Supplementary Material for ‘Multi-Level Cause-Effect Systems’

A SUFFICIENT CAUSAL DESCRIPTION THEOREM, PARTS 1 AND 2

Theorem 1 (Sufficient Causal Description)**.** *Let $(\mathcal{I}, \mathcal{J})$ be a causal ml-system and let C and E be its fundamental cause and effect. Let $\mathbf{E}$ be E applied sample-wise to a sample from the system (so that e.g. $\mathbf{E}(j_1, \cdots, j_k) = (E(j_1), \cdots, E(j_k))$). Then:*

1. *Among all the partitions of $\mathcal{J}$, $\mathbf{E}$ is the* minimal *sufficient statistic for $P(J \mid man(i))$ for any $i \in \mathcal{I}$, and*
2. *C and E losslessly recover $P(j \mid man(i))$. No other (subsidiary) causal variable losslessly recovers $P(j \mid man(i))$. Any other partition is either finer than C, E or does not define unambiguous manipulations. In this sense, the fundamental causal partition corresponds to the coarsest partition that losslessly recovers $P(j \mid man(i))$.*

Proof. 1. We first prove that $\mathbf{E}$ is a sufficient statistic. Recall that we assumed $\mathcal{J}$ to be discrete, although possibly of vast cardinality. For any $j_k \in \mathcal{J}$, write $P(j_k \mid \text{man}(i)) = p_{j_k}$ for the corresponding categorical distribution parameter. Let `range`$(E) = \{E_1, \cdots, E_M\}$ be the set of causal classes of J. By Definition 3 there is a number of “template” probabilities $p_{E_1}, \cdots, p_{E_M}$ such that $p_{j_k} = p_{E_k}$ if and only if $E(j_k) = E_k$. Consider an i.i.d. sample $\mathbf{j} = j_1, \cdots, j_l$ from $P(J \mid \text{man}(i))$. Then

$$\begin{aligned} P(j_1, \cdots, j_l \mid \text{man}(i)) &= \Pi_{k=1}^{l} p_{j_k} \\ &= \Pi_{m=1}^{M} p_{E_m}^{\#(E_m)}, \end{aligned}$$

where $\#(E_m) \triangleq \Sigma_{k=1}^{l} \mathbb{1}\{E(j_k) == E_m\}$ is the number of samples with causal class E_m. Since the sample density depends on the samples only through C and E it follows from Fisher’s factorization theorem that $\mathbf{E}$ is a sufficient statistic for $P(J \mid \text{man}(i))$ for any $i \in \mathcal{I}$.

Now, we prove the minimality of E among all the partitions of $\mathcal{J}$. Consider first any refinement of E. One can directly apply the reasoning above to show that the cell assignment in such a partition is also a sufficient statistic. However, any refinement is not the *minimal* sufficient statistic, as the fundamental causal partition is its coarsening— and thus also its function. Now, consider any partition that is not the fundamental causal partition, and is not its refinement. Call it E'. Assume, for contradiction, that $\mathbf{E}'$ is a sufficient statistic for $P(J \mid \text{man}(i))$. Then, by the factorization theorem, $P(j_1, \cdots, j_k \mid \text{man}(i))$ would factorize as $h(j_1, \cdots, j_k) g(E'(j_1), \cdots, E'(j_k))$, where h does not depend on the parameters p_{j_l}. Now, take some j_1^1, j_1^2 such that $E(j_1^1) \neq E(j_1^2)$ but $E'(j_1^1) = E'(j_1^2)$ (such a pair must exists since E' is not a refinement of E and is not equal to it). Then

$$\frac{P(j_1^1, j_2, \cdots, j_k \mid \text{man}(i))}{P(j_1^2, j_2, \cdots, j_k \mid \text{man}(i))} = \frac{p_{E(j_1^1)}}{p_{E(j_1^2)}},$$

$$\begin{aligned} &\frac{P(j_1^1, j_2, \cdots, j_k \mid \text{man}(i))}{P(j_1^2, j_2, \cdots, j_k \mid \text{man}(i))} = \\ &\qquad = \frac{h(j_1^1, \cdots, j_k) g(E'(j_1^1), \cdots, E'(j_k))}{h(j_1^2, \cdots, j_k) g(E'(j_1^2), \cdots, E'(j_k))} \\ &\qquad = \frac{h(j_1^1, \cdots, j_k)}{h(j_1^2, \cdots, j_k)} \end{aligned}$$

which, as already stated, does not depend on the parameters of the distribution – a contradiction.

2. That $P(J \mid \text{man}(i))$ can be recovered from C and E follows directly from the definition of a causal ml-system and its fundamental causal partition. That it cannot be recovered losslessly from any partition that is not a refinement of C and E follows again from the fact that for any such partitions C' and E' there must be is at least one pair $(i_1, j_1), (i_2, j_2)$ for which $p(E'(j_1) \mid \text{do}(C'(i_1))) = p(E'(j_2) \mid \text{do}(C'(i_2)))$ even though $p(j_1 \mid \text{man}(i_1)) \neq p(j_2 \mid \text{man}(i_2))$. □

We note that the first part of Theorem 1 indicates that $\mathbf{E}$ is only a minimal sufficient statistic among all partitions of $\mathcal{J}$, i.e. among the set of possible causal variables. It is not the minimal sufficient statistic over all possible sufficient statistics for $P(J \mid \text{man}(i))$. In particular, a histogram is a minimal sufficient statistic for the multinomial distribution and is a function of $\mathbf{E}$, but a histogram does not correspond to a partition of $\mathcal{J}$.

B DETAILS AND IMPLEMENTATION OF ALGORITHM 1

First, the algorithm uses a density learning routine to estimate $P(J \mid \text{man}(I))$ given the samples. We don’t specify the density learning routine, as that is highly problem-dependent. In our experiments, dimensionality reduction with autoencoders (Hinton and Salakhutdinov, 2006) followed by kernel density estimation worked well. More sophisticated approaches are readily available, for example

RNADE (Uria et al., 2013). Steps 2 and 3 constitute the core of the algorithm: In Step 2, a vector of (estimated) densities $[P(j_1 \mid \mathrm{man}(i)), \cdots, P(j_N \mid \mathrm{man}(i))]$ is calculated for each i_k in the dataset ($1 \leq k \leq N$). That is, each i_k corresponds to a vector that contains information about the probability of each j_l ($1 \leq l \leq N$) occurring given a manipulation $\mathrm{man}(i_k)$ (note that in the original dataset, i_k might have only appeared as paired with one effect j_k, sampled from $P(J \mid \mathrm{man}(i))$). Similarly, Step 3, computes for each j_l a vector of estimated densities of j_l occurring given an intervention on each i_k.

Clustering these vectors (Step 4 & 5) makes it possible to group together all the i's with similar effects, and all the j's with similar causes — that is, to learn the fundamental causal partition. The number of cells of the fundamental partition is unknown in advance, but it is safe to overcluster the data. Our implementation uses the Dirichlet Process Gaussian Mixture Model (Rasmussen, 1999) for clustering with a flexible number of clusters, but again the algorithm stays clustering-routine-agnostic.

After the initial clustering it should now be easy to merge clusters belonging to the same true causal class, as the probabilistic patterns of mergeable clusters are expected to be similar. The macro-variable cause/effect probability vectors are estimated in Steps 8 and 9. These are analogues to the micro-variable cause/effect density vectors estimated in Step 2. However, instead of estimating the density of the micro-variable data, they count the normalized histograms of conditional probabilities of the $\mathcal{J}$ cluster given the $\mathcal{I}$ clusters. These histograms are aggregates of large numbers of datapoints, and should smooth out errors in the original density estimation. Thus, even if the original clustering algorithm overestimates the number of cells in the fundamental partition, we can hope to be able to merge them based on similarities in the macro-variable histogram vectors. In our experiment, we merge the macro-variable cause/effect probabilities by thresholding the KL-divergence between any two vectors belonging to the same cluster. However, since the number of datapoints to cluster is likely to be very small, the best solution in practice is to cluster them by hand.

By Step 8, the algorithm returns causal labels for the original data samples. These labeled samples can be used to visualize the fundamental causes and effects using the original data samples. To generalize the fundamental cause and effect to the whole $\mathcal{I}$ and $\mathcal{J}$ space, the algorithm trains a classifier using the original data and the learned causal labels.

C THE FUNDAMENTAL CAUSAL COARSENING THEOREM

Theorem 2 (Fundamental Causal Coarsening)**.** *Among all the generative distributions of the form shown in Fig. 2 (main text) which induce given observational partition* $(\Pi_o(\mathcal{I}), \Pi_o(\mathcal{J}))$*:*

1. *The subset of distributions that induce a fundamental causal partition $\Pi_c(\mathcal{I})$ that is* not *a coarsening of the observational partition $\Pi_o(\mathcal{I})$ is Lebesgue measure zero, and*
2. *The subset of distributions that induce a fundamental causal partition $\Pi_c(\mathcal{J})$ that is* not *a coarsening of the observational partition $\Pi_o(\mathcal{J})$ is Lebesgue measure zero.*

Proof. (1) Let E be the fundamental effect of the system. Then $\Pi_c(\mathcal{I})$ and E constitute precisely the "causal partition" and "target behavior" of the system and $\Pi_o(\mathcal{I})$ constitutes the "observational partition" of the system, as defined by Chalupka et al. (2015). Thus, the proof of the Causal Coarsening Theorem by Chalupka et al. (2015) applies directly and proves (1).

(2) While we cannot directly use the proof of Chalupka et al. (2015), we follow a very similar proof strategy. The only difference is in details of the algebra. We first lay out the proof strategy. Let $j_1, j_2 \in \mathcal{J}$. We need to show that if $P(j_1 \mid i) = P(j_2 \mid i)$ *for every* $i \in \mathcal{I}$, then also $P(j_1 \mid \mathrm{man}(i)) = P(j_2 \mid \mathrm{man}(i))$ for every i (for all the distributions compatible with given observational partition, except for a set of measure zero). The proof is split into two parts: (i) Express the theorem as a polynomial constraint on the space of all $P(i, j, h)$ distributions. (ii) Show that the polynomial constraint is not trivial. This, by Meek (1995), implies that among *all* $P(i, j, h)$ distributions, the fundamental causal partition on J is a coarsening of the fundamental observational partition. (iii) Prove that (i) and (ii) apply to "all the distributions which induce a given observational partition" by showing that this restriction results in a simple reparametrization of the distribution space.

(2i) Let H be the hidden variable of the system, with cardinality K; let J have cardinality N and I cardinality M. We can factorize the joint on I, J, H as $P(J, I, H) = P(J \mid H, I)P(I \mid H)P(H)$. $P(J \mid H, I)$ can be parametrized by $(N-1) \times K \times M$ parameters, $P(I \mid H)$ by $(M-1) \times K$ parameters, and $P(H)$ by $K-1$ parameters, all of which are independent.

Call the parameters, respectively,

$$\alpha_{j,h,i} \triangleq P(J = j \mid H = h, I = i)$$
$$\beta_{i,h} \triangleq P(I = i \mid H = h)$$
$$\gamma_h \triangleq P(H = h)$$

We will denote parameter vectors as

$$\begin{aligned}\alpha &= (\alpha_{j_1,h_1,i_1}, \cdots, \alpha_{j_{N-1},h_K,i_M}) \in \mathbb{R}^{(N-1)\times K\times M}\\ \beta &= (\beta_{i_1,h_1}, \cdots, \beta_{i_{N-1},h_K}) \in \mathbb{R}^{(M-1)\times K}\\ \gamma &= (\gamma_{h_1}, \cdots, \gamma_{h_{K-1}}) \in \mathbb{R}^{K-1},\end{aligned}$$

where the indices are arranged in lexicographical order. This creates a one-to-one correspondence of each possible joint distribution $P(J, H, I)$ with a point $(\alpha, \beta, \gamma) \in P[\alpha, \beta, \gamma] \subset \mathbb{R}^{(N-1)\times K^2(K-1)\times M(M-1)}$, where $P[\alpha, \beta, \gamma]$ is the $(N-1)\times K^2(K-1)\times M(M-1)$-dimensional simplex of multinomial distributions.

To proceed with the proof, we pick any point in the $P(J \mid H, I) \times P(H)$ space: that is, we fix the values of α and γ. The only free parameters are now the $\beta_{i,h}$; varying these values creates a subset of the space of all the distributions which we will call

$$P[\beta; \alpha, \gamma] = \{(\alpha, \beta, \gamma) \mid \beta \in [0, 1]^{(M-1)\times K}\}.$$

$P[\beta; \alpha, \gamma]$ is a subset of $P[\alpha, \beta, \gamma]$ isometric to the $[0, 1]^{(M-1)\times K}$-dimensional simplex of multinomials. We will use the term $P[\beta; \alpha, \gamma]$ to refer both the subset of $P[\alpha, \beta, \gamma]$ and the lower-dimensional simplex it is isometric to, remembering that the latter comes equipped with the Lebesgue measure on $\mathbb{R}^{(M-1)\times K}$.

Now we are ready to show that the subset of $P[\beta; \alpha, \gamma]$ which does not satisfy the Fundamental Causal Coarsening constraint on $\mathcal{J}$ is of measure zero with respect to the Lebesgue measure. To see this, first note that since α and γ are fixed, the manipulation probabilities $p(j \mid \text{man}(i)) = \sum_h \alpha_{j,h,i}\gamma_h$ are fixed for each $i \in \mathcal{I}, j \in \mathcal{J}$. The Fundamental Causal Coarsening constraint on $\mathcal{J}$ says "If for some $j_1, j_2 \in \mathcal{J}$ we have $p(j_1 \mid i) = p(j_2 \mid i)$ *for every* $i \in \mathcal{I}$, then also $p(j_1 \mid \text{man}(i)) = p(j_2 \mid \text{man}(i))$ for every i." The subset of $P[\beta; \alpha, \gamma]$ of all distributions that *do not* satisfy the constraint consists of the $P(J, H, I)$ for which for some $j_1, j_2 \in \mathcal{J}$ it holds that

$$\forall_i P(j_1 \mid i) = P(j_2 \mid i) \text{ and } P(j_1 \mid \text{man}(i)) \neq P(j_2 \mid \text{man}(i)).$$

We want to prove that this subset is measure zero. To this aim, *take any pair* j_1, j_2 *and an i for which* $p(j_1 \mid$ *man*$(i)) \neq p(j_2 \mid$ *man*$(i))$ [Assumption 1]. Note that if such a configuration does not exist, then the Fundamental Causal Coarsening constraint holds for all the distributions in $P[\beta; \alpha, \gamma]$ and the proof is done. We can write

$$\begin{aligned}P(j_1 \mid i) &= \sum_h P(j_1 \mid h, i)P(h \mid i)\\ &= \frac{1}{P(i)} \sum_h P(j_1 \mid h, i)P(i \mid h)P(h).\end{aligned}$$

Since the same equation applies to $P(j_2 \mid i)$, the constraint $P(j_1 \mid i) = P(j_2 \mid i)$ can be rewritten as

$$\begin{aligned}&\sum_h P(j_1 \mid h, i)P(i \mid h)P(h)\\ &= \sum_h P(j_2 \mid h, i)P(i \mid h)P(h)\end{aligned}$$

which we can rewrite in terms of the independent parameters as

$$\sum_h [\alpha_{j_1,h,i} - \alpha_{j_2,h,i}]\beta_{h,i}\gamma_h = 0, \tag{1}$$

which is a polynomial constraint on $P[\beta; \alpha, \gamma]$. By a simple algebraic lemma (proven by Okamoto, 1973), if the above constraint is not trivial (that is, if there exists β for which the constraint does not hold), the subset of $P[\beta; \alpha, \gamma]$ on which it holds is measure zero.

(2ii) To see that Eq. (1) does not hold for every distribution, take a distribution distribution with $\beta_{h^*,i} = 1$ (and thus $\beta_{h,i} = 0$ for any $h \neq h^*$). The equation then reduces to

$$(\alpha_{j_1,h^*,i} - \alpha_{j_2,h^*,i})\gamma_{h^*} = 0.$$

Thus, Eq. (1) implies that $\alpha_{j_1,h^*,i} = \alpha_{j_2,h^*,i}$ or $\gamma_{h^*} = 0$. Since h^* is general, $\alpha_{j_1,h,i} = \alpha_{j_2,h,i}$ or $\gamma_h = 0$ for any h. By definition $p(j_1 \mid \text{man}(i)) - p(j_2 \mid \text{man}(i))$ equals $\sum_h (\alpha_{j_1,h,i} - \alpha_{j_2,h,i})\gamma_h$ which, in the context of the previous sentence, is 0. Thus our reasoning so far implies that $p(j_1 \mid \text{man}(i)) = p(j_2 \mid \text{man}(i))$. This is in direct contradiction with Assumption 1.

We have now shown that the subset of $P[\beta; \alpha, \gamma]$ which consists of distributions for which $P(j_1 \mid i) = P(j_2 \mid i)$, but $p(j_1 \mid \text{man}(i)) \neq p(j_2 \mid \text{man}(i)$ for some i, is Lebesgue measure zero. Since there are only finitely many pairs of images j_1, j_2 for which the latter condition holds, the subset of $P[\beta; \alpha, \gamma]$ of distributions which violate the Causal Coarsening constraint is also Lebesgue measure zero (a finite sum of measure zero sets is measure zero). The remainder of the proof is a direct application of Fubini's theorem.

For each α, γ, call the (measure zero) subset of $P[\beta; \alpha, \gamma]$ that violates the Causal Coarsening constraint $z[\alpha, \gamma]$. Let $Z = \cup_{\alpha,\gamma} z[\alpha, \gamma] \subset P[\alpha, \beta, \gamma]$ be the set of all the joint distributions which violate the Causal Coarsening constraint. We want to prove that $\mu(Z) = 0$, where μ is the Lebesgue measure. To show this, we will use the indicator function

$$\hat{z}(\alpha, \beta, \gamma) = \begin{cases} 1 & \text{if } \beta \in z[\alpha, \gamma],\\ 0 & \text{otherwise.}\end{cases}$$

By the basic properties of positive measures we have

$$\mu(Z) = \int_{P[\alpha,\beta,\gamma]} \hat{z}\; d\mu.$$

It is a standard application of Fubini's Theorem for the Lebesgue integral to show that the integral in question equals zero. For simplicity of notation, let

$$\begin{aligned} \mathcal{A} &= \mathbb{R}^{(N-1)\times K\times M} \\ \mathcal{B} &= \mathbb{R}^{(M-1)\times K} \\ \mathcal{G} &= \mathbb{R}^{K-1}. \end{aligned}$$

We have

$$\begin{aligned} \int_{P[\alpha,\beta,\gamma]} \hat{z}\; d\mu &= \int_{\mathcal{A}\times\mathcal{B}\times\mathcal{G}} \hat{z}(\alpha,\beta,\gamma)\, d(\alpha,\beta,\gamma) \\ &= \int_{\mathcal{A}\times\mathcal{G}} \int_{\mathcal{B}} \hat{z}(\alpha,\beta,\gamma)\, d(\beta)\; d(\alpha,\gamma) \\ &= \int_{\mathcal{A}\times\mathcal{G}} \mu(z[\alpha,\gamma])\; d(\alpha,\gamma) \qquad (2) \\ &= \int_{\mathcal{A}\times\mathcal{G}} 0\, d(\alpha,\gamma) \\ &= 0. \end{aligned}$$

Equation (2) follows as $\hat{z}$ restricted to $P[\beta;\alpha,\gamma]$ is the indicator function of $z[\alpha,\gamma]$.

This completes the proof that Z, the set of joint distributions over J, H and I that violate the Causal Coarsening constraint, is measure zero.

(2iii) Finally, we show that (2i) and (2ii) apply if we fix an observational partition on $\mathcal{J}$ *a priori*. Fixing the observational partition means fixing a set of observational constraints (OCs)

$$\begin{aligned} \forall_i p(j_1^1 \mid i) = \cdots = p(j_{N_1}^1 \mid i) = p^1, \\ \vdots \qquad\qquad \\ \forall_i p(j_1^L \mid i) = \cdots = p(j_{N_L}^L \mid i) = p^L, \end{aligned}$$

where $1 \leq L \leq N$ is the number of observational classes of $\mathcal{J}$ and N_l is the cardinality of the lth observational class (so that $N = \sum_l N_l$), and $p^1, \cdots, p^L$ are the numerical values of the observational constraints.

Since $P(J,H,I) = P(H \mid J,I)P(J \mid I)P(I)$, $P(j \mid i)$ is an independent parameter in the unrestricted $P(J,H,I)$, and the OCs reduce the number of independent parameters of the joint by $M\sum_{l=1}^{L}(N_l - 1)$, where M is the cardinality of I. We want to express this parameter-space reduction in terms of the α, β and γ parameterization from (2i) and (2ii). To do this, note first that we can write, for any j_n^l,

$$\sum_h p(j_n^l, h, i) = p^l \sum_h p(h,i).$$

Now, pick any h^* for which $p(h^*, i) \neq 0$. Then we can write

$$p(j_n^l, h^*, i) = p^l \sum_h p(h,i) - \sum_{h\neq h^*} p(j_n^l, h, i).$$

In terms of the α, β, γ parameterization, this equation becomes

$$\alpha_{j_n^l,h^*,i}\beta_{h^*,i}\gamma_h = p^l \sum_h \beta_{h,i}\gamma_h - \sum_{h\neq h^*} \alpha_{j_n^l,h,i}\beta_{h,i}\gamma_h$$

or

$$\alpha_{j_n^l,h^*,i} = \frac{p^l \sum_h \beta_{h,i}\gamma_h - \sum_{h\neq h^*} \alpha_{j_n^l,h,i}\beta_{h,i}\gamma h}{\beta_{h^*,i}\gamma_h}. \qquad (3)$$

The full set of the OCs is equivalent to the full set of equations of this form, one for each possible (j_n^l, i) combination (to the total of $M \times (N - L)$ equations as expected). Thus, we can express the range of $P(J,H,I)$ distributions consistent with a given observational partition $\Pi_o(\mathcal{J})$ in terms of the full range of β, γ parameters and a restricted number of independent α parameters. □

References

K. Chalupka, P. Perona, and F. Eberhardt. Visual Causal Feature Learning. In *Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 181–190. AUAI Press, 2015.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411–418, 1995.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763–765, 1973.

C. E. Rasmussen. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, pages 554–560, 1999.

B. Uria, I. Murray, and H. Larochelle. RNADE: The real-valued neural autoregressive density-estimator. *Advances in Neural Information Processing Systems*, pages 2175–2183, 2013.