Proceedings of Machine Learning Research vol 140:1–5, 2022 1st Conference on Causal Learning and Reasoning

Supplementary Material to “Causal Explanations and XAI”

Sander Beckers SREKCEBREDNAS@GMAIL.COM
University of Tübingen

Editors: Bernhard Schölkopf, Caroline Uhler and Kun Zhang

Appendix A.

Proposition 8 *If $\boldsymbol{X} = \boldsymbol{x}$ is directly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$ then $\boldsymbol{X} = \boldsymbol{x}$ is strongly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$ along some $\boldsymbol{N}$, and if $\boldsymbol{X} = \boldsymbol{x}$ is strongly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$ along some $\boldsymbol{N}$ then $\boldsymbol{X} = \boldsymbol{x}$ is weakly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$.*

Proof: Follows directly from the definitions. ▮

Theorem 12 *If a causal model M that agrees with h satisfies* **Independence** *then the following statements are all equivalent:*

- *$\boldsymbol{X} = \boldsymbol{x}$ is weakly sufficient for $Y = y$ in M.*
- *$\boldsymbol{X} = \boldsymbol{x}$ is strongly sufficient for $Y = y$ in M.*
- *$\boldsymbol{X} = \boldsymbol{x}$ is directly sufficient for $Y = y$ in M.*

Proof: The implications from bottom to top are a direct consequence of Proposition 8.

Assume $\boldsymbol{X} = \boldsymbol{x}$ is weakly sufficient for $Y = y$ in M. This means that for all $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]Y = y$. Let $\boldsymbol{C} = \mathcal{R}(\mathcal{V} - (\boldsymbol{X} \cup \{Y\}))$.

Given that M agrees with h, either the equation for Y is of the form $Y = U$ for some $U \in \mathcal{U}$, or Y only has parents in $\mathcal{V} \setminus \{Y\}$. Since the former contradicts our assumption that in all contexts $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]Y = y$, it has to be the latter.

As a consequence, interventions on all endogenous variables make the particular context irrelevant, i.e., for all $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$ and all $\boldsymbol{u_1}, \boldsymbol{u_2} \in \mathcal{R}(\mathcal{U})$, we have that $(M, \boldsymbol{u_1}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{C} \leftarrow \boldsymbol{c}]\boldsymbol{Y} = \boldsymbol{y}$ iff $(M, \boldsymbol{u_2}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{C} \leftarrow \boldsymbol{c}]\boldsymbol{Y} = \boldsymbol{y}$.

Further, for each $V_i \in \mathcal{V} \setminus \{Y\}$ the equation is of the form $V_i = U_i$. Although technically one could choose to define $\mathcal{R}(V_i)$ such that $\mathcal{R}(V_i) \not\subseteq \mathcal{R}(U_i)$, this comes down to defining a variable with values that it cannot obtain, which serves no purpose. Therefore we can assume that for each $\boldsymbol{c} \in \boldsymbol{C}$ there exists a context $\boldsymbol{u}'$ such that $(M, \boldsymbol{u}') \models \boldsymbol{C} = \boldsymbol{c}$. Given that no members of $\boldsymbol{X}$ are parents of members of $\boldsymbol{C}$, it also holds that $(M, \boldsymbol{u}') \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]\boldsymbol{C} = \boldsymbol{c}$. Since $\boldsymbol{X} = \boldsymbol{x}$ is weakly sufficient for $Y = y$, we also have that $(M, \boldsymbol{u}') \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]\boldsymbol{C} = \boldsymbol{c} \wedge Y = y$, from which it follows that $(M, \boldsymbol{u}') \models [\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{C} \leftarrow \boldsymbol{c}]Y = y$. Taken together with the previous observation that the particular context is irrelevant, we get that for all $\boldsymbol{c} \in \mathcal{R}(\mathcal{V} - (\boldsymbol{X} \cup \boldsymbol{Y}))$ and all $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{C} \leftarrow \boldsymbol{c}]\boldsymbol{Y} = \boldsymbol{y}$, which is what we had to prove. ▮

Theorem 16 *Given a causal setting $(M, \boldsymbol{u})$, the following two statements are equivalent:*

- *there exist $\boldsymbol{W_1}$ and $\boldsymbol{w_1} \in \mathcal{R}(\boldsymbol{W_1})$ so that $Y = y$ counterfactually depends on $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ relative to $\boldsymbol{W_1} = \boldsymbol{w_1}$ in $(M, \boldsymbol{u})$.*
- *there exist $\boldsymbol{W_2}$, $\boldsymbol{w_2} \in \mathcal{R}(\boldsymbol{W_2})$, and $\boldsymbol{N}$ so that $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_2} = \boldsymbol{w_2}, \boldsymbol{N})$ is a good counterfactual explanation of $Y = y$.*

Proof:

Observation 1 *Recall from Definition 3 that exogenous variables only appear in equations of the form $V = U$. Say $\boldsymbol{R} \subseteq \mathcal{V}$ are all variables which have such an equation, and call these the* root *variables. It is clear that if we intervene on all of the root variables, they take over the role of the exogenous variables. Concretely, given strong recursivity, for any setting $\boldsymbol{r} \in \mathcal{R}(\boldsymbol{R})$ there exists a unique setting $\boldsymbol{v} \in \mathcal{R}(\mathcal{V})$ so that for all contexts $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ we have that $(M, \boldsymbol{u}) \models [\boldsymbol{R} \leftarrow \boldsymbol{r}]\mathcal{V} = \boldsymbol{v}$.*

Assume that $Y = y$ counterfactually depends on $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ relative to $\boldsymbol{W_1} = \boldsymbol{w_1}$ in $(M, \boldsymbol{u})$. This means that $(M, \boldsymbol{u}) \models \boldsymbol{X} = \boldsymbol{x} \wedge \boldsymbol{W_1} = \boldsymbol{w_1} \wedge Y = y$, and $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W_1} \leftarrow \boldsymbol{w_1}]Y \neq y$.

Let $\boldsymbol{S} = \boldsymbol{R} \setminus (\boldsymbol{W_1} \cup \boldsymbol{X})$, and let $\boldsymbol{s} \in \mathcal{R}(S)$ be the unique values so that $(M, \boldsymbol{u}) \models \boldsymbol{S} = \boldsymbol{s}$.

As $\boldsymbol{R} \subseteq (\boldsymbol{S} \cup \boldsymbol{W_1} \cup \boldsymbol{X})$, we have that $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{W_1} = \boldsymbol{w_1})$ is strongly sufficient for $Y = y$ along $\boldsymbol{N} = \mathcal{V} \setminus (\boldsymbol{X} \cup \boldsymbol{W_1} \cup \boldsymbol{S})$, and thus $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{W_1} = \boldsymbol{w_1}), \boldsymbol{N})$ is an actual sufficient explanation of $Y = y$.

Furthermore, changing $\boldsymbol{X}$ from $\boldsymbol{x}$ to $\boldsymbol{x}'$ obviously has no effect on any of the other values in $\boldsymbol{R}$. Therefore $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W_1} \leftarrow \boldsymbol{w_1}]\boldsymbol{S} = \boldsymbol{s}$, and thus we get that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W_1} \leftarrow \boldsymbol{w_1}, \boldsymbol{S} \leftarrow \boldsymbol{s}]Y = y'$ for some $y' \neq y$. As before, we can conclude that $(\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{W_1} = \boldsymbol{w_1})$ is strongly sufficient for $Y = y'$ along $\boldsymbol{N}$, and thus $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{W_1} = \boldsymbol{w_1}), \boldsymbol{N})$ is a sufficient explanation of $Y = y'$.

Combining the two previous paragraphs, we get that $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{W_1} = \boldsymbol{w_1})), \boldsymbol{N})$ is a counterfactual explanation of $Y = y$. Let $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_2} = \boldsymbol{w_2}), \boldsymbol{N_2})$ be a dominating counterfactual explanation that is not strictly dominated by any other explanation that contains $\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}')$. (It is easy to see that such an explanation must exist: one can simply keep removing elements from $\boldsymbol{W_1}$, $\boldsymbol{S}$, and $\boldsymbol{N}$ until no further element can be removed while still remaining a counterfactual explanation of $Y = y$.)

Now assume that there exist $\boldsymbol{W_2}$, $\boldsymbol{w_2} \in \mathcal{R}(\boldsymbol{W_2})$, and $\boldsymbol{N}$ so that $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W_2} = \boldsymbol{w_2}), \boldsymbol{N})$ is an actual sufficient explanation of $Y = y$ and $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{W_2} = \boldsymbol{w_2}), \boldsymbol{N})$ is a sufficient explanation of some $Y = y'$ with $y' \neq y$. Since the first explanation is actual, it follows immediately that $(M, \boldsymbol{u}) \models \boldsymbol{X} = \boldsymbol{x} \wedge \boldsymbol{W_2} = \boldsymbol{w_2} \wedge Y = y$. Combining the second explanation with Proposition 8 we get that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W_2} = \boldsymbol{w_2}]Y = y'$.

Note that we did not require $\boldsymbol{X}$ to be minimal in either direction, and thus the conditions as stated without minimality of $\boldsymbol{X}$ are equivalent. Therefore the conditions that include the minimality of $\boldsymbol{X}$ are also equivalent, which is what we had to prove. ∎

Proposition 18 *If $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y$ and there exists a dominating explanation $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{A} = \boldsymbol{a}), \boldsymbol{B})$ for some values $\boldsymbol{x}'$ and $\boldsymbol{a} \subseteq \boldsymbol{w}$, then $\boldsymbol{X} = \boldsymbol{x}'$ can replace $\boldsymbol{X} = \boldsymbol{x}$.*

Proof: Assume $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y$ and $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{A} = \boldsymbol{a}), \boldsymbol{B})$ is a dominating explanation of $Y = y$, with $\boldsymbol{a} \subseteq \boldsymbol{w}$. We show that $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{B})$ is a sufficient explanation of $Y = y$, from which the result follows.

Let $\boldsymbol{C} = \mathcal{V} \setminus (\boldsymbol{X} \cup \boldsymbol{A} \cup \boldsymbol{B})$. From the definition of sufficient explanations, we know that for all $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ and all $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$, we have that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{A} \leftarrow \boldsymbol{a}, \boldsymbol{C} \leftarrow \boldsymbol{c}]\boldsymbol{B} = \boldsymbol{b}$ for some $\boldsymbol{b} \in \mathcal{R}(\boldsymbol{B})$ that includes y.

Let $\boldsymbol{D} = \mathcal{V} \setminus (\boldsymbol{X} \cup \boldsymbol{W} \cup \boldsymbol{B})$, $\boldsymbol{F} = \boldsymbol{W} \setminus \boldsymbol{A}$, and let $\boldsymbol{f}$ be the restriction of $\boldsymbol{w}$ to $\boldsymbol{F}$. Note that $\boldsymbol{C} = \boldsymbol{F} \cup \boldsymbol{D}$. From the previous paragraph it follows that for all $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ and all $\boldsymbol{d} \in \mathcal{R}(\boldsymbol{D}$, we have that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{A} \leftarrow \boldsymbol{a}, \boldsymbol{F} \leftarrow \boldsymbol{f}, \boldsymbol{D} \leftarrow \boldsymbol{d}]\boldsymbol{B} = \boldsymbol{b}$, and thus $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}, \boldsymbol{D} \leftarrow \boldsymbol{d}]\boldsymbol{B} = \boldsymbol{b}$, which is what had to be shown. ▮

Theorem 20 *If $\boldsymbol{X_1} = \boldsymbol{x_1}$ rather than $\boldsymbol{X_1} = \boldsymbol{x_1}'$ is a counterfactual explanation of $Y = y$ in $(M, \boldsymbol{u})$ (relative to some $(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$) then for some $\boldsymbol{X_2} \subseteq \boldsymbol{X_1}$, $\boldsymbol{X_2} = \boldsymbol{x_2}$ rather than $\boldsymbol{X_2} = \boldsymbol{x_2}'$ is an actual cause of $Y = y$ in $(M, \boldsymbol{u})$ (where $\boldsymbol{x_2}$ and $\boldsymbol{x_2}'$ are the relevant restrictions to $\boldsymbol{X_2}$).*

Proof: Assume $\boldsymbol{X_1} = \boldsymbol{x_1}$ rather than $\boldsymbol{X_1} = \boldsymbol{x_1}'$ is a counterfactual explanation of $Y = y$ in $(M, \boldsymbol{u})$ relative to $(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$. This means that $((\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is an actual sufficient explanation of $Y = y$ and $((\boldsymbol{X_1} = \boldsymbol{x_1}', \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y'$ with $y' \neq y$.

Let $(\boldsymbol{T} = \boldsymbol{t}, \boldsymbol{S})$ be a good sufficient explanation of $Y = y$, i.e., an actual sufficient explanation of $Y = y$ that dominates $((\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ and cannot itself be dominated by another actual sufficient explanation of $Y = y$. (It is easy to see that such an explanation must exist: one can simply keep removing elements from $((\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ until no further element can be removed while still remaining a sufficient explanation of $Y = y$.)

Let $\boldsymbol{X_2} = \boldsymbol{X_1} \cap \boldsymbol{T}$. We now show that $\boldsymbol{X_2} \neq \emptyset$ by a reductio.

Assume $\boldsymbol{X_2} = \emptyset$. This means that $\boldsymbol{T} \subseteq \boldsymbol{W}$. Also, $\boldsymbol{S} \subseteq \boldsymbol{N}$. Let $\boldsymbol{s} \in \mathcal{R}(\boldsymbol{S})$ and $\boldsymbol{n} \in \mathcal{R}(\boldsymbol{N})$ be the actual values of $\boldsymbol{S}$ and $\boldsymbol{N}$ in $(M, \boldsymbol{u})$.

Let $\boldsymbol{C} = \mathcal{V} \setminus (\boldsymbol{X_1} \cup \boldsymbol{W} \cup \boldsymbol{N})$. Given that $((\boldsymbol{X_1} = \boldsymbol{x_1}', \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y'$, we have that for all $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$, $(M, \boldsymbol{u}) \models [\boldsymbol{X_1} \leftarrow \boldsymbol{x_1}', \boldsymbol{W} \leftarrow \boldsymbol{w}, \boldsymbol{C} \leftarrow \boldsymbol{c}]Y = y'$.

Let $\boldsymbol{D} = \mathcal{V} \setminus (\boldsymbol{T} \cup \boldsymbol{S})$. Given that $(\boldsymbol{T} = \boldsymbol{t}, \boldsymbol{S})$ is a sufficient explanation of $Y = y$, we have that for all $\boldsymbol{d} \in \mathcal{R}(\boldsymbol{D})$, $(M, \boldsymbol{u}) \models [\boldsymbol{T} \leftarrow \boldsymbol{t}, \boldsymbol{D} \leftarrow \boldsymbol{d}]Y = y$.

By our assumption, $\boldsymbol{X_1} \subseteq (\boldsymbol{W} \setminus \boldsymbol{T})$. Thus $\boldsymbol{D} = \boldsymbol{X_1} \cup \boldsymbol{C} \cup (\boldsymbol{W} \setminus (\boldsymbol{T} \cup \boldsymbol{X_1})) \cup (\boldsymbol{N} \setminus \boldsymbol{S})$. Therefore from the previous paragraph we get that for all $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$, $(M, \boldsymbol{u}) \models [\boldsymbol{X_1} \leftarrow \boldsymbol{x_1}', \boldsymbol{W} \leftarrow \boldsymbol{w}, \boldsymbol{C} \leftarrow \boldsymbol{c}]Y = y$. This contradicts the paragraph before the previous, and therefore $\boldsymbol{X_2} \neq \emptyset$.

Let $\boldsymbol{W_2} = \boldsymbol{T} \setminus \boldsymbol{X_2}$. Then we can conclude that $(\boldsymbol{X_2} = \boldsymbol{x_2}, \boldsymbol{W_2} = \boldsymbol{w_2}, \boldsymbol{S})$ is a good sufficient explanation of $Y = y$, where $\boldsymbol{x_2}$ is the restriction of $\boldsymbol{x_1}$ to $\boldsymbol{X_2}$, and $\boldsymbol{w_2}$ is the restriction of $\boldsymbol{W}$ to $\boldsymbol{W_2}$.

Remains to be shown that $\boldsymbol{x_2}'$ cannot replace $\boldsymbol{x_2}$ in this explanation, where $\boldsymbol{x_2}'$ is the restriction of $\boldsymbol{x_1}'$ to $\boldsymbol{X_2}$. The former just means that $((\boldsymbol{X_2} = \boldsymbol{x_2}', \boldsymbol{W_2} = \boldsymbol{w_2}), \boldsymbol{S_2})$ is not a sufficient explanation of $Y = y$ for any $\boldsymbol{S_2} \subseteq \boldsymbol{S}$.

Again we proceed by a reductio: assume that $((\boldsymbol{X_2} = \boldsymbol{x_2}', \boldsymbol{W_2} = \boldsymbol{w_2}), \boldsymbol{S_2})$ is a sufficient explanation of $Y = y$. Let $\boldsymbol{F} = \mathcal{V} \setminus (\boldsymbol{X_2} \cup \boldsymbol{W_2} \cup \boldsymbol{S_2})$. We have that for all $\boldsymbol{f} \in \mathcal{R}(\boldsymbol{F})$, $(M, \boldsymbol{u}) \models [\boldsymbol{X_2} \leftarrow \boldsymbol{x_2}', \boldsymbol{W_2} \leftarrow \boldsymbol{w_2}, \boldsymbol{F} \leftarrow \boldsymbol{f}]Y = y$. In particular, we have that $(M, \boldsymbol{u}) \models [\boldsymbol{X_1} \leftarrow \boldsymbol{x_1}', \boldsymbol{W} \leftarrow \boldsymbol{w}]Y = y$.

Recall that $((\boldsymbol{X_1} = \boldsymbol{x_1}', \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y'$ with $y' \neq y$. Using Proposition 8, we get that $(M, \boldsymbol{u}) \models [\boldsymbol{X_1} \leftarrow \boldsymbol{x_1}', \boldsymbol{W} \leftarrow \boldsymbol{w}]Y = y'$. This contradicts the result in the previous paragraph, which concludes the proof. ∎

Proposition 22 *If $\boldsymbol{X} = \boldsymbol{x}$ is a direct cause of $Y = y$ in $(M, \boldsymbol{u})$ then there exist values $\boldsymbol{x}'$ such that $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an actual cause of $Y = y$.*

Proof: Assume $\boldsymbol{X} = \boldsymbol{x}$ is a direct cause of $Y = y$ in $(M, \boldsymbol{u})$, i.e., it is part of a direct good sufficient explanation $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w})$ of $Y = y$ in $(M, \boldsymbol{u})$. This means that all that remains to be shown, is that there exist values $\boldsymbol{x}' \in \mathcal{R}(\boldsymbol{X})$ such that $(\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{W} = \boldsymbol{w})$ is not a direct sufficient explanation of $Y = y$.

We know that $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w})$ is directly sufficient for $Y = y$, and this does not hold if we remove any subset from either $\boldsymbol{X}$ or $\boldsymbol{W}$. Let $\boldsymbol{C} = \mathcal{V} \setminus (\boldsymbol{X} \cup \boldsymbol{W} \cup \{Y\})$. Then we have that for all $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$, $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{W} \leftarrow \boldsymbol{w}, \boldsymbol{C} \leftarrow \boldsymbol{c}]Y = y$. From the minimality of $\boldsymbol{X}$, it follows that there exists a $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$ and $\boldsymbol{x}' \in \mathcal{R}(\boldsymbol{X})$ so that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}, \boldsymbol{C} \leftarrow \boldsymbol{c}]Y \neq y$. Therefore $(\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{W} = \boldsymbol{w})$ is not a direct sufficient explanation of $Y = y$. ∎

Theorem 23 *If a causal model M satisfies* **Independence** *then for any $\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{N_1}, \boldsymbol{N_2}$, it holds that there exist values $\boldsymbol{w_1} \in \mathcal{R}(\boldsymbol{W_1})$ so that $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_1} = \boldsymbol{w_1}, \boldsymbol{N_1})$ is a good counterfactual explanation of $Y = y$ iff there exist values $\boldsymbol{w_2} \in \mathcal{R}(\boldsymbol{W_2})$ so that $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_2} = \boldsymbol{w_2}, \boldsymbol{N_2})$ is a good counterfactual explanation of $Y = y$.*

Proof: Given the symmetry, it suffices to prove the implication in one direction. We invoke Theorem 16 so that we can use counterfactual dependence instead of good counterfactual explanations.

Assume $Y = y$ counterfactually depends on $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ relative to $\boldsymbol{W_1} = \boldsymbol{w_1}$. This means that $\boldsymbol{X}$ is a minimal set such that $(M, \boldsymbol{u}) \models \boldsymbol{X} = \boldsymbol{x} \wedge \boldsymbol{W_1} = \boldsymbol{w_1} \wedge Y = y$ and $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W_1} \leftarrow \boldsymbol{w_1}]Y \neq y$. Take any set $\boldsymbol{W_2} \subseteq (\mathcal{V} \setminus (\boldsymbol{X} \cup \{Y\}))$, and let $\boldsymbol{w_2}$ be the values of $\boldsymbol{W_2}$ in $(M, \boldsymbol{u})$. First, note that we have $(M, \boldsymbol{u}) \models \boldsymbol{X} = \boldsymbol{x} \wedge \boldsymbol{W_2} = \boldsymbol{w_2} \wedge Y = y$. Second, since none of the members of $\boldsymbol{X}$ are parents of any of the members of $\boldsymbol{W_2}$, we also have that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}']\boldsymbol{W_2} = \boldsymbol{w_2} \wedge Y = y'$. Thus we also have that $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W_2} \leftarrow \boldsymbol{w_2}]Y = y'$. As we did not use the fact that $\boldsymbol{X}$ is a minimal set, and thus both statements are equivalent when ignoring minimality on both sides, they are also equivalent when including minimality on both sides, which is what we have to prove. ∎

Theorem 24 *If a causal model M satisfies* **Independence** *then the following statements are all equivalent:*

- *$\boldsymbol{X} = \boldsymbol{x}$ is a direct cause of $Y = y$ in $(M, \boldsymbol{u})$.*
- *there exist values $\boldsymbol{x}'$ so that $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an actual cause of $Y = y$ in $(M, \boldsymbol{u})$.*
- *$\boldsymbol{X} = \boldsymbol{x}$ is part of a good sufficient explanation of $Y = y$ in $(M, \boldsymbol{u})$.*

Proof: The implication from the first statement to the second is a direct consequence of Proposition 22.

The implication from the second statement to the third follows from the definition of actual cause.

The implication from the third statement to the first follows from Theorem 12, which shows that under **Independence** we may replace sufficiency with direct sufficiency, and thus the result follows from the definition of direct cause. ∎