Manuscript version: Author’s Accepted Manuscript
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/170095

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher’s statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
Towards Computing an Optimal Abstraction for Structural Causal Models

Fabio Massimo Zennaro  
Paolo Turri
Theodoros Damoulas

1University of Warwick, Coventry, United Kingdom

Abstract

Working with causal models at different levels of abstraction is an important feature of science. Existing work has already considered the problem of expressing formally the relation of abstraction between causal models. In this paper, we focus on the problem of learning abstractions. We start by defining the learning problem formally in terms of the optimization of a standard measure of consistency. We then point out the limitation of this approach, and we suggest extending the objective function with a term accounting for information loss. We suggest a concrete measure of information loss, and we illustrate its contribution to learning new abstractions.

1 INTRODUCTION

Understanding causality is a key challenge for modern artificial intelligence (AI) [Schölkopf et al., 2021]. Structural causal models (SCM) [Pearl, 2009] are well-established tools used in statistics and computer science to describe causal systems and to express causal assumptions in a graphical form. One could, for instance, describe the relation between smoking and cancer in a model with few variables of interest represented as nodes (e.g., environment stress, smoking, and cancer), and with directed arcs denoting the hypothesised causal relationships (see Figure 1a).

An SCM is formulated over a set of relevant variables, some of which may not be as important for analysing the problem at hand. In our example, for instance, we may not be interested in explicitly modelling the role of the environment in causing smoking or cancer, but it could be sufficient to have a smaller model incorporating only the latter variables (see Figure 1b). Being able to work with SCMs at different levels of abstraction allows us to adjust to our available computational resources, while still getting meaningful results; it would also allow us to integrate data that may have been collected with models at different resolutions.

But what is the “right” abstraction of an SCM? Answering this question in a rigorous fashion requires tackling several challenges, among which how to define mathematically a relation of abstraction and how to formalize a notion of consistency among models. Answers to these two problems have recently been proposed in the literature [Rubenstein et al., 2017, Beckers and Halpern, 2019, Rischel, 2020] and, building on these contributions, we can now establish a relationship between SCMs and assess their consistency, as sketched in Figure 1c. Despite the recent progress in the literature, the question still remains of how to order abstractions in terms of gains and losses with respect to the original model and how to compute an optimal abstraction.

Contribution. In this paper we address the problem of computing an optimal abstraction for SCMs. We define several concrete subproblems, ranging from the simpler question of finding an abstraction between two fully specified SCMs (Figure 2a) to the harder question of being given a starting model and jointly finding an abstraction and an abstracted model (Figure 2b). We phrase the problem of learning an abstraction as an optimization problem where the objective is to maximize key properties of an abstraction. We start by considering the optimization of a standard measure of consistency; however, this approach could lead to optimal, yet doubtfully useful solutions, such as identities or the collapse of all the variables and outcomes onto a single value (Figure 2c). Therefore, we suggest the introduction of a measure that accounts for the information being lost in an abstraction, and which will be used in combination with consistency. We provide the definition of an optimization problem and we illustrate preliminary results on our motivating example showing how a measure of consistency and a measure of information loss capture different aspects and properties of an abstraction. Throughout the paper, we will illustrate ideas relying on a motivating example, for which

Accepted for the Causal Representation Learning workshop at the 38th Conference on Uncertainty in Artificial Intelligence (UAI CRL 2022).
Figure 1: Smoking (S) and cancer (C), at different levels. (a) A simple model, considering the environment (E). (b) Abstracting away from the environment. (c) A sketch of the relationship between the original model and its abstracted version.

Figure 2: Different abstraction problems. (a) Given two SCMs, learn their relative abstraction. (b) Given an SCM, learn an abstracted model and their relative abstraction. (c) A trivial solution to (b).

2 PRELIMINARIES

2.1 STRUCTURAL CAUSAL MODELS

Definition 1 (SCM [Pearl, 2009]) A structural causal model (SCM) \( \mathcal{M} \) is a tuple \((X, \mathcal{U}, \mathcal{F}, \mathbf{P}(\mathcal{U}))\) with an underlying directed acyclic graph (DAG) \( G_{\mathcal{M}} \) where:

- \( X \) is a finite set of \( N \) endogenous random variables \( X_i \); each variable \( X_i \) is associated with a set \( \mathcal{M}[X_i] = \{x_1, x_2, ..., x_N\} \) of outcomes; we use the boldface notation \( \mathbf{X} \subseteq X \) for subsets of variables, and \( \mathcal{M}[X] = \prod_{X_i \in X} \mathcal{M}[X_i] \) for the Cartesian product of the sets of outcomes of the variables in \( X \).
- \( \mathcal{U} \) is a finite set of \( N \) exogenous random variables \( U_i \); one for each endogenous variable \( X_i \); each variable \( U_i \) is associated with a set \( \mathcal{M}[U_i] = \{u_1, u_2, ..., u_N\} \) of outcomes.
- \( \mathcal{F} \) is a finite set of \( N \) modular measurable structural functions \( f_i \), one for each endogenous variable \( X_i \); a structural function \( f_i : \mathcal{M}[\mathbf{PA}(X_i)] \times \mathcal{M}[U_i] \to \mathcal{M}[X_i] \), where \( \mathbf{PA}(X_i) \subseteq X \setminus X_i \) defines deterministically the value of the random variable \( X_i \) given the values of the variables in the set \( \mathbf{PA}(X_i) \) and \( U_i \).
- \( \mathbf{P}(\mathcal{U}) \) is a joint probability distributions over the exogenous variable \( U_i \).

Endogenous variables represents variables of interest, explained by deterministic mechanisms; exogenous variables capture stochastic factors of variance beyond the control of a modeler. Several common assumptions underlying this definition are explicitly stated in Appendix A.

Example. Let us define a simple toy SCM \( \mathcal{M} \) for the lung cancer scenario we introduced earlier:

- \( X = \{E, S, C\} \) is the set of endogenous variables containing three binary variables representing respectively level of stress due to the environment, habit of smoking, and presence of lung cancer;
• $\mathcal{U} = \{U_E, U_S, U_C\}$ is the set of exogenous variables containing three binary variables;
• $\mathcal{F} = \{f_E, f_S, f_C\}$ is the set of structural functions such that $E = f_E(U_E)$, $S = f_S(E, U_S)$, and $C = f_C(E, S, U_C)$;
• $P(\mathcal{U}) = P(U_E, U_S, U_C)$ is the joint probability distribution over the exogenous variables.

Figure 1a can now be given a formal reading as the DAG $\mathcal{G}_{\mathcal{M}}$ underlying the model we have just defined. Notice that, according to the conventions in the field, the figure shows only the endogenous variables. □

SCMs allow for the rigorous definition of interventions:

**Definition 2 (Intervention)** Given a SCM $\mathcal{M}$, a set of variable $X \in \mathcal{X}$ together with an associated set of values $x$, such that for each $X_i \in X$ there is a $x_i \in \mathcal{M}[X_i]$, an intervention $\iota : \text{do}(X = x)$ is an operator on a SCM that replaces the structural functions $f_i$ with the constants $x_i$.

Graphically, an intervention $\iota$ mutilates the original DAG $\mathcal{G}_{\mathcal{M}}$ by removing all incoming edges in $X_i$ and replacing $f_i$ with $x_i$. Thus, the intervention $\iota$ on the SCM $\mathcal{M}$ induces a new post-interventional model $\mathcal{M}_\iota$.

In a SCM, the probability distributions over the exogenous variables can be pushforwarded over the endogenous variables, thus defining joint probabilities $P_\mathcal{M}(X)$ over $X \subseteq \mathcal{X}$. Furthermore, the finite dimensionality of the outcome sets of the variables in $\mathcal{M}$ allows us to represent a SCM as a tuple $\langle \mathcal{M}[X], \mathcal{M}[\phi_X] \rangle$, where $\mathcal{M}[X]$ is the set of binary variables in $\mathcal{M}_{\iota}$ and $\mathcal{M}[\phi_X]$ is the set of mechanisms encoding the conditional distribution of an outcome as a stochastic matrix $\mathcal{M}[\phi_X]$. We can now institute an abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$ by defining:

• $R = \{S, C\}$, evaluating only nodes $S$ and $C$ in $\mathcal{M}$ as relevant to our abstraction;
• $a : R \to X'$ mapping $S \mapsto S'$, $C \mapsto C'$ specifying how variables in the two levels are related;
• $\alpha_{S'} : \mathcal{M}[S] \to \mathcal{M}'[S']$ and $\alpha_{C'} : \mathcal{M}[C] \to \mathcal{M}'[C']$.

A SCMs. An abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$ is a tuple $(R, a, \alpha_i)$ where:

• $R \subseteq X$ defines a subset of relevant variables in $\mathcal{M}$;
• $a : R \to X'$ is a surjective function mapping relevant variables $R$ in $\mathcal{M}$ to variables in $\mathcal{M}'$;
• $\alpha_i : \mathcal{M}[a^{-1}(X_i)] \to \mathcal{M}'[X_i]$ is a collection of surjective functions, one for each variable in $\mathcal{M}'$, mapping the outcomes of variable(s) $a^{-1}(X_i)$ onto the outcomes of variable $X_i$.

An abstraction $\alpha$ defines an (asymmetric) relation from a base or low-level model model $\mathcal{M}$ to an abstracted or high-level model model $\mathcal{M}'$.

**Example.** Let us consider again our toy model $\mathcal{M}$ along with a simplified SCM $\mathcal{M}'$ defined over two binary sets $\mathcal{M}'[S'], \mathcal{M}'[C']$ and two stochastic matrices $\mathcal{M}'[\phi_{S'}], \mathcal{M}'[\phi_{C'}]$. We can now institute an abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$ by defining:

• $R = \{S, C\}$, evaluating only nodes $S$ and $C$ in $\mathcal{M}$ as relevant to our abstraction;
• $a : R \to X'$ mapping $S \mapsto S'$, $C \mapsto C'$ specifying how variables in the two levels are related;
• $\alpha_{S'} : \mathcal{M}[S] \to \mathcal{M}'[S']$ and $\alpha_{C'} : \mathcal{M}[C] \to \mathcal{M}'[C']$.

Figure 1c is an illustration of the abstraction we have just defined. The formal definition of $\mathcal{M}'$ is available in Appendix D.2, while the definition of $\alpha$ is in Appendix D.3 □

The definition of abstraction is paired with a requirement of interventional consistency.

**Definition 4 (Zero-error abstraction [Rischel, 2020])** An abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$ is a zero-error abstraction if, for all disjoint sets $X', Y'$ in $\mathcal{X}'$, the following diagram commute:

\[
\begin{array}{c}
\mathcal{M}_i[a^{-1}(X')] \xrightarrow{\mathcal{M}[\phi_{a^{-1}(Y')}] \times a^{-1}(Y')] \\
\alpha_X \downarrow \alpha_{Y'} \\
\mathcal{M}'[a^{-1}(X')] \xrightarrow{\mathcal{M}'[\phi_{Y'}]} \mathcal{M}'[Y']
\end{array}
\]

that is, $\alpha_{Y'} \circ \mathcal{M}[\phi_{a^{-1}(Y')}] = \mathcal{M}'[\phi_{Y'}] \circ \alpha_X$, for all possible interventions $\iota$ on $a^{-1}(X')$, where $\mathcal{M}[\phi_{a^{-1}(Y')}]$ and $\mathcal{M}'[\phi_{Y'}]$ are the stochastic matrices derived from the SCMs encoding the relevant distribution.
The interpretation of commutativity is straightforward: an abstraction $\alpha$ is a zero-error abstraction if, for any intervention $\iota$ on $a^{-1}(X')$, we can obtain the same result in two ways: (i) by abstracting to the high-level post-interventional model and then computing the distribution of interest via a high-level mechanism; or, (ii) by computing a distribution via a low-level mechanism first, and then abstracting to high-level.

**Example.** Let us consider the abstraction $\alpha$ between $\mathcal{M}$ and $\mathcal{M}'$ defined above. Let us consider the following two disjoint subsets in $X': X' = \{S'\}$ and $Y' = \{C'\}$. This implies we will be considering interventions $\iota$ of the form $d_{\alpha}(S = s)$. To evaluate commutativity we then consider the following diagram:

$$
\begin{array}{c}
\mathcal{M}_i[S] \\
\downarrow \begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix} \\
\mathcal{M}_i[C]
\end{array}
\quad
\begin{array}{c}
\mathcal{M}'_i[S'] \\
\downarrow \begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix} \\
\mathcal{M}'_i[C']
\end{array}
$$

It is immediate to see that the diagram commute. A detailed explanation of the diagram is in Appendix D.4 \(\square\)

**Non-commutativity.** In case an abstraction diagram were not to commute, we could quantify the discrepancy between the upper and the lower path using Jensen-Shannon distance (JSD) [Cover, 1999] as:

$$E_\alpha(X', Y') = \max_\iota D_{\text{JSD}}(\alpha_{Y'} \circ M(\phi_{a^{-1}(X')},)\ |
\mathcal{M}'(\phi_{Y'} \circ \alpha_{X'})), (1)$$

over all interventions $\iota$ on $a^{-1}(X')$. A definition of JSD is given in Appendix B.

The choice of using JSD was proposed in Rischel [2020], and justified on the ground that, when composing abstractions, JSD guarantees that the overall error is bounded by the sum of the component errors [Rischel, 2020; Rischel and Weichwald, 2021]. From this measure of error on a single diagram, it is possible to define an overall abstraction error as follows.

**Definition 5 (Abstraction Error [Rischel, 2020])** Let $\alpha$ be an abstraction from a model $\mathcal{M}$ to a model $\mathcal{M}'$. Then the abstraction error is:

$$e(\alpha) = \sup_{X', Y' \subseteq X'} E_\alpha(X', Y')$$

for all disjoint non-empty and non-independent subsets $X', Y' \subseteq X'$.

**Example.** Let us consider the same base model $\mathcal{M}$ and suppose we are given an alternative abstracted model $\mathcal{M}'$, identical to $\mathcal{M}$ except for the mechanism $\mathcal{M}''[\phi_C]$ which is now encoded by the matrix $\begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$. Let us relate $\mathcal{M}$ and $\mathcal{M}''$ via the previous abstraction $\alpha$. We then obtain $E_\alpha(S'', C'') \approx 0.077$. Moreover, since the $S''$, $C''$ are the only two disjoint subsets in $\mathcal{M}'$, we also get that the overall abstraction error $e(\alpha) \approx 0.077$. The formal definition of $\mathcal{M}''$ is available in Appendix D.5, the computation of the abstraction error in Appendix D.6 \(\square\)

### 3 LEARNING ABSTRACTIONS

The definition of abstraction error provides us with a rigorous way to analyze the quality of an abstraction. We could then consider expressing the problem of learning new abstractions (or improving on existing ones) by defining the optimization problem:

$$\min_\alpha e(\alpha) \quad (3)$$

over the space of abstractions $\alpha = \langle R, a, \alpha \rangle$, and, implicitly, over the space of SCMs $\mathcal{M}'$ implied by such an abstraction.

**Hierarchy of problem.** If we make the optimization variables in Equation 3 explicit, we obtain:

$$\begin{align*}
\min_\alpha & \quad e(\alpha) \\
\text{s.t.} & \quad |X'| \in \mathbb{N} \\
& \quad |\mathcal{M}'[X']| \in \mathbb{N} \\
& \quad \mathcal{M}'[\phi_{X'}] \in \mathbb{S}(|\mathcal{M}'[X']|, |\mathcal{M}'[PA(X')]|) \\
& \quad R \subseteq X' \\
& \quad a \in \mathbb{S}_{(0,1)}(|X'|, |X|) \\
& \quad \alpha_{X'} \in \mathbb{S}_{(0,1)}(|\mathcal{M}'[X']|, |\mathcal{M}[a^{-1}(X')]|)
\end{align*} \quad (4)$$

where $\mathbb{S}((a, b))$ is the space of column-stochastic matrices with dimension $a \times b$, $\mathbb{S}_{(0,1)}((a, b))$ is the space of binary column-stochastic matrices with dimension $a \times b$, $1^T_k$ is a column vector of length $k$ of ones.

Notice how the first three lines of optimization variables account for the learning of model $\mathcal{M}'$, while the last three lines account for the learning of the abstraction $\alpha$; moreover, the first constraint enforces acyclicity, while the last two constraints enforce surjectivity. We can identify different classes of problems according to the variables that are given, as summarized in Table 1 in Appendix C.
The problem in Equation 4 is defined over integer domains, and constitutes a combinatorial optimization problem. We will leave the discussion of its complexity and of efficient algorithms to future work; in our motivating example, given its limited size, we are able to find solutions by enumeration.

**Loss function.** The objective in the optimization problem of Equation 4 might be insufficient for learning a meaningful abstraction. In a problem where we can learn the abstraction $\alpha$ and the abstracted model $M'$ (like in Figure 2b), a trivial optimal solution would be to learn an abstraction $\alpha$ that maps everything to a singleton SCM (as in Figure 2c). By mapping all the variables in $M$ to a single variable, and mapping all possible outcomes of the variables $M[X_i]$ to a single value, commutativity is trivially preserved and $e(\alpha) = 0$. This is due to the fact that a zero abstraction error only guarantees that by commuting abstraction and mechanisms we will obtain the same result, but it does not take into account the amount of information that is given up in an abstraction.

We then suggest rewriting the objective function as:

$$
\min_{\alpha} e(\alpha) + \lambda i(\alpha)
$$

where $i(\alpha)$ is a measure of information loss due to the abstraction, and $\lambda \in \mathbb{R}$ is a trade-off parameter.

**Measure of information loss.** Different measure of information loss may be considered; customized measures may weight the information loss proportionally to the importance of different subsystems, emphasizing the contribution of specific (observational or interventional) conditional distributions.

Here we propose a simple generic measure based on the discrepancy between the observational joint distribution $P_M(X)$ of the low-level model $M$ and the observational joint distribution $P_{M'}(X)$ that we would reconstruct inverting the abstraction $\alpha$. We define the inverse of a function $\alpha_{X_i}$ as:

$$
\alpha_{X_i}^* = \ell_{1, col}(\alpha_{X_i}^T)
$$

where $\ell_{1, col}$ is an $\ell_1$-normalization along the columns, and $\cdot^T$ is the transpose operator. Although for binary column-stochastic matrices this inverse is just the conventional Moore-Penrose pseudoinverse, the formulation in Equation 6 highlights the rationale behind this choice. By using a transpose, we require to map back a high-level outcome to a low-level outcome; however, multiple low-level outcomes may be mapped to a single high-level outcome; by using a $\ell_1$-normalization we evenly spread the probability among all possible low-level outcomes, in accordance with Laplace’s principle of insufficient reason [Jaynes, 1957].

If $R = X_M$, we can then define a global inverse as:

$$
\alpha^* = \bigotimes_{X_i' \in X_M} \alpha_{X_i'}^*,
$$

where $\bigotimes$ is the Kronecker product. If $R \subseteq X_M$, we need to account for non-relevant variables. Let $\bar{R}$ be the set of non-relevant variables, and let $r$ be the cardinality $|M[\bar{R}]|$; then we can compute the global inverse as:

$$
\alpha^* = \frac{1^T_r}{r} \bigotimes_{X_i' \in X_M'} (\bigotimes_{X_i' \in X_M'} \alpha_{X_i'})
$$

where $1^T_r$ is a column vector of length $r$ of ones.

**Example.** Let us consider our motivating example for the abstraction $\alpha$ from $M$ to $M'$, and compute the global inverse $\alpha^*$:

$$
\alpha^* = \begin{bmatrix} .5 \\ .5 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
$$

The explicit computation of $\alpha^*$ is given in Appendix D.7.

Finally, in analogy with abstraction error, we can define our information loss measure as:

$$
i(\alpha) = D_{JSD}(P_M(X), \alpha^*(P_{M'}(X))).
$$

**Example.** The information loss for our motivating example is:

$$
i(\alpha) \approx 0.44.
$$

Exact computations are reported in Appendix D.8.

Information loss provides a different criterion for evaluating abstraction, and it allows us to quantify two ways in which a base model and an abstracted model may diverge.

**Information loss as discrepancy between distributions.** Interven tional consistency is concerned with interventional quantities and mechanisms; it does not take into account marginal distributions on root variables in the DAG of the models, or conditional distributions that do not correspond to any mechanism. Information loss, on the other hand, is measured with respect to the joint distribution of the models and it is sensitive to the values of all the distributions.

Disregarding the value of marginal distributions makes an interventionally-consistent abstraction more robust with respect to shifts in the underlying population which are encoded in probabilities over the root nodes; in an interventional settings this a desirable properties. However, if we were to work in an observational setting, and we were interested in making predictions, especially in the anti-causal direction, a proper reconstruction of the populations may be in order.
Example. The abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$ turned out to have zero abstraction error $\epsilon(\alpha) = 0$, but quite a high information loss $i(\alpha) \approx 0.44$. This is not surprising if we look at the difference between the marginal distribution over the variable $S$: the two models were likely inferred over populations with almost diametrically opposed smoking patterns. Exact values for the marginals are given in Appendix D.9. This would of course impact observational inferences that will be no information loss in coarsening these variables to-

Example. If in our optimization we can learn a different abstracted model, it may come to no surprise that an alternative abstraction $\beta$ that swaps the outcomes of the variables in $\mathcal{M}$ and $\mathcal{M}'$ could achieve a lower information loss of $i(\beta) \approx 0.31$, although at the cost of not being interventionaly consistent anymore, $\epsilon(\beta) = 0.22$. Complete definition of $\beta$ is given in Appendix D.11.

Example. If we keep our models $\mathcal{M}$ and $\mathcal{M}'$ fixed, it may come to no surprise that an alternative abstraction $\beta$ that swaps the outcomes of the variables in $\mathcal{M}$ and $\mathcal{M}'$ could achieve a lower information loss of $i(\beta) \approx 0.31$, although at the cost of not being interventionally consistent anymore, $\epsilon(\beta) = 0.22$. Complete definition of $\beta$ is given in Appendix D.11.

Information loss as quantification of uncertainty due to reduction in resolution. Furthermore, information loss may act as a proxy to quantify how much detail is lost through abstraction. Reducing the number of variables in an abstracted model, or restricting the range of outcomes of the same variables, implies more uncertainty in the reconstruction of the joint distribution over the base model via the inverse $\alpha^*$. Notice, however, that information loss is a function of the reconstructed probability, not of the number of variables or their cardinality; if the base joint distribution over a set of variables is already maximally uncertain, there will be no information loss in coarsening these variables together. Abstracting a model to a singleton like in Figure 2 would then be sensible when the uncertainty of the base model is so high that we would not lose much by working on the abstracted singleton model and then reconstructing the original distribution.

Example. Let us consider the original base low-level model $\mathcal{M}$, and let us instantiate a singleton model $\mathcal{M}^*$, together with the trivial abstraction $\gamma$ from $\mathcal{M}$ to $\mathcal{M}^*$. This abstraction has zero abstraction error $\epsilon(\gamma) = 0$, and an information loss of $i(\gamma) \approx 0.37$. Notice that the information loss $i(\gamma)$ is less than the information loss for the abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$, despite $\mathcal{M}'$ being defined on more variables; this is due to the fact that the joint distribution reconstructed via the inverse $\alpha^*$ is further from the original joint distribution than the maximally uncertain distribution reconstructed via the inverse $\gamma^*$. However, information loss $i(\gamma)$ is higher than the information loss for the abstraction $\alpha$ from $\mathcal{M}$ to $\mathcal{M}'$; in this case, the mapping to $\mathcal{M}^*$ successfully exploits the higher number of variables and their cardinality to retain statistical information from the base model. The exact definition of the singleton model $\mathcal{M}^*$ is available in Appendix D.14, the abstraction $\gamma$ in Appendix D.15.

4 DISCUSSION

In this paper, we considered the problem of learning abstractions between SCMs: we introduced a taxonomy of optimization problems, we highlighted the limitation of focusing only on consistency, we proposed a tentative definition of an information loss quantity, and we illustrated the relevance of such a measure. Future work will take into account evaluating the complexity of the identified problem, justifying a proper information loss measure, evaluating its properties and trade-offs, and proposing heuristics for the learning problem.

References


**A SCM ASSUMPTIONS**

In our definition of SCM we make the following assumptions:

1. (Finite Variables) We explicitly assumed that a SCM is defined on a finite number of $N$ endogenous variables.

2. (Unique Exogenous Variable (UEV)) With no loss of generality [Beckers and Halpern, 2019], we assumed a single exogenous variable associated with each endogenous variable.

3. (Non-independent Exogenous Variables) We do not assume exogenous variables to be independent. This allows the exogenous variables to define a latent structure.

4. (Modularity) The mechanisms encoded by the structural functions are independent of each other. This assumption is necessary to specify perfect interventions.

5. (Measurability) The structural functions encoding the mechanisms are measurable functions. This assumption is necessary to guarantee that we can pushforward the probability distributions over the exogenous variables onto the endogenous variables.

6. (Acyclicity) Every SCM $\mathcal{M}$ admits an underlying graph $G_{\mathcal{M}} = \langle V, E \rangle$, where $V = \mathcal{X} \cup \mathcal{U}$ is the set of vertices given by endogenous and exogenous variables, and $E$ is the set of edges determined by the structural functions in $\mathcal{F}$; precisely, for each structural function $f_i$, and for each variable $Y \in \text{PA}(X_i) \cup \{U_i\}$, we will introduce and edge from $Y$ to $X_i$. Notice that, under this construction, the set $\text{PA}(X_i)$ can be given the graph-theoretic reading of parents of $X_i$. We assume that the graph $G_{\mathcal{M}}$ is acyclic.

7. (Finite Domains) Following [Rischel 2020] we will assume that the domain of each endogenous variable $\mathcal{M}[X_i] = \{x_1, x_2, \ldots, x_m\}$ is finite. This assumption is necessary to admit a representation of a SCM in terms of sets and stochastic matrices.

Notice that Assumption (3) and (6) imply that our SCM is semi-Markovian. For further discussion of these properties see, for instance, [Pearl 2009].

**B DEFINITION OF JENSEN-SHANNON DISTANCE**

Here we provide a summary definition of Kullback–Leibler divergence and Jensen–Shannon distance between two discrete probability distributions. For a more generic treatment of probability distances and their properties, we refer the reader to [Cover 1999].

**Definition 6** (Kullback–Leibler (KL) divergence) Let $p$ and $q$ be two probability mass functions defined on the same domain $\mathcal{X}$, such that $p(x) > 0$ for all $x \in \mathcal{X}$. The Kullback–Leibler (KL) divergence from $p$ to $q$ is defined as:

$$d_{KL}(p; q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)}.$$

**Definition 7** (Jensen–Shannon (JSD) distance) Let $p$ and $q$ be two probability mass functions defined on the same domain $\mathcal{X}$, such that $p(x) > 0$ and $q(x) > 0$ for all $x \in \mathcal{X}$. The Jensen–Shannon distance between $p$ and $q$ is defined as:

$$D_{JSD}(p, q) = \frac{1}{2} d_{KL}(p; m) + \frac{1}{2} d_{KL}(q; m),$$

where $m = \frac{1}{2} p + \frac{1}{2} q$.

**C HIERARCHY OF LEARNING PROBLEMS**

Table 1 reports a listing of relevant abstraction learning problems.

**D MOTIVATING EXAMPLE**

Here is a full specification of the motivating example we used throughout the paper. Code for these models is available at [https://github.com/FM2ennaro/CategoricalCausalAb](https://github.com/FM2ennaro/CategoricalCausalAb).

**D.1 LOW-LEVEL MODEL $\mathcal{M}$**

Let our low-level model $\mathcal{M}$ be defined by
<table>
<thead>
<tr>
<th>(M)</th>
<th>(M')</th>
<th>Abstraction</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>X'</td>
<td>)</td>
<td>(</td>
</tr>
<tr>
<td>given</td>
<td>given</td>
<td>given</td>
<td>given</td>
</tr>
<tr>
<td>given</td>
<td>given</td>
<td>given</td>
<td>given</td>
</tr>
<tr>
<td>given</td>
<td>given</td>
<td>given</td>
<td>given</td>
</tr>
<tr>
<td>given</td>
<td>given</td>
<td>given</td>
<td>-</td>
</tr>
<tr>
<td>given</td>
<td>given</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>given</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>given</td>
<td>given</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: Hierarchy of abstraction learning problems.
• $M[X]$ containing the following sets:
  - $\{\ast\}$
  - $M[E] = \{0, 1\}$
  - $M[S] = \{0, 1\}$
  - $M[C] = \{0, 1\}$
• $M[\phi_X]$ given by the following column-stochastic matrices:
  - $M[\phi_E]$; stochastic map from $\{\ast\}$ to $M[E]$ encoded by matrix $\begin{bmatrix} 0.8 & 0.2 \end{bmatrix}$ representing $PM(E)$.
  - $M[\phi_S]$; stochastic map from $M[E]$ to $M[S]$ encoded by matrix $\begin{bmatrix} 0.8 & 0.6 \\ 0.2 & 0.4 \end{bmatrix}$ representing $PM(S|E)$.
  - $M[\phi_C]$; stochastic map from $M[E] \times M[S]$ to $M[C]$ encoded by matrix $\begin{bmatrix} 0.9 & 0.8 & 0.4 & 0.3 \\ 0.2 & 0.4 & 0.6 & 0.7 \end{bmatrix}$ representing $PM(C|E,S)$.

D.2 HIGH-LEVEL MODEL $M'$

Let the high-level model $M'$ be defined by:

• $M'[X']$ containing the following sets:
  - $\{\ast\}$
  - $M'[S'] = \{0, 1\}$
  - $M'[C'] = \{0, 1\}$
• $M'[\phi_{X'}]$ given by the following column-stochastic matrices:
  - $M'[\phi_{S'}]$; stochastic map from $\{\ast\}$ to $M'[S']$ encoded by matrix $\begin{bmatrix} 0.2 & 0.8 \end{bmatrix}$ representing $PM'(S')$.
  - $M'[\phi_{C'}]$; stochastic map from $M'[S']$ to $M'[C']$ encoded by matrix $\begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix}$ representing $PM'(C'|S')$.

D.3 ABSTRACTION $\alpha$

Let the abstraction $\alpha$ from $M$ to $M'$ be defined by

• $R = \{S, C\}$;
• $\alpha: R \rightarrow X'$ mapping $S \rightarrow S'$, $C \rightarrow C'$;
• $\alpha$ given by the collection of maps:
  - $\alpha_{S'}: M[S] \rightarrow M'[S']$ encoded by matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
  - $\alpha_{C'}: M[C] \rightarrow M'[C']$ encoded by matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

D.4 COMMUTING DIAGRAM FOR ABSTRACTION $\alpha$ WHEN CONSIDERING SETS $S'$ AND $C'$

Let us evaluate the commutativity of abstraction $\alpha$ from $M$ to $M'$ when considering the disjoint sets $S'$ and $C'$. We consider the following diagram:

\[
\begin{array}{ccc}
M_i[S] & \xrightarrow{\begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix}} & M_i[C] \\
\downarrow \ & \ & \downarrow \\
M_i'[S'] & \xrightarrow{\begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix}} & M_i'[C']
\end{array}
\]

where:

• the left vertical arrow $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ encodes the abstraction $\alpha_{S'}$ mapping deterministically values of $S$ to values of $S'$;
• the upper horizontal arrow $\begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix}$ encodes a (virtual) mechanism $M[\phi_{C}]$ from $S$ to $C$; notice that this mechanism is computed as $PM(C|do(S))$, as such, it is different from the given mechanism $M[\phi_{C}]$ which instead encodes $PM(C|E,S)$;
• the right vertical arrow $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ encodes the abstraction $\alpha_{C'}$ mapping deterministically values of $C$ to values of $C'$;
• the lower horizontal arrow $\begin{bmatrix} 0.88 & 0.38 \\ 0.12 & 0.62 \end{bmatrix}$ encodes a (virtual) mechanism $M'[\phi_{C'}]$ from $S'$ to $C'$; this mechanism is computed as $PM'(C'|do(S'))$ and, in this case, it is the same as the given mechanism $M'[\phi_{C'}]$ which encodes $PM'(C'|S')$.

D.5 HIGH-LEVEL MODEL $M''$

Let an alternative high-level model $M''$ be defined by:

• $M''[X'']$ containing the following sets:
  - $\{\ast\}$
  - $M''[S''] = \{0, 1\}$
  - $M''[C''] = \{0, 1\}$
• $M''[\phi_{X''}]$ given by the following column-stochastic matrices:
  - $M''[\phi_{S''}]$; stochastic map from $\{\ast\}$ to $M''[S'']$ encoded by matrix $\begin{bmatrix} 0.2 & 0.8 \end{bmatrix}$ representing $PM'(S'')$.
\(- M''[\varnothing_C]:\) stochastic map from \(M''[S'']\) to \(M''[C'']\) encoded by matrix \[
\begin{bmatrix}
0.8 & 0.3 \\
0.2 & 0.7
\end{bmatrix}
\]
representing \(P_{M''}(C''|S'')\).

### D.6 ABSTRACTION ERROR \(e(\alpha)\) FOR \(\alpha\) FROM \(M\) TO \(M''\)

Let us consider the two disjoint subsets in \(\lambda''\): \(\{S''\}\) and \(\{C''\}\). To evaluate the abstraction error \(E_\alpha(S'', C'')\) of the abstraction \(\alpha\) from \(M\) to \(M''\) we consider the following diagram:

\[
\begin{array}{c}
M_i[S] \\
\downarrow \begin{bmatrix} 0.88 & 0.38 \\
0.12 & 0.62
\end{bmatrix} \\
\downarrow \begin{bmatrix} 1 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix} \\
M_i[C]
\end{array}
\]

\[
\begin{array}{c}
M''[S''] \\
\downarrow \begin{bmatrix} 0.8 & 0.3 \\
0.2 & 0.7
\end{bmatrix} \\
\downarrow \begin{bmatrix} 1 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix} \\
M''[C'']
\end{array}
\]

and we evaluate:

\[
E_\alpha(S'', C'') = \max \{D_{\text{JSD}}([0.88, 0.12], [0.8, 0.2]), D_{\text{JSD}}([0.38, 0.62], [0.3, 0.7])\}
\]

\[
\approx 0.077.
\]

### D.7 INVERSE \(\alpha^*\)

Let us compute the global inverse \(\alpha^*\):

\[
\alpha^* = \frac{1}{2} \begin{bmatrix} 1 & 1 \\
1 & 1 \end{bmatrix} \otimes \alpha_{S'}^* \otimes \alpha_{C'}^*.
\]

The inverses \(\alpha_{S'}^*\) and \(\alpha_{C'}^*\) are trivially identities. The global inverse \(\alpha^*\) is then given by:

\[
\alpha^* = \begin{bmatrix}
.5 & .5 \\
.5 & .5
\end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\
0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\
0 & 1 \end{bmatrix}
\]

\[
\begin{bmatrix}
.5 & .5 \\
.5 & .5
\end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\
0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\
0 & 1 \end{bmatrix}
\]

\[
= \begin{bmatrix}
.5 & .5 \\
.5 & .5
\end{bmatrix}
\]

where, for readability, we omitted writing zeros in the last matrix. Notice how the matrix \(\alpha^*\) expresses our uncertainty in reconstructing \(M\) from \(M'\): for instance, the first column of the matrix \(\alpha^*\) encodes the fact that the joint values \((S' = 0, C' = 0)\) could be evenly mapped to the joint values \((E = 0, S = 0, C = 0)\) or \((E = 1, S = 0, C = 0)\).

### D.8 INFORMATION LOSS \(i(\alpha)\) FOR \(\alpha\) FROM \(M\) TO \(M'\)

In order to compute the information loss, we need to evaluate first the joint distribution on the base model:

\[
P_M(E, S, C) = \begin{bmatrix}
0.576 \\
0.064 \\
0.064 \\
0.096 \\
0.096 \\
0.024 \\
0.024 \\
0.056
\end{bmatrix},
\]

then the joint distribution on the abstracted model:

\[
P_{M'}(S', C') = \begin{bmatrix}
0.176 \\
0.024 \\
0.304 \\
0.496
\end{bmatrix},
\]

and last reconstruct the distribution over \(E, S, C\) via \(\alpha^*\):

\[
\alpha^*(P_{M'})(E, S, C) = \begin{bmatrix}
0.088 \\
0.12 \\
0.152 \\
0.248 \\
0.088 \\
0.012 \\
0.152 \\
0.248
\end{bmatrix}.
\]

Finally, we can compute the actual information loss as:

\[
D_{\text{JSD}}(P_M(E, S, C), \alpha^*(P_{M'})(E, S, C)) \approx 0.44.
\]

### D.9 MARGINAL \(P(S)\) IN \(M\) AND \(M'\)

Let us evaluate the marginal distribution for the smoking variable \((S)\) in the two models \(M\) and \(M'\). In the base model we have:

\[
P_M(S) = \sum_{E, C} P_M(E, S, C) = \begin{bmatrix}
0.76 \\
0.24
\end{bmatrix}.
\]

In the abstracted model we are given:

\[
P_{M'}(S') = \begin{bmatrix}
0.2 \\
0.8
\end{bmatrix}.
\]

The marginal distribution reconstructed via \(\alpha^*\) is trivially:

\[
\alpha^*(P_M)(S) = \sum_{E, C} \alpha^*(P_{M'})(E, S, C) = \begin{bmatrix}
0.2 \\
0.8
\end{bmatrix}.
\]
D.10 CONDITIONAL $P(S|C)$ IN $\mathcal{M}$ AND $\mathcal{M}'$

Let us evaluate the conditional distribution of the smoking variable ($S$) given the cancer variable in the two models $\mathcal{M}$ and $\mathcal{M}'$. In the base model we have:

$$P_{\mathcal{M}}(S|C) = \frac{P_{\mathcal{M}}(S,C)}{P_{\mathcal{M}}(C)} = \begin{bmatrix} 0.88 & 0.37 \\ 0.12 & 0.63 \end{bmatrix}.$$  

In the base model, no-cancer is highly correlated with not-smoking, and having cancer is correlated with smoking. If we were to make this inference in the base model $\mathcal{M}$ and then abstract the outcome via $\alpha_S$, we would map the outcome $S = 0$ to $S' = 0$, and $S = 1$ to $S' = 1$.

However, if we were to abstract the condition via $\alpha_C$, we would first map the condition $C = 0$ to $C' = 0$, and $C = 1$ to $C' = 1$; then if we were to compute the conditional in the abstracted model we would get:

$$P_{\mathcal{M}'}(S'|C') = \frac{P_{\mathcal{M}'}(S',C')}{P_{\mathcal{M}'}(C')} = \begin{bmatrix} 0.37 & 0.63 \\ 0.05 & 0.95 \end{bmatrix}.$$  

Thus, in this case we would infer smoking with high probability for any value of the cancer variable.

D.11 ABSTRACTION $\beta$

Let the abstraction $\beta$ from $\mathcal{M}$ to $\mathcal{M}'$ be defined by

- $R = \{S, C\}$;

- $b : R \rightarrow X'$ mapping $S \mapsto S'$, $C \mapsto C'$;

- $\beta$ given by the collection of maps:
  
  - $\beta'_S : \mathcal{M}[S] \rightarrow \mathcal{M}'[S']$ encoded by matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$;
  
  - $\beta'_C : \mathcal{M}[C] \rightarrow \mathcal{M}'[C']$ encoded by matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

D.12 INVERSE $\beta^*$

The inverses $\beta'_S^*$ and $\beta'_C^*$, remain exchange matrices. The global inverse $\beta^*$ is then given by:

$$\beta^* = \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

where, for readability, we omitted writing zeros in the last matrix.

D.13 HIGH-LEVEL MODEL $\mathcal{M}'''$

Let us consider a third high-level model $\mathcal{M}'''$ be defined by:

- $\mathcal{M}'''[X'''$ containing the following sets:
  
  - $\{\ast\}$
  
  - $\mathcal{M}'''[S'''] = \{0, 1\}$
  
  - $\mathcal{M}'''[C'''] = \{0, 1\}$

- $\mathcal{M}'''[\phi_{X'''}$ given by the following column-stochastic matrices:
  
  - $\mathcal{M}'''[\phi_{S'''}$ : stochastic map from $\{\ast\}$ to $\mathcal{M}'''[S''']$ encoded by matrix $\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$ representing $P_{\mathcal{M}'''}(S''')$
  
  - $\mathcal{M}'''[\phi_{C'''}$ : stochastic map from $\mathcal{M}'''[S''']$ to $\mathcal{M}'''[C''']$ encoded by matrix $\begin{bmatrix} 0.88 & 0.12 \\ 0.38 & 0.62 \end{bmatrix}$ representing $P_{\mathcal{M}'''}(C'''|S''')$

D.14 HIGH-LEVEL MODEL $\mathcal{M}^*$

Let the singleton high-level model $\mathcal{M}^*$ be defined by

- $\mathcal{M}^*[X]$ containing the following set:
  
  - $\{\ast\}$

- $\mathcal{M}^*[\phi_*$ given by the following column-stochastic matrix:
  
  - $\mathcal{M}^*[\phi_*$ : stochastic map from $\{\ast\}$ to $\{\ast\}$ encoded by matrix $\begin{bmatrix} 1 \\ \end{bmatrix}$

D.15 ABSTRACTION $\gamma$

Let the abstraction $\gamma$ from $\mathcal{M}$ to $\mathcal{M}^*$ be defined by:
• $R = \{E, S, C\};
• $c : R \rightarrow \mathcal{X}^*$ mapping $E \mapsto \{\ast\}, S \mapsto \{\ast\}, C \mapsto \{\ast\};$
• $\gamma$ given by the map:
  - $\gamma_* : \mathcal{M}[S] \times \mathcal{M}[E] \times \mathcal{M}[C] \rightarrow \{\ast\}$ encoded by matrix $[1, 1, 1, 1, 1, 1, 1, 1].$

D.16 INVERSE $\gamma^*$

The global inverse $\gamma^*$ is trivially:

$$
\gamma^* = \begin{bmatrix}
.125 \\
.125 \\
.125 \\
.125 \\
.125 \\
.125 \\
.125 \\
.125
\end{bmatrix}.
$$