# CSDT

October 30, 2019

## 1 Invariance and Capital-C Causality

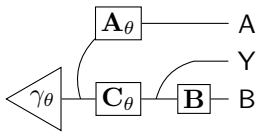
CSDT features *consequences* - that is, probabilistic relations between decisions and results - but it does not feature *causal effects*, which seem to be probabilistic relations between random variables on the observation space $E$ that are not necessarily disintegrations of a joint distribution. Here I propose the following notion of a causal effect in CSDT: if I have prior knowledge about the consequences of my decisions $D$ on some random variable $\mathsf{B}$ via $\mathbf{C}_0\mathbf{B}$ and I can extend this to the consequences on $\mathsf{Y}$ using some Markov kernel $\mathbf{G}_\theta : B \to \Delta(\mathcal{Y})$ via composition $\mathbf{C}_\theta\mathbf{Y} = \mathbf{C}_0\mathbf{B}\mathbf{G}_\theta$ then we'll say that $\mathbf{G}_\theta$ is the "causal effect" of $\mathsf{B}$ on $\mathsf{Y}$ in state $\theta$.

This notion accords with intuitions about do-interventions - if I have the option to $do(\mathsf{B})$ then I definitely know the effect of my decision on $\mathsf{B}$. Furthermore, using do interventions I assume that the effect of $do(\mathsf{B} = x)$ on $\mathsf{Y}$ is given by fixing $\mathsf{B}$ to $x$ and then computing $P(\mathsf{Y}|do(\mathsf{B}) = x)$, a special case of the composition above if we let $\mathbf{G}_\theta = P(\mathsf{Y}|do(\mathsf{B}))$.

However, the notion of causal effect proposed here doesn't make additional commitments that $do(\mathsf{B})$ does - namely, that I also know $do(\mathsf{B})$ has no direct effect on any other variable. Unlike the assumption of prior knowledge, the "no direct effects" assumption is cannot even be formulated within CSDT.

The notion of "causal effect" given here permits a (to my knowledge) novel set of assumptions under which this type of causal effect of $\mathsf{B}$ on $\mathsf{Y}$ can be inferred. Importantly, it avoids any assumptions of "conditional independence of unobservables", and is to my knowledge the only known case of "causal identifiability" that achieves this.

Suppose $\mathbf{C} : D \times \Theta \to \Delta(\mathcal{E})$ is the consequence of interest, and furthermore given some $\theta \in \Theta$ we have $\mathbf{A}_\theta : D \to \Delta(\mathcal{A})$ and $\mathbf{B} : E \to \Delta(\mathcal{B})$ such that the observations are distributed according to

$\mathbf{H}_\theta :=$ 

(1)

That is, the observations $\mathbf{H}_\theta$ are the inputs and outputs of $\mathbf{C}_\theta$ where the inputs are masked by $\mathbf{A}_\theta$ and we run one copy of the outputs through a fixed $\mathbf{B}$. Without the mask, we could recover $\mathbf{C}_\theta$ from $\mathbf{H}_\theta$ via disintegration $\mathsf{D} \dashrightarrow \mathsf{Y}$.

For all $\theta$, assume $\mathbf{A}_\theta$ is $\gamma_\theta$-almost surely right invertible and that $\gamma_\theta$ is strictly positive. Together these are strong assumptions. Note that as $\theta$ is unknown, despite the fact that $\mathbf{A}_\theta$ is right invertible, we can't recover it's inputs, so we can't simply disintegrate.

This setup is something like an instrumental variables (IV) setup where $\gamma_\theta$ is a source of "exogenous" variation. It is stronger than a typical IV setup in that $\gamma_\theta$ is strictly positive and $\mathbf{A}_\theta$ is right invertible, but it is weaker than a typical IV setup in that we make no assumptions at all about model class and fewer assumptions about the relationships between $\mathsf{A}$, $\mathsf{B}$ and $\mathsf{Y}$.

It is also somewhat similar to Arjovsky et al. (2019), itself based on Peters et al. (2016), if $\mathsf{A}$ is understood as an environment indicator. Under this interpretation, the assumption that $\gamma_\theta$ is strictly positive is much stronger than Arjovsky et al. (2019), but on the other hand we do not assume any particular relationship between $\mathsf{B}$ and $\mathsf{Y}$ where Arjovsky et al. (2019) assumes a particular form of SEM. An interpretational difference is that while we regard $D$ as a set of feasible decisions, Arjovsky et al. (2019) considers the set of environments thus:

> Here, the set of all environments contains all possible experimental conditions concerning oursystem of variables, both observable and hypothetical.

As an aside, we could potentially interpret such a set of environments as an rich set of decisions which may be coarsened to a realistic set of decisions.

Under certain conditions, this setup allows for the extension of causal knowledge via observed data. In particular, if we find the conditional probability of $\mathsf{Y}$ on $\mathsf{A}$ and $\mathsf{B}$ (written $[\mathsf{Y}|\mathsf{A}\underline{\otimes}\mathsf{B}]_\theta$) is independent of $\mathsf{A}$, then given prior knowledge for the effect of a decision on $\mathsf{B}$ we can deduce the full consequence map $\mathbf{C}_\theta$ from our prior knowledge and the disintegration $[\mathsf{Y}|\mathsf{B}]_\theta$ (see Theorem 1.2). Note that no assumptions have been made about "causal" relationships between $\mathsf{B}$ and $\mathsf{Y}$ - $[\mathsf{Y}|\mathsf{B}]_\theta$ is an ordinary disintegration, not a platonic Markov kernel/structural equation/FFRCISTG/whatever.

A kernel $\mathsf{B}$ that throws away more information is advantageous in the sense that less prior knowledge is needed to determine $\mathbf{C}_\theta$.

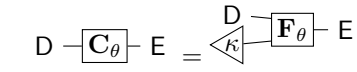
**Example 1.1** (Waste collection)**.** A council is deciding on how to implement a waste collection service to reduce littering $\mathsf{Y}$. For every possible service $d \in D$, the collection schedule $\mathsf{S}$ that will be achieved is known prior to any investigation (services may differ in other ways - e.g. the bin types may differ, and these differences may or may not be known in advance). In addition, the council has obtained weekly collection and littering data from a set $A$ of other councils with their own waste collection services that are known to have faced the same unknown consequence map $\mathbf{C}_\theta$. Each other council has implemented exactly one possible service $d \in D$ and enough councils were surveyed that all possible choices of service have been sampled, though they do not know which councils

have implemented which systems. These conditions ensure that for each service $d$ the set of councils $A_d \subset A$ implementing that service is disjoint from the set of councils implementing any other service, and hence the unknown map from services to councils $\mathbf{A}_\theta : D \to \Delta(\mathcal{A})$ is right invertible.

It is found by the council's statisticians that the disintegration $[\mathsf{Y}|\mathsf{A}\underline{\otimes}\mathsf{S}]_\theta$ is independent of $\mathsf{A}$. Thus by theorem 1.2 the impact of any service $d$ on the rate of littering $\mathsf{Y}$ can be found via the collection schedule $\mathsf{S}$ that will be achieved by that service and the disintegration $[\mathsf{Y}|\mathsf{S}]_\theta$.

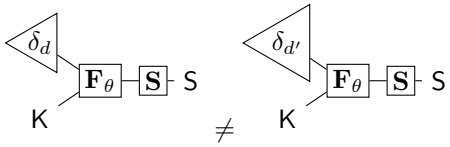
This is a surprisingly strong conclusion from the assumptions made. We appear to have the ability to deduce a "causal effect" from observational data in a context that looks remarkably similar to standard examples of when this *can't* be done. In fact, the conclusion is stronger than a *do*-style causal relationship, as *do* interventions assume we know a decision has no "direct effect" on any variable other than the target, whereas here we only assume the consequence of $d$ on $\mathsf{S}$ is known.

An objection might be that other councils' waste collection service choices are determined, in part, by some unobserved background factors. Suppose that these background factors take values in some space $K$. One means of formalising this is that $\mathbf{C}_\theta$ factorises:




(2)

Where $\kappa \in \Delta(\mathcal{K})$ is a distribution on background factors and $\mathbf{F}_\theta : D \times K \to \Delta(\mathcal{E})$ maps decisions and background factors to consequences. For $d \neq d'$ we may have:



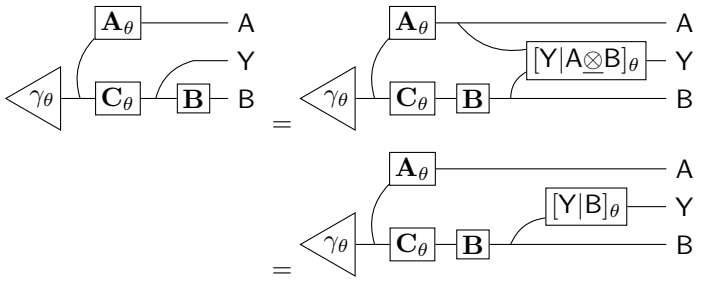

(3)

That is, different decisions may induce different relationships between background characteristics $\mathsf{K}$ and collection schedules $\mathsf{S}$. Formally, introducing this extra complication has not violated any of our assumptions - the causal inference is still valid! Practically, we would typically need a much larger set of decisions $D$ in this case to account for the number of plausible relationships between $\mathsf{K}$ and $\mathsf{S}$, and this may give us more reason to question the assumption that $\gamma_\theta$ is strictly positive - i.e. that the set of councils implementing each decision has positive measure. Nonetheless, this appears to be quite different to existing conditions for observational causal inference: we have allowed for $\mathsf{K}$ to affect both $\mathsf{S}$ and $\mathsf{Y}$ but in contrast to existing approaches **we do not need to see** $\mathsf{K}$ **in order to – sometimes – infer the "causal effect" of** $\mathsf{S}$ **on** $\mathsf{Y}$.

**Theorem 1.2.** *Suppose we have a causal theory* $\mathbf{T} : \Theta \times D \to \Delta(\mathcal{E}^2)$ *where for* $\theta \in \Theta$ *we have consequence* $\mathbf{C}_\theta$ *and experiment* $\mathbf{H}_\theta$ *given by 1. Suppose for some* $\mathbf{C}_0 : D \to \Delta(\mathcal{E})$ *we have* $\mathbf{C}_\theta\mathbf{B} = \mathbf{C}_0\mathbf{B}$ *for all* $\theta \in \Theta$.

*For all $\theta \in \Theta$ such that $[\mathsf{Y}|\mathsf{A}\underline{\otimes}\mathsf{B}]_\theta$ is independent of $\mathsf{A}$ we have $\mathbf{C}_\theta = \mathbf{C}_0\mathbf{B}[\mathbf{Y}|\mathbf{B}]_\theta$*

*Proof.* By the definition of disintegration

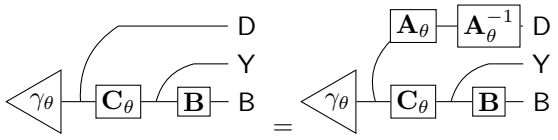



(4)

(5)

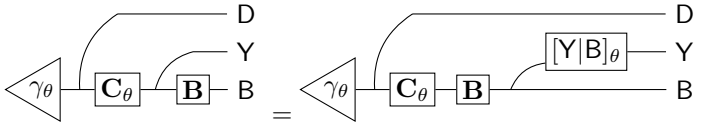
Where 5 follows from the independence of $[\mathsf{Y}|\mathsf{A}\underline{\otimes}\mathsf{B}]_\theta$ from $\mathsf{A}$.

In addition, we have



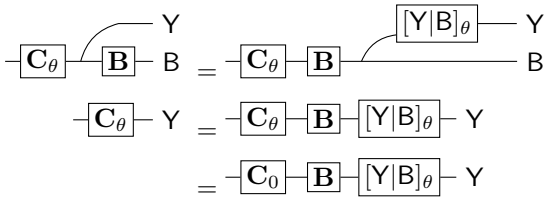

(6)

Thus




(7)

From Lemma 1.3 and by the assumption of strict positivity on $\gamma_\theta$, we therefore have




(8)

(9)

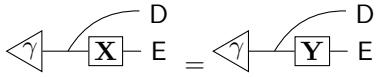
(10)

Where 9 follows from marginalisation of 8.

□

A non-invertible $\mathbf{A}_\theta$ means 6 doesn't hold. Given that $\mathbf{A}_\theta$ is arbitrary apart from the fact that it is invertible, I wonder if a channel capacity lower boundon $\mathbf{A}_\theta$ could give an upper bound on the "distance" between the kernels on the left and right of 6 using an appropriate replacement for $\mathbf{A}_\theta^{-1}$.

**Lemma 1.3.** *Given strictly positive probability measure $\gamma \in \Delta(\mathcal{D})$ and Markov kernels $\mathbf{X} : D \to \Delta(\mathcal{E})$ and $\mathbf{Y} : D \to \Delta(\mathcal{E})$ if*




(11)

*Then $\mathbf{X} = \mathbf{Y}$.*

*Proof.* We note that both $\mathsf{X}$ and $\mathsf{Y}$ are $\mathsf{D} \dashrightarrow \mathsf{E}$ disintegrations of 11, and so they must be $\gamma$-almost surely equal. Strict positivity means they must in fact be equal. □

## References


Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, July 2019. URL `http://arxiv.org/abs/1907.02893`. arXiv: 1907.02893.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12167. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167`.

## Appendix: