know that conditioned on $\{A_i \xi_i\}_{i \neq j}$, $A_j \xi_j$ is almost surely a constant. This implies that $P_{A_j \xi_j} \neq P_{A_j \xi_j | \{A_i \xi_i\}_{i \neq j}}$ which contradicts the assumption of mutual independence.

Suppose v is almost surely not a constant vector. Then $P_v \neq P_{v|\{A_i\xi_i\}_{i\in[n]}}$ as v is almost surely a constant vector conditioned on $\{A_i\xi_i\}_{i\in[n]}$. This contradicts mutual independence. $\square$

Lemma C.3. *Let $\Theta_x := [A, B]^T$, $\hat{\Theta}_x := [\hat{A}, \hat{B}]^T$, $\Theta_u := [C, D]^T$, and $\hat{\Theta}_u := [\hat{C}, \hat{D}]^T$. Let z_x, z_u and $\hat{z}_x, \hat{z}_u$ be the vectors of random variables corresponding to the state and control variables respectively of two data generating processes corresponding to the same DAG. Let ξ_x, ξ_u and $\hat{\xi}_x, \hat{\xi}_u$ correspond to independent, exogenous noise where $\xi_x \overset{d}{=} \hat{\xi}_x$ and $\xi_u \overset{d}{=} \hat{\xi}_u$. Let Q and $\hat{Q}$ be defined such that the edge relationships of the DAG can be represented as $z_v - \xi_v \overset{d}{=} Q_v\Theta_v$ and $\hat{z}_v - \hat{\xi}_v \overset{d}{=} \hat{Q}_v\hat{\Theta}_v$ for $v = x$ and $v = u$.*

Then for both $v = x$ and $v = u$, if $z_v - \xi_v \overset{d}{=} Q_v\hat{\Theta}_v$, then $z_v - \xi_v \overset{d}{=} Q_v\hat{\Theta}_v \overset{d}{=} \hat{Q}_v\hat{\Theta}_v \overset{d}{=} \hat{z}_v - \hat{\xi}_v$ for both $v = x$ and $v = u$.

Proof Combine the z_x and z_u matrices to form an ordered set z; do the same with $\hat{z}_x$ and $\hat{z}_u$ to form $\hat{z}$. These objects are sets to bypass dimensionality issues. Define $\sigma : \mathbb{Z} \to \mathbb{Z}$ such that $z_{\sigma(i)}$ is in sorted DAG order (i.e, the parents of $z_{\sigma(i)}$ have σ indices smaller than $\sigma(i)$ and its children have σ indices larger than $\sigma(i)$). We proceed inductively to show that $z_{\sigma(i)} \overset{d}{=} \hat{z}_{\sigma(i)}$.

Base case: $z_{\sigma(1)} \overset{d}{=} L(\xi, \hat{\Theta}_x, \hat{\Theta}_u)$, where L a tri-linear function in each of the parameters. Since $\xi \overset{d}{=} \hat{\xi}$, we have that $z_{\sigma(1)} \overset{d}{=} L(\hat{\xi}, \hat{\Theta}_x, \hat{\Theta}_u) = \hat{z}_{\sigma(1)}$; the last equality follows from definition.

Inductive step: suppose $z_{\sigma(j)} \overset{d}{=} \hat{z}_{\sigma(j)}$ jointly over all j. We know that $z_{\sigma(j+1)} \overset{d}{=} L(\{z_{\sigma(i)}\}_{i<j}, \xi, \hat{\Theta}_x, \hat{\Theta}_u)$ where L is linear in $\{z_{\sigma(i)}\}_{i<j}$, linear in ξ, linear with respect to $\hat{\Theta}_x$, and linear in $\hat{\Theta}_u$. By the inductive hypothesis we know that $z_{\sigma(j+1)} \overset{d}{=} L(\{z_{\sigma(i)}\}_{i<j}, \xi, \hat{\Theta}_x, \hat{\Theta}_u) \overset{d}{=} L(\{\hat{z}_{\sigma(i)}\}_{i<j}, \hat{\xi}, \hat{\Theta}_x, \hat{\Theta}_u) \overset{d}{=} \hat{z}_{\sigma(j+1)}$, as all the inputs to the function are equal in distribution. .

Because the entries of Q ($\hat{Q}$ respectively) are comprised of entries of z ($\hat{z}$ respectively), we have that $\hat{Q} \overset{d}{=} Q$. This proves the desired result. $\square$

Now that we have established our supporting lemmas, we can now prove our desired result.

Claim 1: Necessity and sufficiency when $x_{-1} = u_{-1} = \xi_t = 0$ for $t \geq 1$. Define the following variables:

$$\Theta_x := \begin{bmatrix} A^T \\ B^T \end{bmatrix} \qquad \Theta_u := \begin{bmatrix} C^T \\ D^T \end{bmatrix} \qquad \xi_x := \begin{bmatrix} \xi_0^T \\ \vdots \\ 0 \end{bmatrix} \qquad \xi_u := 0$$

Let $\hat{\Theta}_x, \hat{\Theta}_u$ be defined with respect to $\hat{A}, \hat{B}, \hat{C}, \hat{D}$. Let $\hat{\xi}_x \overset{d}{=} \xi_x$ and $\hat{\xi}_u \overset{d}{=} \xi_u$. Let (A, B, C, D) and ξ_x, ξ_u induce P_K and let $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ and $\hat{\xi}_x, \hat{\xi}_u$ induce $\hat{P}_K$.

Let $x := [x_0, \ldots, x_K]^T$ and $u := [u_0, \ldots, u_{K-1}]^T$ be observations from P_K and let $\hat{x}$ and $\hat{u}$ defined with hat variables be observations from $\hat{P}_K$. Let $X_t \in \mathbb{R}^{d,d}$ and $U_t \in \mathbb{R}^{p,d}$ be defined such that $x_t = X_t\varepsilon_0$ and $u_t = U_t\varepsilon_0$. Further define the following variables:

$$Q_x := \begin{bmatrix} x_{-1}^T & u_{-1}^T \\ x_0^T & u_0^T \\ \vdots & \vdots \\ x_{K-1}^T & u_{K-1}^T \end{bmatrix} \qquad Q_u := \begin{bmatrix} x_0^T & u_{-1}^T \\ x_1^T & u_0^T \\ \vdots & \vdots \\ x_{K-1} & u_{K-2} \end{bmatrix}$$

Define hat versions of all variables accordingly.

We have that $x - \xi_x \overset{d}{=} \hat{x} - \hat{\xi}_x$ and $u - \xi_u \overset{d}{=} \hat{u} - \hat{\xi}_u$. Moreover, since Q_u and Q_x are comprised of entries of x and u, $Q_u \overset{d}{=} \hat{Q}_u$ and $Q_x \overset{d}{=} \hat{Q}_x$ (jointly). Thus

$$x - \xi_x \overset{d}{=} \hat{x} - \xi_x \overset{d}{=} \hat{Q}\hat{\Theta}_x \overset{d}{=} Q\hat{\Theta}_x$$
$$u - \xi_u \overset{d}{=} \hat{u} - \xi_u \overset{d}{=} \hat{Q}\hat{\Theta}_u \overset{d}{=} Q\hat{\Theta}_u$$

Letting $X \coloneqq [X_0, \dots, X_K]^T$ and $U \coloneqq [U_0, \dots, U_{K-1}]^T$

$$Q_X \coloneqq \begin{bmatrix} X_{-1}^T & U_{-1}^T \\ X_0^T & U_0^T \\ \vdots & \vdots \\ X_{K-1}^T & U_{K-1}^T \end{bmatrix} \qquad Q_U \coloneqq \begin{bmatrix} X_0^T & U_{-1}^T \\ X_1^T & U_0^T \\ \vdots & \vdots \\ X_{K-1} & U_{T-2} \end{bmatrix},$$

we can rewrite the relationship as follows

$$\begin{aligned} \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot X &= \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot Q_X \hat{\Theta}_x \\ \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot U &= \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot Q_U \hat{\Theta}_u \end{aligned} \tag{4}$$

Using Lemma C.2 we know the above equality holds if and only if both the following holds

$$\begin{aligned} X &= Q_X \hat{\Theta}_x \\ U &= Q_U \hat{\Theta}_u \end{aligned} \tag{5}$$

The solution to these linear systems is unique if and only if Q_X and Q_U are full column rank.

Thus, if Q_X and Q_U are not full column rank, we can construct a $(\hat{\Theta}_x, \hat{\Theta}_u) \neq (\Theta_x, \Theta_u)$ and use Lemma C.3 to show that $\hat{P}_K = P_K$. If Q_X and Q_U are full column rank, by uniqueness, $(\hat{\Theta}_x, \hat{\Theta}_u) = (\Theta_x, \Theta_u)$. Thus, our system is identifiable if and only if Q_X and Q_U are full column rank. We conclude by showing equivalent conditions to Q_X and Q_U being full column rank.

Observe the following fact where ~ denotes rank equality

$$Q_X = \begin{bmatrix} I & C^T \\ X_1^T & (CX_1 + DU_0)^T \\ \vdots & \vdots \\ X_{K-1}^T & (CX_{K-1} + DU_{T-2})^T \end{bmatrix} \sim \begin{bmatrix} I & C^T \\ 0 & (DU_0)^T \\ \vdots & \vdots \\ 0 & (DU_{T-2})^T \end{bmatrix} \sim \begin{bmatrix} I & C^T \\ 0 & (DC)^T \\ \vdots & \vdots \\ 0 & (D^{K-1}C)^T \end{bmatrix}$$

The first ~ comes from elementary row operations. The second ~ comes from repeatedly applying the fact $U_t = CX_t + DU_{t-1}$ and using elementary operations. Thus, Q_X is full column rank if and only if $[DC, \dots, D^{K-1}C]$ has full row rank. We now apply the same logic for Q_U,

$$Q_U = \begin{bmatrix} I & 0 \\ X_1^T & U_0^T \\ \vdots & \vdots \\ X_{K-1}^T & U_{T-2}^T \end{bmatrix} \sim \begin{bmatrix} I & 0 \\ 0 & U_0^T \\ \vdots & \vdots \\ 0 & U_{T-2}^T \end{bmatrix} \sim \begin{bmatrix} I & 0 \\ 0 & C^T \\ 0 & (DC)^T \\ \vdots & \vdots \\ 0 & (D^{K-2}C)^T \end{bmatrix}$$

Thus, Q_U is full column rank if and only if $[C, \dots, D^{K-2}C]$ has full row rank. Tracing back all the if and only if statements gives the result.

Claim 2: Sufficiency even when $x_{-1} \neq 0$ and $u_{-1} \neq 0$. In this setting, the proof for Claim 1 holds up to Equation (4). Equation (4) changes to the following

$$\begin{aligned} \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot X + w_1(x_{-1}, u_{-1}, \xi_1, \xi_2) &= \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot Q_X \hat{\Theta}_x + w_2(x_{-1}, u_{-1}, \xi_1, \xi_2) \\ \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot U + w_3(x_{-1}, u_{-1}, \xi_1, \xi_2) &= \begin{bmatrix} \xi_0^T \\ \vdots \\ \xi_0^T \end{bmatrix} \odot Q_U \hat{\Theta}_u + w_4(x_{-1}, u_{-1}, \xi_1, \xi_2) \end{aligned}$$

By Lemma C.2, we know that these equalities hold if and only if Equation (5) holds, $w_1(x_{-1}, u_{-1}, \xi_1, \xi_2) = w_2(x_{-1}, u_{-1}, \xi_1, \xi_2)$ holds, and $w_3(x_{-1}, u_{-1}, \xi_1, \xi_2) = w_4(x_{-1}, u_{-1}, \xi_1, \xi_2)$ holds. Q_X and Q_U being full column rank *suffices* (but is no longer necessary as there are two other relationships we are not accounting for) in showing there is a unique solution to the linear system in Equation (5). Finally, as we show in Claim 1, Q_X is full column rank if and only if $[DC, \dots, D^{K-1}C]$ has full row rank, and Q_U is full column rank if and only if $[C, \dots, D^{K-2}C]$ has full row rank. This gives the result.

D Proofs of Appendix A

D.1 Proof of Theorem 4

The proof proceeds by bounding each of the following terms.

$$
\begin{aligned}
&\left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) Z^n(x) - \mathbb{E}[x_2 | do(u_1 \coloneqq u)] \right\| \leq \\
&\quad \left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) Z^n(x) - \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \right\| \\
&\quad + \left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) - \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \right\| \\
&\quad + \left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) - \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 = u, x_1 = x] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \right\| \\
&\quad + \left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 = u, x_1 = x] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) - \mathbb{E}[x_2 | do(u_1 \coloneqq u)] \right\|
\end{aligned}
$$

D.1.1 Supporting lemmas

We begin with a series of supporting lemmas that will aid us in bounding these terms.

Lemma D.1. *Let $P = P_2$. Let the conditions of Theorem 1 hold and let λ denote the Lebesgue measure for $\mathbb{R}^{d+p}$. Let $\mathcal{N}_{\mathcal{X}}$ be a cover of $\mathcal{X}$ and $\mathcal{N}_{\mathcal{U}}$ be a cover of $\mathcal{U}$. If for all $A \in \mathcal{N}_{\mathcal{X}}$ and $B \in \mathcal{N}_{\mathcal{U}}$, $\lambda(A \times B) > 0$, then $P(x_1 \in A, u_1 \in B) > 0$ for all $A \in \mathcal{N}_{\mathcal{X}}, B \in \mathcal{N}_{\mathcal{U}}$.*

Proof

$$
P(x_1 \in A, u_1 \in B) = \int_B \int_A p_{x_1, u_1}(x, u) dx du > 0
$$

We know the RHS is positive because the function being integrated is positive by Theorem 1 and the set it's being integrated over has measure greater than 0. □

Lemma D.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^p$, and let $f : \mathcal{X} \to \mathcal{Y}$ be a L-Lipschitz function. Let $\mathcal{N}$ be any ε-cover of $\mathcal{X}$ for $\varepsilon > 0$. Then for all $U \in \mathcal{N}$, for all $x, y \in U$, $\|f(x) - f(y)\| \leq L\varepsilon$.*

Proof Follows directly from definitions of cover and Lipschitz Continuity. □

Lemma D.3. *Consider the data generation model of* (1)*. Let Assumption 4 hold. Let x_1 have full coverage. Let $\mathcal{R}_{\mathcal{X}}$ denote the ε-cover representatives of $\mathcal{X} \subset \mathbb{R}^d$. Then,*

$$
\left\| \mathbb{E}[x_2 | do(u_1 \coloneqq u)] - \sum_{r \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 \mid u_1 = u, x_1 = r] P(x_1 \in \mathcal{N}_{\mathcal{X}}(r)) \right\| \leq L\varepsilon
$$

Proof

Let $B \coloneqq \mathbb{R}^d \setminus \cup_{A \in \mathcal{N}_{\mathcal{X}}} A$ denote the set of points not covered by the epsilon cover.

$$
\begin{aligned}
&\left\| \mathbb{E}[x_2 | do(u_1 \coloneqq u)] - \sum_{r \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 \mid u_1 = u, x_1 = r] P(x_1 \in \mathcal{N}_{\mathcal{X}}(r)) \right\| \\
&\quad \leq \left\| \sum_{r \in \mathcal{R}_{\mathcal{X}}} \int_{\mathcal{N}_{\mathcal{X}}(r)} \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz - \sum_{r \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 \mid u_1 = u, x_1 = r] P(x_1 \in \mathcal{N}_{\mathcal{X}}(r)) \right\| \\
&\qquad + \left\| \int_B \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz \right\| \\
&\quad \leq \sum_{r \in \mathcal{R}_{\mathcal{X}}} \left\| \int_{\mathcal{N}_{\mathcal{X}}(r)} \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz - \mathbb{E}[x_2 \mid u_1 = u, x_1 = r] P(x_1 \in \mathcal{N}_{\mathcal{X}}(r)) \right\| \\
&\qquad + \left\| \int_B \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz \right\| \\
&\quad \leq \sum_{r \in \mathcal{R}_{\mathcal{X}}} \int_{\mathcal{N}_{\mathcal{X}}(r)} \| \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] - \mathbb{E}[x_2 \mid u_1 = u, x_1 = r] \| \, p_{x_1}(z) dz \\
&\qquad + \left\| \int_B \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz \right\| \\
&\quad \leq L\varepsilon + \left\| \int_B \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz \right\|
\end{aligned}
$$

The first and second inequality is from triangle inequality. The third comes from Jensen's inequality. The fourth inequality comes Assumption 4 and Lemma D.2.

□

Lemma D.3 tells us that it suffices to create an estimator that estimates $\sum_{r \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 \mid u_1 = u, x_1 = r] P(x_1 \in \mathcal{N}_{\mathcal{X}}(r))$.

Lemma D.4. *Consider the data generating process from* (1)*. Let Assumption 5 hold, then*

$$
\| \mathbb{E}[Y^n(u, x)] - \mathbb{E}[x_2 | u_1 = u, x_1 = x] \| \leq \frac{2\eta}{1 - \eta} \mathbb{E}[\| x_2 \| \, | u_1 = u, x_1 = x].
$$

Proof Fix any $u \in \mathcal{U}, x \in \mathcal{X}$. Let $Z \coloneqq (u_1, x_1)$, $z \coloneqq (u, x)$, and $A \coloneqq \mathcal{N}_{\mathcal{U}}(u) \times \mathcal{N}_{\mathcal{X}}(x)$. Observe that $\mathbb{E}[Y^n(u, x)] = \mathbb{E}[x_2 \mid Z \in A]$.

$$
\begin{aligned}
\| \mathbb{E}[x_2 \mid Z \in A] - \mathbb{E}[x_2 \mid Z = z] \| &= \left\| \int_{\mathbb{R}^d} x \left[\frac{P(Z \in A | x_2 = x)}{P(Z \in A)} - \frac{p(Z = z | x_2 = x)}{p(Z = z)} \right] p(x_2 = x) dx \right\| \\
&\leq \int_{\mathbb{R}^d} \| x \| \left| \frac{P(Z \in A | x_2 = x)}{P(Z \in A)} - \frac{p(Z = z | x_2 = x)}{p(Z = z)} \right| p(x_2 = x) dx \\
&= \int_{\mathbb{R}^d} \| x \| \left| \frac{\int_A p(Z = a | x_2 = x) da}{\int_A p(Z = a) da} - \frac{p(Z = z | x_2 = x)}{p(Z = z)} \right| p(x_2 = x) dx \\
&\leq \frac{2\eta}{1 - \eta} \int_{\mathbb{R}^d} \| x \| \frac{p(Z = z | x_2 = x)}{p(Z = z)} p(x_2 = x) dx \\
&= \frac{2\eta}{1 - \eta} \mathbb{E}[\| x_2 \| \, | Z = z].
\end{aligned}
$$

The first inequality is an application of Jensen's inequality. The second inequality is an application of Assumption 5 and the fact that the diameter of A is no more than ε. □

D.1.2 Applying lemmas to bound terms

Armed with these lemmas we can proceed with bounding each of the aforementioned terms.

First term: Recall that the following holds for a τ^2-subgaussian random variable X.

$$
P(|X - \mathbb{E}[X]| > \delta |\mathbb{E}[X]|) \leq 2 \exp\left(\frac{-\delta^2 \mathbb{E}[X]^2}{2\tau^2} \right)
$$

$Z^n(x)$ is $\frac{1}{4n}$ subgaussian. This means we need $n = \frac{1}{2\delta^2 P(x_1 \in \mathcal{N}_{\mathcal{X}}(x))^2} \log(4|\mathcal{R}_{\mathcal{X}}|/\rho)$ samples to get $Z^n(x)$ within error of $\delta P(x_1 \in \mathcal{N}_{\mathcal{X}}(x))$ of $P(x_1 \in \mathcal{N}_{\mathcal{X}}(x))$ with probability $\rho/(2|\mathcal{R}_{\mathcal{X}}|)$. Using union bound, we have that with probability with at least $1 - \rho/2$,

$$
\begin{aligned}
\left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) Z^n(x) - \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \right\| &\leq \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|Y^n(u,x)\| \, |Z^n(x) - P(x_1 \in \mathcal{N}_{\mathcal{X}}(x))| \\
&\leq \delta \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|Y^n(u,x)\| \, P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\leq \delta \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|Y^n(u,x) - \mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)]\| \, P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\quad + \delta \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|\mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)]\| \, P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\leq \delta\gamma + \delta \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|\mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)]\| \, P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\leq \delta\gamma + \delta c_0 D + \delta c_0 \mathbb{E}[\|g(u_1)\| \mid u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)] \\
&\leq \delta\gamma + 2\delta c_0 D
\end{aligned}
$$

where the first inequality comes from triangle inequality. The second inequality comes from subgaussianity. The third inequality is from triangle inequality. The fourth inequality is from the bound of the **Second term** below. The fifth and sixth inequalities are from Assumption 6 and compactness and from the fact $\mathbb{E}[\xi_t] = 0$.

Second term: $Y^n(u,x)$ is $\frac{\sigma^2}{n_{u,x}}$ subgaussian, which means its $\frac{d\sigma^2}{n_{u,x}}$ norm-subgaussian by Lemma 1 from Jin et al. [2019]. Thus, the following inequality holds

$$
P(\|Y^n(u,x) - \mathbb{E}[Y^n(u,x)]\| \geq t) \leq 2 \exp\left(-\frac{t^2 n_{u,x}}{2d\sigma^2} \right)
$$

This means we need $n_{u,x} = \frac{2d\sigma^2}{\gamma^2} \log(4|\mathcal{R}_{\mathcal{X}}|/\rho)$ samples to get $Y^n(u,x)$ with error γ of $\mathbb{E}[Y^n(u,x)]$ with probability $\rho/(2|\mathcal{R}_{\mathcal{X}}|)$. Moreover, because the conditions of Lemma D.1 are met, we know these requirements will hold for all $n_{u,x}$ for large enough n. Using union bound, we have that with probability with at least $1 - \rho/2$,

$$
\begin{aligned}
&\left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} Y^n(u,x) P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) - \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \right\| \\
&\quad \leq \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|Y^n(u,x) - \mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)]\| \, P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\quad \leq \gamma
\end{aligned}
$$

The first inequality comes from Jensen's inequality. The second comes from subgaussianity.

Third term:

$$
\begin{aligned}
&\left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) - \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 = u, x_1 = x] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \right\| \\
&\quad \leq \sum_{x \in \mathcal{R}_{\mathcal{X}}} \|\mathbb{E}[x_2 | u_1 \in \mathcal{N}_{\mathcal{U}}(u), x_1 \in \mathcal{N}_{\mathcal{X}}(x)] - \mathbb{E}[x_2 | u_1 = u, x_1 = x])\| \, P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\quad \leq \frac{2\eta}{1-\eta} \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[\|x_2\| \, | u_1 = u, x_1 = x] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) \\
&\quad \leq \frac{2\eta}{1-\eta} \left(c_0 D + c_0 \|g(u)\| + c_1 \right)
\end{aligned}
$$

The first inequality comes from Jensen's inequality. The second comes from Lemma D.4. The third inequality comes from Assumption 6.

Fourth term:

$$\left\| \sum_{x \in \mathcal{R}_{\mathcal{X}}} \mathbb{E}[x_2 | u_1 = u, x_1 = x] P(x_1 \in \mathcal{N}_{\mathcal{X}}(x)) - \mathbb{E}[x_2 | do(u_1 := u)] \right\| \leq L\varepsilon + \left\| \int_B \mathbb{E}[x_2 \mid u_1 = u, x_1 = z] p_{x_1}(z) dz \right\|$$
$$\leq L\varepsilon + c_0 \mathbb{E}[\|f(x_1)\| \mathbf{1}\{x_1 \in B\}] + c_0 P_{x_1}(B) \|g(u)\|.$$

The first inequality comes from Lemma D.3. The second inequality comes from $\mathbb{E}\xi_2 = 0$ and Assumption 6 and Jensen's inequality.

Union bounding over the two events in bounding the first and second terms and combining all the inequalities gives the result.

D.2 Proof of Theorem 5

We first introduce a helpful supporting lemma.

Lemma D.5. *Suppose n samples are drawn iid from P_2. If $X_0 X_0^T$ is invertable, then $\widehat{C} = C$ and $\widehat{H} = A + BC + E_1 X_0^T (X_0 X_0^T)^{-1}$. If $X_0 X_0^T$ is invertable and $DCX_0 X_0^T C^T D^T$ is invertable, then $\widehat{B} = B - E_1 X_0^T (X_0 X_0^T)^{-1} X_1 (DCX_0)^T (DCX_0 X_0^T C^T D^T)^{-1}$.*

Proof Substituting $U_0 = CX_0$ and $X_1 = (A + BC)X_0 + E_1$ into the closed form solutions of $\widehat{C}$ and $\widehat{H}$ respectively gives the first result.

To get the second result, we use the fact that $\widehat{C} = C$ and $\widehat{H} = A + BC + E_1 X_0^T (X_0 X_0^T)^{-1}$ by the first result. We observe that $U_1 = CX_1 + DU_0 = CX_1 + DCX_0$ to get that $\widehat{B} = (X_2 - \widehat{H} X_1)(DCX_0)^T (DCX_0 X_0^T C^T D^T)^{-1}$. Then we use the fact that subtracting BCX_1 from both sides of the relationship $X_2 - AX_1 = BU_1$ gives us that $X_2 - (A + BC)X_1 = B(U_1 - CX_1)$. Using our invertability assumptions, this gives us $\widehat{B} = B - E_1 X_0^T (X_0 X_0^T)^{-1} X_1 (DCX_0)^T (DCX_0 X_0^T C^T D^T)^{-1}$. □

With this, we can analyze the quantities of interest. Let $\hat{\Sigma}_0 = \frac{1}{n} X_0 X_0^T$. Let $Q := DC\hat{\Sigma}_0 C^T D^T$.

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{B} - B\right\|_{\mathrm{Fr}}^2 \mid \mathcal{G}\right] &= \frac{1}{n^2} \operatorname{tr}(\mathbb{E}[Q^{-1} DCX_0 X_1^T (X_0 X_0^T)^{-1} X_0 E_1^T E_1 X_0^T (X_0 X_0^T)^{-1} X_1 X_0^T C^T D^T Q^{-1}]) \\
&= \frac{\sigma_1^2 d}{n} \operatorname{tr}(\mathbb{E}[Q^{-1} DC\hat{\Sigma}_0 (A + BC)^T \hat{\Sigma}_0^{-1} (A + BC) \hat{\Sigma}_0 C^T D^T Q^{-1}]) \\
&\leq \frac{\sigma_1^2 pd}{n} \kappa_{DC}^2 \left(\frac{\|A + BC\|_{\mathrm{op}}}{\sigma_{\min}(DC)} \right)^2 \mathbb{E}\left[\frac{\kappa_{\hat{\Sigma}_0}^2}{\lambda_{\min}(\hat{\Sigma}_0)} \right].
\end{aligned}$$

Rearranging and using the definition of τ_0 gives the result.

If $p = d$, then DC is a square, invertible matrix,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{B} - B\right\|_{\mathrm{Fr}}^2 \mid \mathcal{G}\right] &= \frac{\sigma_1^2 d}{n} \operatorname{tr}[(C^T D^T)^{-1} (A + BC)^T \mathbb{E}\left[\hat{\Sigma}_0^{-1}\right] (A + BC)(DC)^{-1}] \\
&\leq \frac{\sigma_1^2 d^2}{n} \left(\frac{\|A + BC\|_{\mathrm{op}}}{\lambda_{\min}(DC)} \right)^2 \left\|\mathbb{E}\left[\hat{\Sigma}_0^{-1}\right]\right\|_{\mathrm{op}}.
\end{aligned}$$

Rearranging and using the definition of τ_1 gives the result.