
Semi-Supervised Learning, Causality and the Conditional Cluster Assumption

Julius von Kügelgen^{1,2} Alexander Mey³ Marco Loog^{3,4} Bernhard Schölkopf²

¹University of Cambridge, ²Max Planck Institute for Intelligent Systems,
³Delft University of Technology, ⁴University of Copenhagen

Abstract

While the success of semi-supervised learning (SSL) is still not fully understood, Schölkopf et al. (2012) have established a link to the principle of independent causal mechanisms. They conclude that SSL should be impossible when predicting a target variable from its causes, but possible when predicting it from its effects. Since both these cases are somewhat restrictive, we extend their work by considering classification using cause and effect features at the same time, such as predicting a disease from both risk factors and symptoms. While standard SSL exploits information contained in the marginal distribution of the inputs (to improve our estimate of the conditional distribution of target given inputs), we argue that in our more general setting we can use information in the conditional of effect features given causal features. We explore how this insight generalizes the previous understanding, and how it relates to and can be exploited for SSL.

1 Introduction

Due to the scarcity and often high acquisition cost of labelled data, machine learning methods that make effective use of large quantities of unlabelled data are crucial. One such method is semi-supervised learning (SSL) [1, 2] where in addition to labelled data, possibly large numbers of unlabelled observations are available to the learner at training time.

While positive results have been obtained on a range of problems, a shortcoming of SSL is that it can actually degrade performance if certain assumptions are not met [2, Chapter 4]. For example, Ben-David, Lu, and Pál [3, Section 3] show that the cluster assumption, commonly used in SSL settings, can lead to degraded performance even in simple cases, e.g., for binary classification with data generated from two unimodal Gaussians. Such examples make it clear that many aspects of SSL are, as of yet, not well understood.

Building on the *principle of independent causal mechanisms* (ICM) [4, 5, 6], Schölkopf et al. [7] have pointed out a link between the possibility of SSL and the causal structure underlying a given learning problem. Specifically, they argue that SSL should be impossible when predicting a target variable from its causes (referred to as *causal learning*), but possible when predicting it from its effects (referred to as *anticausal learning*)—see Sec. 2 for details. Empirical evidence from a meta-analysis of various SSL scenarios supports this claim [7].

In this work, we extend the investigation of connections between SSL and causality to a more general setting. Rather than treating causal and anticausal learning in isolation, we consider predicting a target variable from both causes and effects *at the same time*. As an example, consider the setting of predicting disease from medical data where both types of features are commonly found. A patient’s age, sex, medical family history, genetic information, diet, and other risk factors such as smoking all constitute (possible) causal features, whereas examples of effect features include the clinical symptoms exhibited by the patient, as well results of medical tests such as imaging results, serum tests, or tissue samples.

As our main result we show that for this general setting of learning with both cause and effect features, *the relevant information that additional unlabelled data may provide for prediction is contained in the conditional distribution of effect features given causal features* (Sec. 3). We then argue how this new insight may be used to reformulate classical SSL assumptions (Sec. 4), and propose algorithms based on these assumptions (Sec. 5). Results from evaluating our methods against well-established SSL algorithms on synthetic and medical datasets (Sec. 6) provide empirical support for our analysis.

2 Background and related work

We start by reviewing previous work and key concepts upon which our work builds. Throughout, we use X to denote a random variable taking values in \mathcal{X} , which is assumed to be a subset of \mathbb{R}^d . P denotes a probability measure and $P(X)$ the probability distribution of X with density p . We write $x \in \mathbb{R}$ for a scalar, $\mathbf{x} \in \mathbb{R}^d$ for a vector, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ for a matrix or collection of samples.

2.1 Semi-supervised learning

SSL describes a learning setting where, in addition to a labelled sample $(\mathbf{X}^l, \mathbf{y}^l) = \{(\mathbf{x}^i, y^i)\}_{i=1}^{n_l}$, we have access to an unlabelled sample $\mathbf{X}^u = \{\mathbf{x}^i\}_{i=n_l+1}^{n_l+n_u}$ at training time. It is usually assumed that $n_l \ll n_u$. At test time, the task is to predict targets Y from inputs X . If predictions are made on the unlabelled training data only we speak of *transductive* learning [8].

The aim and hope of SSL is that additional unlabelled data helps in making better predictions. \mathbf{X}^u can improve the estimate of $P(X)$, but SSL aims at improving $P(Y|X)$. This can only work if there is a link between $P(X)$ and $P(Y|X)$. Indeed, many approaches to SSL establish such a link through additional assumptions [1, 2, 7]. Two common ones are the *cluster assumption*, positing that points in the same cluster of $P(X)$ have the same label Y ; and *low-density separation*, stating that class boundaries of $P(Y|X)$ should lie in an area where $P(X)$ is small. For original references, as well as for discussion how these assumption relate to various SSL methods, refer to [2].

We briefly mention four of the more common methods, starting with self-learning (sometimes also called the Yarowsky-algorithm). This is a wrapper algorithm that initializes the learner based on the labeled data, updates the labels for the unlabelled data, and then retrains based on all labeled data available, possibly iterating this procedure [9, 10, 11]. Secondly, generative model approaches maximise the likelihood of a generative model

$$\hat{\theta}^{\text{MLE}} = \arg \max_{\theta} p(\mathbf{X}^l, \mathbf{y}^l, \mathbf{X}^u | \theta) = \sum_{\mathbf{y}^u \in \mathcal{Y}^{n_u}} p(\mathbf{X}^l, \mathbf{y}^l, \mathbf{X}^u, \mathbf{y}^u | \theta). \quad (1)$$

While this is a hard optimization problem due to the latent variables \mathbf{y}^u , a local optimum can be found via the expectation maximisation algorithm (EM) [12]. The third class of common methods we mention are the graph-based approaches. These construct a similarity-based graph representation of the data and propagate labels to neighbours in this graph [13, 14, 15]. Finally, transductive SVMs assign labels which maximise a (soft) margin over labelled and unlabelled data while minimizing a regularized risk on the labelled data [8, 16].

2.2 Causality

Despite data showing a positive correlation between chocolate consumption and the number of Nobel prizes per capita [17], we would not expect that force-feeding the population chocolate would result in higher research output. The correlation in this example may make chocolate consumption a useful predictor in an i.i.d. setting, but it does not allow one to answer interventional questions of the form “what would happen if we actively changed some of the variables?”.

This notion of intervention is at the heart of the difference between correlation and causation. While much of machine learning is concerned with using correlations between variables to make predictions, Reichenbach [18] has argued that such correlations always result from underlying causal relationships: statistical dependence is an epiphenomenon—a by-product of a causal process.¹

¹For the given example, a possible causal explanation for the observed correlation would be that a healthy economy acts as a common cause for both chocolate consumption and a good education system.

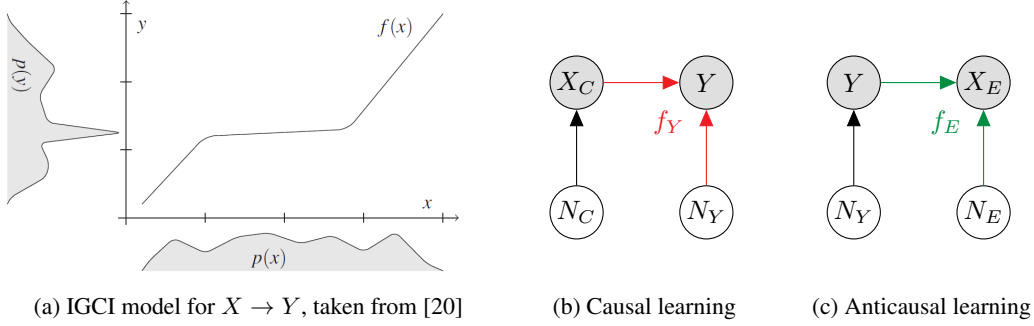


Figure 1: (a) Illustration of the idea of ICM: if the distribution of the cause $p(x)$ is chosen independently of the mechanism $f : \mathcal{X} \rightarrow \mathcal{Y}$, then this independence is violated in the backward direction as $p(y)$ contains information about f^{-1} . In a causal learning setting (b) SSL is thus impossible as $P(X_C)$ contains no information about $P(Y|X_C)$, whereas in an anticausal learning setting (c) $P(X_E)$ may contain information about $P(Y|X_E)$ and SSL is, in principle, possible.

Structural causal model (SCM) To reason about causality in SSL, we adopt the SCM framework [19] which defines a causal model over variables X_1, \dots, X_d as: (i) a collection of structural assignments $\{X_i := f_i(\text{PA}_i, N_i)\}_{i=1}^d$, where f_i are deterministic functions computing variable i from its causal parents $\text{PA}_i \subseteq \{X_1, \dots, X_d\} \setminus X_i$; and (ii) a factorizing joint distribution over the unobserved noise variables $\{N_1, \dots, N_d\}$. Together, (i) and (ii) define a causal generative process and imply an observational joint distribution over X_1, \dots, X_d which factorises over the causal graph² \mathcal{G} as:

$$P(X_1, \dots, X_d) = \prod_{i=1}^d P(X_i | \text{PA}_i), \quad (2)$$

Principle of independent causal mechanisms (ICM) Motivated by viewing the f_i in the definition of SCMs as independent physical mechanisms of nature, ICM states that “the causal generative process (...) is composed of independent and autonomous modules that do not inform or influence each other” [6]. In other words, the conditional distributions of each variable given its parents, $P(X_i | \text{PA}_i)$, in eq. (2) are independent modules which do not share information. Note that this notion of independence is different from statistical independence (indeed, variables can still be statistically dependent), but it can be formalized as an *algorithmic* independence of distributions. Intuitively, two distributions are considered algorithmically independent if encoding them jointly does not admit a shorter description than describing each of them separately. In this case we say that they do not share information. This notion has been formalized in terms of Kolmogorov complexity (or Algorithmic Information) $K(\cdot)$ by Janzing and Schölkopf [21] who show that when X is a cause of Y , $X \rightarrow Y$,

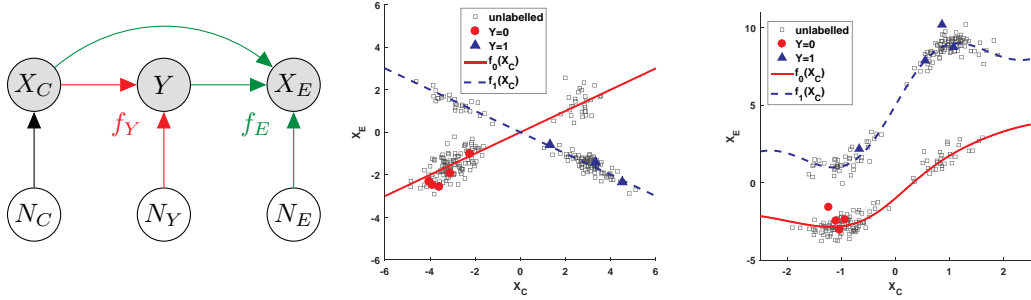
$$K(P(X)) + K(P(Y|X)) \stackrel{+}{\leq} K(P(Y)) + K(P(X|Y)).$$

Here, the notation $\stackrel{+}{\leq}$ refers to a constant due to the choice of a Turing machine in the definition of algorithmic information. For the bivariate setting, ICM reduces to an *independence of cause and mechanism* [4, 22]. This is illustrated in Fig. 1a using the information geometric causal inference (IGCI) model [20] for $X \rightarrow Y$ in which a deterministic invertible function f generates effect Y from cause X . If the input distribution of the cause, $p(x)$, is chosen independently from the mechanism $f(x)$ (or more generally, $p(y|x)$), then this independence is violated in the backward (non-causal) direction, since $p(y)$ has large density where f has small slope and thereby contains information about $f^{-1}(y) = p(x|y)$.

2.3 Causal and anticausal learning

For the task of predicting a target Y from features X , Schölkopf et al. [7] distinguish between *causal learning* where X is a cause of Y (see Fig. 1b), and *anticausal learning* where X is an effect of Y (see Fig. 1c). In a causal learning setting, it then follows from the independence of cause and mechanism

²The causal graph \mathcal{G} is obtained by drawing a directed edge from each node in PA_i (i.e. the direct causes of X_i) to X_i for all i . We assume throughout that \mathcal{G} is acyclic.



(a) Learning with cause (X_C) and effect (X_E) features (b) Example of linear mechanisms f_0, f_1 between X_C and X_E (c) Example of non-linear mechanisms f_0, f_1 between X_C and X_E

Figure 2: (a) Causal setting assumed in this paper, and two datasets which could arise from it (b),(c). As shall be explained below, the conditional cluster assumption links class labels to membership in clusters of the conditional $p(X_E|X_C)$, suggesting to classify unlabelled points according to whether they are better explained by the red or the blue function. Best viewed in color.

that $P(X_C)$ and $P(Y|X_C)$ are algorithmically independent. Recalling the goal of SSL—improving $P(Y|X)$ from $P(X)$ —SSL should thus be impossible. In the anticausal direction, on the other hand, this independence relation is between $P(Y)$ and $P(X_E|Y)$ whereas $P(X_E)$ may (and in some cases provably will [4]) share information with $P(Y|X_E)$ and SSL is thus, in principle, possible.

3 Semi-supervised learning with cause and effect features

In this work, we assume that we are given a small labelled sample $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l) = \{(\mathbf{x}_c^i, y^i, \mathbf{x}_e^i)\}_{i=1}^{n_l}$ and a large unlabelled sample $(\mathbf{X}_C^u, \mathbf{X}_E^u) = \{(\mathbf{x}_c^i, \mathbf{x}_e^i)\}_{i=n_l+1}^{n_l+n_u}$ generated from the following SCM:

$$X_C := N_C \quad (3)$$

$$Y := f_Y(X_C, N_Y) \quad (4)$$

$$X_E := f_E(X_C, Y, N_E). \quad (5)$$

This causal model is shown in Figure 2a. We will refer to X_C as causal features and X_E as effect features and assume this partitioning to be known (e.g., think of the medical example with risk factors and diagnostic tests). Analogous to eq. (2), the SCM eqs. (3)-(5) induces an observational distribution which factorises into *independent causal mechanisms* as

$$P(X_C, Y, X_E) = P(X_C)P(Y|X_C)P(X_E|Y, X_C). \quad (6)$$

Note that this setting generalises the cases of only causes or effects considered by Schölkopf et al. [7] without positing any new statistical independences. It thus remains widely applicable.³

Our goal is to predict the target Y from features (X_C, X_E) , so we are interested in estimating

$$P(Y|X_C, X_E) = \frac{P(X_C, Y, X_E)}{P(X_C, X_E)} = \frac{P(Y|X_C)P(X_E|Y, X_C)}{\sum_{y \in \mathcal{Y}} P(Y = y|X_C)P(X_E|Y = y, X_C)} \quad (7)$$

while having additional information about $P(X_C, X_E)$ from unlabelled data.

In analogy to the case of causal learning (see Sec. 2.3 and [7]), by the principle of independent causal mechanisms, the distribution over causes $P(X_C)$ does not contain any information about $P(Y|X_C)$ or $P(X_E|Y, X_C)$ (see RHS of eq. (6)) and thereby also not about $P(Y|X_C, X_E)$ (see RHS of eq. (7)). Indeed, $P(Y|X_C, X_E)$ is completely determined by the structural equations for Y as a function of X_C , eq. (4), and X_E as a function of X_C and Y , eq. (5), and does not depend on what distribution of causal features X_C , eq. (3), is fed into this generative process.

³Note that, for example, omitting the link $X_C \rightarrow X_E$ renders the two feature sets conditionally independent given Y [23], which can be a restrictive assumption for realistic scenarios and can already be well addressed by approaches like co-training [10].

Having established that $P(X_C)$ does not contain useful information for our task,⁴ we are left with $P(Y, X_E|X_C)$ which according to the chain rule of probability admits two possible factorisations,

$$P(Y, X_E|X_C) = P(Y|X_C)P(X_E|X_C, Y) \quad (8)$$

$$P(Y, X_E|X_C) = P(X_E|X_C)P(Y|X_C, X_E). \quad (9)$$

Eq. (8) is a causal factorization into independent mechanisms which do not share any information. Eq. (9), however, corresponds to a *non-causal* factorization, implying that the factors on the RHS may share information. Since we care about estimating $P(Y|X_C, X_E)$ and we have additional information about $P(X_E|X_C)$ from unlabelled data, it is precisely this potential dependence or link between $P(Y|X_C, X_E)$ and $P(X_E|X_C)$ that SSL approaches should aim to exploit in our setting (Figure 2a). We formulate this result as follows.

Main insight. When learning with both causes and effects of a target as captured by the causal model in eqs. (3)-(5), $P(X_E|X_C)$ contains all relevant information provided by additional unlabelled data $(\mathbf{x}_C, \mathbf{x}_E)$ about $P(Y|X_C, X_E)$. Therefore, SSL approaches for such a setting should aim at exploiting this information and linking these two distributions via suitable assumptions.

We note that this contains previous results for causal and anticausal learning as special cases: in absence of causal features (i.e., for anticausal learning) $P(X_E|X_C)$ reduces to the known setting of $P(X_E)$ containing information about $P(Y|X_E)$, whereas in absence of effect features, $P(X_E|X_C)$ becomes meaningless and SSL thereby impossible, both consistent with the findings of [7].

However, our result goes further than this since having additional unlabelled data of both cause and effect features can be strictly more informative than having only unlabelled effects. To illustrate this point, consider the example where X_E is a possibly noisy copy, or proxy label, of Y . In this case, unlabelled data contains information which is very similar to the information contained in the labelled data so that learning to predict X_E from X_C can be very helpful in solving the problem.

4 New assumptions for semi-supervised classification

While the previous analysis (Sec. 3) applies to general prediction tasks including regression, we now focus our attention on classification. For conceptual simplicity and ease of illustration, we will assume binary classification in what follows, but extensions to the multi-class setting are straightforward.

First, we note that for a binary label $Y \in \{0, 1\}$ we can rewrite eqs. (4) and (5) as follows:

$$Y := \mathbb{I}\{g(X_C) > U\} \quad (10)$$

$$X_E := \begin{cases} f_0(X_C, N_E) & \text{if } Y = 0, \\ f_1(X_C, N_E) & \text{if } Y = 1, \end{cases} \quad (11)$$

where \mathbb{I} is the indicator function and U is a uniform random variable on $[0, 1]$, so that $g(X_C)$ computes $P(Y = 1|X_C)$. Allowing arbitrary g, f_0, f_1 and N_E , this comes without loss of generality.

Next, we use our insight to reformulate standard SSL assumptions (see Sec. 2.1) for the setting of Fig. 2a where both X_C and X_E are observed. Our aim is to adapt these assumptions such that they make use of potential information shared between $P(X_E|X_C)$ and $P(Y|X_C, X_E)$.

4.1 Conditional cluster assumption

While the standard cluster assumption advocates for sharing labels within clusters in the marginal distribution of all features, in view of the above we postulate that *points in the same cluster of $P(X_E|X_C)$ share the same label Y* . We refer to this as the *conditional cluster assumption*.

Here, one can think of clusters of $P(X_E|X_C)$ as clusters in the space \mathcal{F} of functions $f : \mathcal{X}_C \rightarrow \mathcal{X}_E$ computing effects from causes. Different functions in this space can arise from different choices of Y and N_E in the structural equation for X_E , eq. (11). The conditional cluster assumption can then be understood as saying that the two class-dependent mechanisms f_0 and f_1 form clusters in \mathcal{F} .⁵

⁴Since we generally aim to minimise an *expected* loss, $P(X_C)$ can still be helpful in getting a better estimate of the expectation operator [6]. By useful information here we mean information about $P(Y|X_C, X_E)$.

⁵Note that due to the general form of eq. (11), it is possible to have more than one cluster per class in \mathcal{F} . For handwritten digits, for example, N_E could act as a switch between 7s with and without the horizontal stroke.

This idea is illustrated for two cases of linear and non-linear functions f_0, f_1 with additive and unimodal noise in Figs. 2b and 2c, respectively (best viewed in colour). These are simple examples where the asymmetry introduced by knowing the causal partitioning of features can help identify the true mechanisms (shown in solid and dashed lines) and therefore the correct labelling. Standard SSL approaches agnostic to the causally-induced asymmetry between features, on the other hand, can easily fail in these situations. For the data shown in Fig. 2b, for example, large-margin methods or graph-based approaches (s. Sec. 2.1) operating in the joint feature space (X_C, X_E) will learn to classify Y by the sign of X_C (i.e., the boundary $X_C = 0$), leading to an error rate of almost 50%.

4.2 Low conditional-density separation

In a similar vein, we adapt the low-density separation assumption to our setting. While in its original form, low-density separation is a statement about the joint density of all features, we have argued that (subject to the ICM principle) $P(X_C)$ contains no information about $P(Y|X_C, X_E)$, but that the conditional density $P(X_E|X_C)$ may do so. We therefore propose that a more justified notion of separation is that *class boundaries of $P(Y|X_C, X_E)$ should lie in regions where $P(X_E|X_C)$ is small*. We refer to this as *low conditional-density separation*.

5 Algorithms

While the main contribution of the present paper is conceptual, it is illustrative to discuss the implication of our assumptions from Sec. 4 for some of the standard approaches to SSL introduced in Sec. 2.1 and propose variations thereof which explicitly aim to make use of the information shared between $P(X_E|X_C)$ and $P(Y|X_C, X_E)$.

5.1 Semi-generative models

While a naive generative model would unnecessarily model the full distribution—including the uninformative part $P(X_C)$ —this approach to SSL is easily adapted to our new assumptions by only modelling the informative part of the generative process, $P(Y, X_E|X_C)$. This type of *semi-generative model* has been introduced by von Kügelgen, Mey, and Loog [23] in the context of domain adaptation and under the additional assumption of conditionally independent feature sets.

Given a model parameterised by θ , a maximum likelihood approach similar to eq. (1) then yields

$$\hat{\theta}^{\text{MLE}} = \arg \max_{\theta} p(\mathbf{y}^l, \mathbf{X}_E^l | \mathbf{X}_C^l; \theta) \sum_{\mathbf{y}^u \in \mathcal{Y}^{n_u}} p(\mathbf{y}^u, \mathbf{X}_E^u | \mathbf{X}_C^u; \theta). \quad (12)$$

Equivalently, we minimise the negative log-likelihood (NLL) which *for fixed labels* decomposes according to eq. (6) into separate terms which can be optimised independently for θ_Y and θ_E :

$$\text{NLL}(\theta | \mathbf{X}_C, \mathbf{y}, \mathbf{X}_E) := -\log p(\mathbf{y}, \mathbf{X}_E | \mathbf{X}_C; \theta) = -\log p(\mathbf{y} | \mathbf{X}_C; \theta_Y) - \log p(\mathbf{X}_E | \mathbf{X}_C, \mathbf{y}; \theta_E)$$

This separation leads us to an EM approach [12] to find a local optimum of eq. (12) by iteratively computing the expected label given the current parameters (E-step) and then minimising the NLL w.r.t. to the parameters keeping the labels fixed (M-step). This is summarized in Algorithm 1. For the specific case of logistic regression for $P(Y|X_C)$ and a class-dependent linear Gaussian model for $P(X_E|X_C, Y)$ we provide a more detailed procedure for both soft and hard labels in Algorithm 3.

5.2 Conditional self-learning

The second algorithm we propose is loosely related to the ideas of label propagation and self-learning [9]. However, instead of propagating labels based on similarities between points computed in the joint feature space (X_C, X_E) as in the conventional approach, we argue that $P(X_C)$ contains no information about $P(Y|X_C, X_E)$ and that we should instead focus on the information contained in $P(X_E|X_C)$. To achieve this, we assume an *additive noise model* [24] for f_0, f_1 in eq. (11), i.e.,

$$f_i(X_C, N_E) = f_i(X_C) + N_{E,i} \quad \text{for } i = 0, 1. \quad (13)$$

Note, however, that unlike in the probabilistic approach of Sec. 5.1 we do not make additional assumptions about the exact noise distribution, such as Gaussianity. We do however assume that the noise has zero mean and is unimodal, so that there is one function from X_C to X_E for each label.

Algorithm 1: EM-like algorithm for fitting a semi-generative model by maximum likelihood

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; models $p(y|\mathbf{x}_C; \boldsymbol{\theta}_Y)$ and $p(\mathbf{x}_E|\mathbf{x}_C, y; \boldsymbol{\theta}_E)$

Output: fitted labels \mathbf{y}^u ; estimates $\hat{\boldsymbol{\theta}}_Y, \hat{\boldsymbol{\theta}}_E$

```

1  $t \leftarrow 0$ 
2  $\hat{\boldsymbol{\theta}}_Y^{(0)} \leftarrow \arg \min \text{NLL}(\boldsymbol{\theta}_Y | \mathbf{y}^l)$ 
3  $\hat{\boldsymbol{\theta}}_E^{(0)} \leftarrow \arg \min \text{NLL}(\boldsymbol{\theta}_E | \mathbf{y}^l)$ 
4 while not converged do
5    $\mathbf{y}^{(t)} \leftarrow \mathbb{I}\{p(\mathbf{y}|\mathbf{X}_C^u, \mathbf{X}_E^u; \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}) > 0.5\}$ 
6    $\hat{\boldsymbol{\theta}}_Y^{(t+1)} \leftarrow \arg \min \text{NLL}(\boldsymbol{\theta}_Y | \mathbf{y}^l, \mathbf{y}^{(t)})$ 
7    $\hat{\boldsymbol{\theta}}_E^{(t+1)} \leftarrow \arg \min \text{NLL}(\boldsymbol{\theta}_E | \mathbf{y}^l, \mathbf{y}^{(t)})$ 
8    $t \leftarrow t + 1$ 
9 end
10 return  $\mathbf{y}^{(t-1)}, \hat{\boldsymbol{\theta}}_Y^{(t)}, \hat{\boldsymbol{\theta}}_E^{(t)}$ 

```

Algorithm 2: Conditional self-learning

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data; $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; `Regress()` method

Output: fitted labels \mathbf{y}^u ; functions \hat{f}_0, \hat{f}_1

```

1  $t \leftarrow 0$ 
2 while unlabelled data left do
3   for  $i = 0, 1$  do
4      $\hat{f}_i^{(t)} \leftarrow \text{Regress}(\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l)$ 
5      $\mathbf{r}_i \leftarrow \|\mathbf{X}_E^u - \hat{f}_i^{(t)}(\mathbf{X}_C^u)\|^2$ 
6   end
7    $(i, j) \leftarrow \arg \min \{\mathbf{r}_i^j : j = 1, \dots, n_u\}_{i=0}^1$ 
8    $y^{n_i+j} \leftarrow i$ 
9    $\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l \leftarrow \text{append}(\mathbf{x}_E^{n_i+j}, \mathbf{x}_C^{n_i+j})$ 
10   $t \leftarrow t + 1$ 
11 end
12 return  $\mathbf{y}^u, \hat{f}_0^{(t-1)}, \hat{f}_1^{(t-1)}$ 

```

Our approach then aims at learning these functions and can be summarised as follows. We first initialise two functions \hat{f}_0 and \hat{f}_1 from the labelled sample by regressing X_E on X_C . We then proceed to iteratively compute the predictions of the \hat{f}_i on unlabelled data, label the point with smallest prediction error as the respective class, and then use this point to update the corresponding \hat{f}_i until all points are labelled. We refer to this approach as *conditional self-learning*, see Algorithm 2.

Soft labels and connection to probabilistic approach It is also possible to use the above approach with soft labels (as often done in conventional label propagation [14, 15]) by using a weighted regression scheme. This would require a method of computing regression weights from prediction errors of \hat{f}_0 and \hat{f}_1 , though, and therefore needs additional assumptions or heuristics. We note that choosing a particular noise distribution for N_E and using $P(Y|X_C)$ as a class prior leads to a soft-label EM approach, see Algorithm 3. We therefore presently restrict ourselves to hard labels.

While it is conceptually based on the ICM assumption and an analysis of the causal structure among the feature set, the conditional self-learning approach is linked to a number of known methods, including not only self-learning, but also methods building on *competition of experts*, as recently applied to the problem of learning causal mechanisms. In this work [25], the functions f_0, f_1 are generative models competing for data that has undergone unknown transformations, and eventually each specializing on how to invert one of those transformations.

6 Experiments

To corroborate our analysis with empirical evidence, we evaluate our algorithms from Sec. 5 on synthetic data as well as on two medical datasets from the UCI repository. We compare with T-SVMs [8, 16] with linear and RBF kernels using the q3svm implementation [26], and with label propagation [14, 15] using the implementation in sci-kit learn [27]. We use the default parameters in all cases. For our conditional self-learning algorithm we use linear ridge regression [27] with default regularization strength 1 and for the EM algorithms we use a logistic regression model for $Y|X_C$ and linear, class-depedent Gaussians for $X_E|X_C, Y$, see Algorithm 3 and Appendix B.2 for details.

Synthetic data As controlled environments, we generate three different types of datasets S1, S2, and S3 with cause and effect features: S1 represents linearly-separable data; S2 corresponds to a non-linear decision boundary similar to Fig.2c; and S3 is a version of S2 with multi-dimensional features. Details of how exactly synthetic data was generated are provided in Appendix B.2

Medical data As real world data, due to the fact that both plausibly contain cause and effect features, we chose the two medical datasets *Pima Indians Diabetes* [28] and *Heart Disease* [29].

Table 1: Average accuracies on unlabelled data (higher is better) \pm one standard deviation across 100 runs, each time randomly drawing 10 (for S1, S2, S3) or 20 (for Diabetes, Heart) new labelled and 200 new unlabelled samples. Results refer to transductive evaluation for ease of comparison with other methods. Best method for each dataset highlighted in bold. Last four rows are our causally-motivated methods. The “-” indicates that label propagation did not converge on S3.

Method	S1 (lin)	S2 (non-lin)	S3 (mult.dim)	Diabetes	Heart Disease
Lin. log. reg.	.968 \pm .023	.823 \pm .080	.945 \pm .039	.626 \pm .058	.526 \pm .066
Lin. T-SVM	.865 \pm .093	.878 \pm .074	.822 \pm .117	.602 \pm .065	.746 \pm .060
RBF T-SVM	.863 \pm .094	.876 \pm .075	.821 \pm .116	.601 \pm .064	.745 \pm .060
RBF label propag.	.924 \pm .082	.909 \pm .065	-	.650 \pm .030	.528 \pm .068
Semi-gen. (sup.)	.968 \pm .076	.935 \pm .074	.949 \pm .082	.669 \pm .064	.550 \pm .096
Semi-gen.+soft EM	.986 \pm .081	.989 \pm .024	.991 \pm .067	.661 \pm .063	.518 \pm .050
Semi-gen.+hard EM	.985 \pm .079	.972 \pm .058	.987 \pm .076	.695 \pm .064	.518 \pm .050
Cond. self-learning	.980 \pm .052	.923 \pm .090	.961 \pm .069	.659 \pm .079	.719 \pm .076

We select those features which are most strongly correlated with the target variable ($p < 0.01$) and categorise them into cause and effect features to the best of our knowledge (see Appendix B.1).

Results The results of our experiments are summarised in Table 1, see the table caption for details. On the synthetic datasets, our causally-motivated methods outperform the purely supervised logistic regression baseline as well as the other SSL approaches, which in the case of S1 and S3 actually lead to a decrease in performance. The probabilistic approaches perform particularly well on the synthetic datasets, which is expected since the generative model for these cases was specified by us and thus known (see Appendix B.2). However, our conditional self-learning algorithm is rather competitive as shown by the results on the real data. Notably, it is the only method which improves upon the baseline (i.e., achieves SSL) for each of the five datasets considered. Moreover, it also leads to significant improvements on the Heart dataset where—likely due to a violation of the underlying assumptions on the generative model—the EM approaches fail.

7 Discussion

The present paper looks at SSL from the point of view of causal modelling. Its main contribution is conceptual rather than algorithmic. We argue that if we know how the feature set in the input X can be partitioned into cause and effect features (X_C and X_E), then this has surprising theoretical implications for how SSL should utilize unlabelled data: rather than simply exploiting links between $P(X)$ and $P(Y|X)$, as formalized for instance in the standard SSL cluster assumption, one should exploit links between $P(X_E|X_C)$ and $P(Y|X_C, X_E)$. Note that we view this not as a contradiction to the usual cluster assumption, but rather as an explication or refinement thereof, taking into account the causal structure; indeed, it subsumes SSL in the anticausal setting [7] as a special case. It does not subsume SSL in the causal setting, but as argued by [7], SSL fails in this case.

We do not mean to claim that all assumptions underlying these insights always apply in practice. We may not know the causal structure of the features, and in practice, some of the features may be neither causes nor effects, but correlated to the target by unobserved confounders (a case which would be interesting to study in future work). Moreover, the principle of independent causal mechanisms (ICM) underlying both our analysis and the one of [7] may not strictly hold for a task at hand.

While the present analysis is intriguing and points out a previously unexplored link between two conditional distributions, the jury is still out on how to best exploit unlabelled data in machine learning. The present insight is but one step, and in particular, while encouraging, the algorithms and experiments based upon it can only be a starting point. We hope that they will lead to new approaches that make explicit use of causal structure and exploit the conditional cluster assumption in more elegant and effective ways. Ultimately, the value of novel assumptions and conceptual models lies in whether they provide a fertile basis to inspire further algorithm development and theoretical understanding, and we expect that the present ideas and analysis will constitute such a contribution.

References

- [1] Xiaojin Zhu. *Semi-Supervised Learning Literature Survey*. Tech. rep. 1530. Computer Sciences, University of Wisconsin-Madison, 2005.
- [2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. 1st. The MIT Press, 2010.
- [3] Shai Ben-David, Tyler Lu, and Dávid Pál. “Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning”. In: *Proceedings of the The 21st Annual Conference on Learning Theory (COLT-08)*. Helsinki, Finland, 2008.
- [4] Povilas Daniušis et al. “Inferring deterministic causal relations”. In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2010, pp. 143–150.
- [5] D. Janzing and B. Schölkopf. “Causal inference using the algorithmic Markov condition”. In: *IEEE Transactions on Information Theory* 56.10 (2010), pp. 5168–5194.
- [6] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press, 2017.
- [7] Bernhard Schölkopf et al. “On causal and anticausal learning”. In: *29th International Conference on Machine Learning (ICML 2012)*. International Machine Learning Society. 2012, pp. 1–8.
- [8] V Vapnik. *Statistical learning theory*. Wiley, New York, 1998.
- [9] H. J. Scudder. “Probability of Error of Some Adaptive Pattern-Recognition Machines”. In: *IEEE Transactions on Information Theory* 11 (1965), pp. 363–371.
- [10] Avrim Blum and Tom Mitchell. “Combining labeled and unlabeled data with co-training, semi-supervised”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, pp. 92–100.
- [11] Steven Abney. “Understanding the Yarowsky Algorithm”. In: *Comput. Linguist.* 30.3 (Sept. 2004), pp. 365–395. ISSN: 0891-2017. DOI: 10 . 1162 / 0891201041850876. URL: <http://dx.doi.org/10.1162/0891201041850876>.
- [12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.
- [13] Xiaojin Zhu and Zoubin Ghahramani. *Learning from labeled and unlabeled data with label propagation*. Tech. rep. Carnegie Mellon University, 2002.
- [14] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. “Semi-supervised learning using gaussian fields and harmonic functions”. In: *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 2003, pp. 912–919.
- [15] Dengyong Zhou et al. “Learning with local and global consistency”. In: *Advances in neural information processing systems*. 2004, pp. 321–328.
- [16] Thorsten Joachims. “Transductive inference for text classification using support vector machines”. In: *International Conference on Machine Learning*. Vol. 99. 1999, pp. 200–209.
- [17] Franz H Messerli. “Chocolate Consumption, Cognitive Function, and Nobel Laureates”. In: *The New England Journal of Medicine* 367.16 (2012), pp. 1562–1564.
- [18] Hans Reichenbach. *The direction of time*. Vol. 65. Univ of California Press, 1956.
- [19] Judea Pearl. *Causality*. Cambridge university press, 2000.
- [20] Dominik Janzing et al. “Information-geometric approach to inferring causal directions”. In: *Artificial Intelligence* 182 (2012), pp. 1–31.
- [21] Dominik Janzing and Bernhard Schölkopf. “Causal inference using the algorithmic Markov condition”. In: *IEEE Transactions on Information Theory* 56.10 (2010), pp. 5168–5194.
- [22] Jan Lemeire and Erik Dirkx. *Causal models as minimal descriptions of multivariate systems*. 2006.
- [23] Julius von Kügelgen, Alexander Mey, and Marco Loog. “Semi-Generative Modelling: Covariate-Shift Adaptation with Cause and Effect Features”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 1361–1369.
- [24] P. O. Hoyer et al. “Nonlinear causal discovery with additive noise models”. In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. Vancouver, BC, Canada: Curran, 2009, pp. 689–696. URL: <http://nips.cc/Conferences/2008/>.

- [25] Giambattista Parascandolo et al. “Learning Independent Causal Mechanisms”. In: *International Conference on Machine Learning*. 2018, pp. 4033–4041.
- [26] Fabian Gieseke et al. “Fast and simple gradient-based optimization for semi-supervised support vector machines”. In: *Neurocomputing* 123 (2014), pp. 23–32.
- [27] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [28] Jack W Smith et al. “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus”. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1988, p. 261.
- [29] Robert Detrano et al. “International application of a new probability algorithm for the diagnosis of coronary artery disease”. In: *The American journal of cardiology* 64.5 (1989), pp. 304–310.

A Algorithms

First, we make some clarifying remarks regarding Algorithms 1 and 2 presented in the main paper.

In Algorithm 1, the negative log-likelihood terms (NLL) implicitly also depend on (X_C, X_E)

$$\begin{aligned} \text{NLL}(\boldsymbol{\theta}_Y | \mathbf{y}^l, \mathbf{y}^{(t)}) &:= -\log p(\mathbf{y}^l | \mathbf{X}_C^l; \boldsymbol{\theta}_Y) - \log p(\mathbf{y}^{(t)} | \mathbf{X}_C^u; \boldsymbol{\theta}_Y) + \text{const.} \\ \text{NLL}(\boldsymbol{\theta}_E | \mathbf{y}^l, \mathbf{y}^{(t)}) &:= -\log p(\mathbf{X}_E^l | \mathbf{X}_C^l, \mathbf{y}^l; \boldsymbol{\theta}_E) - \log p(\mathbf{X}_E^u | \mathbf{X}_C^u, \mathbf{y}^{(t)}; \boldsymbol{\theta}_E) + \text{const.} \end{aligned}$$

where the additive constants do not depend on $\boldsymbol{\theta}_Y$ and $\boldsymbol{\theta}_E$, respectively, and can thus be ignored for finding a minimum. For sake of brevity and since only the values of fitted labels and parameter estimates change throughout the algorithm while the observed feature values are held fixed, we omit explicitly conditioning on (X_C, X_E) .

In Algorithm 2, the notation $\mathbf{X}_{C,i}$ and $\mathbf{X}_{E,i}$ refers to those samples of causal and effect features with label $Y = i$. Moreover, \mathbf{r}_i denote the residuals from regressing \mathbf{X}_E on \mathbf{X}_C using \hat{f}_i .

Algorithm 3 below describes concrete soft and hard labelling versions of the EM approach proposed in Algorithm 1 for the model assumption of a logistic regression for $Y|X_C$ and linear Gaussian distributions for $X_E|X_C, Y = 0$ and $X_E|X_C, Y = 1$ in some more detail. It was used for our experiments in Sec. 6.

Algorithm 3: Soft and hard label EM algorithms for a semi-generative model with logistic regression for $P(Y|X_C)$ and class-dependent linear Gaussian models for $P(X_E|Y, X_C)$.

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; regularisation strength λ for ridge regression; labelling type (hard/soft)

Output: fitted labels \mathbf{y}^u ; parameters $\boldsymbol{\theta}_Y, \boldsymbol{\theta}_E$

```

1  $t \leftarrow 0$ 
   // initialise parameter estimates using only the labelled sample
2  $\boldsymbol{\theta}_Y^{(0)} \leftarrow \text{LogisticRegression}(\mathbf{X}_C^l, \mathbf{y}^l; \boldsymbol{\theta}_Y)$ 
3  $\boldsymbol{\theta}_{E,0}^{(0)} \leftarrow ((\mathbf{X}_{C,0}^l)^T \mathbf{X}_{C,0}^l + \lambda I)^{-1} (\mathbf{X}_{C,0}^l)^T \mathbf{X}_{E,0}^l$ 
4  $\boldsymbol{\theta}_{E,1}^{(0)} \leftarrow ((\mathbf{X}_{C,1}^l)^T \mathbf{X}_{C,1}^l + \lambda I)^{-1} (\mathbf{X}_{C,1}^l)^T \mathbf{X}_{E,1}^l$ 
5 while not converged do
   | // compute soft (class probabilities) and hard labels (E-step)
6    $\mathbf{q}^{(t)} \leftarrow p(\mathbf{y} | \mathbf{X}_C^u, \mathbf{X}_E^u; \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_{E,0}^{(t)}, \boldsymbol{\theta}_{E,1}^{(t)})$ 
7    $\mathbf{y}^{(t)} \leftarrow \mathbb{I}\{\mathbf{q}^{(t)} > 0.5\}$ 
   | // compute weights (1 for hard labels, class prob. for soft labels)
8   if hard labelling then
9     |  $\mathbf{w}^{(t)} \leftarrow \mathbf{1}$ 
10  else
11    |  $\mathbf{w}^{(t)} \leftarrow [\mathbf{1}; \mathbf{y}^{(t)} \odot \mathbf{q}^{(t)} + (1 - \mathbf{y}^{(t)}) \odot (1 - \mathbf{q}^{(t)})]$ 
12  end
13   $\mathbf{W}_1^{(t)} \leftarrow \text{diag}([\mathbf{y}^l; \mathbf{q}^{(t)}])$ 
14   $\mathbf{W}_0^{(t)} \leftarrow \text{diag}([\mathbf{1} - \mathbf{y}^l; \mathbf{1} - \mathbf{q}^{(t)}])$ 
   | // Update parameter estimates keeping estimated labels fixed (M-step)
15   $\boldsymbol{\theta}_Y^{(t)} \leftarrow \text{WeightedLogisticRegression}([\mathbf{X}_C^l; \mathbf{X}_C^u], [\mathbf{y}^l; \mathbf{y}^{(t)}], \mathbf{w}^{(t)}; \boldsymbol{\theta}_Y)$ 
16   $\boldsymbol{\theta}_{E,1}^{(t)} \leftarrow (\mathbf{X}_C^T \mathbf{W}_1^{(t)} \mathbf{X}_C + \lambda I)^{-1} \mathbf{X}_C^T \mathbf{W}_1^{(t)} \mathbf{X}_E$ 
17   $\boldsymbol{\theta}_{E,0}^{(t)} \leftarrow (\mathbf{X}_C^T \mathbf{W}_0^{(t)} \mathbf{X}_C + \lambda I)^{-1} \mathbf{X}_C^T \mathbf{W}_0^{(t)} \mathbf{X}_E$ 
18   $t \leftarrow t + 1$ 
19 end
20 return  $\mathbf{y}^{(t-1)}, \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}$ 

```

B Experimental details

B.1 Real-world data

For the Pima Indians Diabetes dataset we used the partitioning $X_C = \{\text{DiabetesPedigreeFunction, Pregnancies, BMI}\}$ and $X_E = \{\text{Glucose}\}$. DiabetesPedigreeFunction is a measure of the family history of diabetes and BMI stands for body mass index.

For the (Coronary) Heart Disease dataset we used the partitioning $X_C = \{\text{sex, ca, thal}\}$ and $X_E = \{\text{chest pain}\}$. Here, ‘‘ca’’ refers to the number of major vessels (0-3) that contained calcium (colored by flouroscopy), and ‘‘thal’’ to thallium scintigraphy results, a nuclear medicine test that images the blood supply to the muscles of the heart.

For further details we refer to Smith et al. [28] and Detrano et al. [29].

B.2 Synthetic data

The synthetic datasets used in our experiments were generated as follows. First, we draw $X_C \in \mathbb{R}^{d_C}$ from a mixture of m d_C -dimensional Gaussians. Next, we draw $Y \in \{0, 1\}$ and $X_E \in \mathbb{R}^{d_E}$ according to the SCM

$$Y := \mathbb{I}[\sigma(\mathbf{a}^T X_C + b) > N_Y], \quad N_Y \sim U[0, 1]$$

$$X_E := \begin{cases} \mathbf{A}_0 X_C + \mathbf{b}_0 + \mathbf{D}_0 N_E & \text{if } Y = 0, \\ \mathbf{A}_1 X_C + \mathbf{b}_1 + \mathbf{D}_1 N_E & \text{if } Y = 1, \end{cases} \quad N_E \sim \mathcal{N}_{d_E}(0, I)$$

with $\mathbf{a} \in \mathbb{R}^{d_C}$; $b \in \mathbb{R}$; $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{R}^{d_E \times d_C}$; and $\mathbf{D}_0, \mathbf{D}_1 \in \mathbb{R}^{d_E \times d_E}$ are diagonal matrices of standard deviations. $\sigma(x) = (1 + e^{-x})^{-1}$ denotes the logistic sigmoid function.

This induces the distributions

$$Y|X_C \sim \text{Bernoulli}(\mathbf{a}^T X_C + b)$$

and

$$X_E|(X_C, Y = i) \sim \mathcal{N}(\mathbf{A}_i X_C + \mathbf{b}_i, D_i^2) \quad \text{for } i = 0, 1.$$

For experiments on synthetic data we draw a new dataset according to the above generative process in each run, keeping parameters fixed at the following values.

S1: Linear synthetic dataset

We use the following parameters to generate S1.

- feature dimensions: $d_C = d_E = 1$
- X_C : $m = 3$ components with weights $w = [0.3, 0.4, 0.3]$, means $\mu_C = [-5, 0, 5]$ and standard deviations $\sigma = [0.5, 0.5, 0.5]$
- Y : $a = 0.5, b = 0$
- X_E : $A_0 = A_1 = 1, b_0 = -b_1 = 2, D_0 = D_1 = 0.25$

S2: Nonlinear synthetic dataset

We use the following parameters to generate S2.

- feature dimensions: $d_C = d_E = 1$
- X_C : $m = 2$ components with weights $w = [0.5, 0.5]$, means $\mu_C = [-3, 3]$ and variances $\sigma^2 = [0.5, 0.5]$
- Y : $a = 0.5, b = 0$
- X_E : $A_0 = -A_1 = 0.5, b_0 = b_1 = 0, D_0 = D_1 = 0.25$

S3: Nonlinear multidimensional synthetic dataset

We use the following parameters to generate S3.

- feature dimensions: $d_C = d_E = 2$
- X_C : $m = 2$ components with weights $w = [0.5, 0.5]$, means $[-3, -3], [3, 3]$ and covariances $\Sigma = \text{diag}([0.5, 0.5])$
- Y : $a = 0.5, b = 0$
- X_E : $a_0 = -a_1 = 0.5, b_0 = b_1 = 0, D_0 = D_1 = \text{diag}(0.25, 0.25)$

C Code

We will release Python code for our algorithms, as well as scripts to reproduce our results with the camera ready version of this paper.