Causal vs. Anticausal merging of predictors

Sergio Hernan Garrido Mejia
Max Planck Institute for Intelligent Systems
Amazon
Tübingen, Germany
shgm@tuebingen.mpg.de

Patrick Blöbaum
Amazon
Tübingen, Germany

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Dominik Janzing
Amazon
Tübingen, Germany

Abstract

We study the differences arising from merging predictors in the causal and anti-causal directions using the same data. In particular we study the asymmetries that arise in a simple model where we merge the predictors using one binary variable as target and two continuous variables as predictors. We use Causal Maximum Entropy (CMAXENT) as inductive bias to merge the predictors, however, we expect similar differences to hold also when we use other merging methods that take into account asymmetries between cause and effect. We show that if we observe all bivariate distributions, the CMAXENT solution reduces to a logistic regression in the causal direction and Linear Discriminant Analysis (LDA) in the anticausal direction. Furthermore, we study how the decision boundaries of these two solutions differ whenever we observe only some of the bivariate distributions implications for Out-Of-Variable (OOV) generalisation.

1 Introduction

A common problem in machine learning and statistics consists of estimating or combining models or (expert opinions) of a target variable of interest into a single, hopefully better, model [14]. There are several reasons of why this problem is important. For example, experts might have access to different data when creating their models, but might not have access to the data available to other experts, while there might be a modeller who can access the expert's opinions and put them together into a single model. Furthermore, experts might specialise in certain areas of the support of the input space, so that a modeller with access to the expert's opinions could potentially produce a single model exploiting the strengths of each modeller. This problem is commonly known as "mixture of experts", "expert aggregation", "merging of experts" or "expert pooling" [40, 22, 6, 31, 30].

The merging of experts problem is usually ill-defined, in the sense that there are multiple joint models (that is, models that include *all* covariates) that after marginalisation would render the same prediction as the individual experts (that is, those which include only *some* of the covariates). This ill-definedness of the problem requires strong inductive biases. One way to provide this inductive bias in a principled way is through the Maximum Entropy (MAXENT) principle [20]. In brief, MAXENT suggests finding the distribution with maximum Shannon entropy subject to moment constraints. This turns out to be the same as choosing the distribution closest to the uniform distribution having the same moments as those given by the constraints. In Section 2.2 we introduce MAXENT and Causal MAXENT (CMAXENT) in more detail, the latter being an extension that allows to include causal information when available [17].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Most of the research on the merging of predictors focuses on finding a meta-predictor that uses the given models and best fits to the data [41], whereas the focus of the present article is understanding the implications of causal assumptions in the merging of experts instead of aiming for models with best performance. Furthermore, most of this research focuses on predictors that use the same predicting variables for each of the models. To the best of our knowledge, the only exception is the random subspace method, notable for being the basis of random forests [15].

What if in addition to the different models, a researcher has some causal knowledge of the underlying system? For example, they could know whether a variable or set of variables used in a particular model are causes or effects of the target variable, potentially changing the resulting predictor of interest. Causal knowledge produces asymmetries that have been exploited in the past to understand some common machine learning tasks like transfer learning, semi-supervised learning or distribution shift [35, 39, 18, 21].

In the present work, we investigate how including causal knowledge produces asymmetric results when merging predictors. In particular, we are interested in how solutions to the CMAXENT principle [17] differ when we assume different causal relations for the same data. That is, we are interested in the asymmetries produced by *causal* assumptions on CMAXENT inferences. In particular, we will study the differences in the solutions when we assume the causal data generation process (so that the covariates or predictors are causal parents of our target variable) in contrast with the anticausal generation process (so that the predictors are causal children of our target variable). We are going to study these asymmetries in the case where we do not observe all the variables jointly; one of the differentiating characteristics of this research with respect to other merging of predictors work.

Including the right causal assumptions when merging predictors is relevant, for instance, in the medical domain. Suppose we are interested in the presence or absence of a disease, and we have models from hospitals and labs relating risk factors and symptoms to the disease we are interested in. Combining the predictors would be valuable to predict the disease but it also requires to include the right causal assumptions, if the direction matters for the merging of predictors: risk factors cause diseases and diseases cause symptoms. The literature of merging of predictors has focused on important aspects of the resulting models like generalisation bounds or speed of estimation but the relation to causality has remained largely unexplored.

Previous approaches have considered the problem of merging of experts using MAXENT [25, 27, 34]. However, such research considers the problem from a purely statistical perspective and does not study the ramifications of different causal assumptions. In fact, the way they study the aggregation problem is by merging the probability of the outcome given by each expert without regarding how these probabilities were produced.

The main contributions of this article can be summarised as follows

- We study the differences in causal and anticausal merging of predictors whenever the inductive bias used to merge the predictors allows causal information to be included.
- In particular, we find that CMAXENT with a binary target and continuous covariates, reduces to logistic regression and LDA, two classic classification algorithms, when merging predictors in causal and anticausal directions, respectively.
- Furthermore, we study the implications of these asymmetries Out Of Variable (OOV) generalisation whenever we do not observe all the first and second moments as constraints in the CMAXENT problem.

The remainder of the paper is organised as follows. In Section 2 we introduce basic notation and give a brief overview of MAXENT and CMAXENT. Then, in Section 3 we present the optimisation problem in the causal and anticausal direction and give the explicit solutions of the problems, thereby connecting the solutions of CMAXENT to well-known classification algorithms. In addition, we prove that the decision boundary in the causal and anticausal directions, with full knowledge of the moments (as defined in the section itself) renders equal slopes of the predictors. In Section 4 we weaken the assumption of full knowledge of the moments and instead assume knowledge of a subset of the moments in Section 3. Partial knowledge of the moments have implications for Out Of Variable (OOV) generalisation and resulting in differences in decision boundaries. We close with Section 5 with some discussion and concluding remarks. All the proofs are left to the appendix for the sake of brevity and clarity of the main text.

2 Notation and preliminaries

2.1 Notation

Let Y be a binary random variable taking values in $\mathcal{Y} = \{-1, 1\}$, and $\mathbf{X} = \{X_1, X_2\}$ be a pair of continuous variables, so that $x_i \in \mathbb{R}$. Let $f : \mathcal{Y} \times \mathbb{R}^2 \to \mathbb{R}$ be a measurable function, P a measure on $\mathcal{Y} \times \mathbb{R}^2$, and p the density of the distribution of a random variable with respect to the Lebesgue measure in the case of real valued random variables, and with respect to the counting measure in the case of discrete random variables. To be precise, $\mathrm{p}(Y, \mathbf{X})$ is a density with respect to the product of the Lebesgue measure and the counting measure. We denote $\mathbb{E}_{\mathrm{p}}[f(Y, \mathbf{X})]$ the expectation of f with respect to p. We restrict ourselves to the scenario with two continuous variables and one binary outcome given that we can already observe asymmetries in the merging of experts, and can visualise such asymmetries without having to project such space into 2 dimensions. The results here can be easily generalised into a discrete outcome variable (and indeed we do, in Corollary 7). Throughout the article we will care about finding a predictor of Y using $\mathbf{X}$ as covariates. That is, we are interested in the density $\mathrm{p}(Y \mid \mathbf{X})$.

2.2 Maximum Entropy and Causal Maximum Entropy

The Maximum Entropy (MAXENT) principle was born in the statistical mechanics literature as a way to find a distribution consistent with a set of expectation constraints [20]. That is, given observed sample averages $\tilde{f} = \frac{1}{N}\sum_{i=1}^{N} f(y_i, x_i)$ we find the density $\mathrm{p}(Y, \mathbf{X})$ so that the expectations with respect to $\mathrm{p}(Y, \mathbf{X})$ are equal to those observed.

Notice that MAXENT does not attempt to find the 'true' distribution of the data, but instead the distribution closest to the uniform distribution so that the expectation constraints are satisfied. We will see examples of such optimisation problems in subsequent sections. Using the Lagrange multiplier formalism for constrained optimisation, one can prove that the solution to the MAXENT problem belongs to the exponential family. The MAXENT distribution and its properties have been studied widely, see Grünwald and Dawid [10] and Wainwright et al. [38] and references therein.

In **Causal MAXENT** (CMAXENT, Janzing [17]), the optimisation is performed in an assumed causal order; that is, we first find the MAXENT distribution of causes and then the Maximum Conditional Entropy of the effects given the inferred distribution of the causes. As argued in [36] this typically results in distributions that are more plausible for the respective causal direction. One can think of CMAXENT as usual MAXENT with the distribution of the cause as additional constraint, where the latter has been obtained via separate entropy maximization.

3 Known predictor covariances

We will begin by studying the solution of the CMAXENT problem when we observe all the bivariate distributions and summarise them with first and second moments. The restriction to first and second moments has several reasons: First, these simple constraints are already sufficient to explain the interesting asymmetries between causal and anticausal. Second, including higher order moments makes the problem computationally harder and increases the risk of overfitting on noisy finite sample results. Last, including more moments decreases the asymmetries between the causal directions. Mathematically, we have the following (estimated) expectations and their respective sample averages:

$$\hat{\mathbb{E}}[Y] = q, \qquad \hat{\mathbb{E}}[\mathbf{X}Y] = \boldsymbol{\phi} = \begin{bmatrix}\phi_1 \\ \phi_2\end{bmatrix}, \tag{1}$$

$$\hat{\mathbb{E}}[\mathbf{X}] = \bar{\mathbf{x}} = \begin{bmatrix}\bar{x}_1 \\ \bar{x}_2\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}, \qquad \hat{\mathbb{E}}[\mathbf{X}\mathbf{X}^\top] = \boldsymbol{\Sigma}_{\mathbf{X}} = \begin{pmatrix}\bar{s}_1^2 & \bar{s}_{1,2} \\ \bar{s}_{1,2} & \bar{s}_2^2\end{pmatrix}, \tag{2}$$

where we assumed the mean of $\mathbf{X}$ is zero.

3.1 The causal direction

Consider the causal graph in Figure 1a and the expectations given in Equations (1) and (2). As mentioned on Section 2.2, CMAXENT suggests finding the density $\mathrm{p}(\mathbf{X})$ with maximum entropy

(a) Graph in the causal direction (b) Graph in the anticausal direction

Figure 1: Causal graphs analysed throughout the article

consistent with the first and second moments of $\mathbf{X}$, and then finding the density $\mathrm{p}(Y \mid \mathbf{X})$ with maximum entropy, with the estimated $\mathrm{p}(\mathbf{X})$ and the moments that involve Y as constraints.

These steps can be summarised in the following optimisation problems. First, for $\mathrm{p}(\mathbf{X})$,

$$\begin{aligned} \max_{\mathrm{p}(\mathbf{x})} \quad & H(\mathbf{X}) = -\int_{\mathbb{R}^2} \mathrm{p}(\mathbf{x}) \log \mathrm{p}(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ \text{s.t.} \quad & \mathbb{E}[X_i] = \bar{x}_i, \text{ with } i \in \{1,2\} \\ & \mathbb{E}[X_i^2] = \bar{s}_i^2, \text{ with } i \in \{1,2\} \\ & \mathbb{E}[X_1 X_2] = \bar{s}_{1,2} \\ & \int_{\mathbb{R}^2} \mathrm{p}(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1. \end{aligned} \tag{3}$$

On the other hand, the Maximum Conditional Entropy optimisation problem is as follows

$$\begin{aligned} \max_{\mathrm{p}(y \mid \mathbf{x})} \quad & H(Y \mid \mathbf{X}) = -\int_{\mathbb{R}^2} \sum_y \mathrm{p}(y \mid \mathbf{x}) \mathrm{p}(\mathbf{x}) \log \mathrm{p}(y \mid \mathbf{x})\, \mathrm{d}\mathbf{x} \\ \text{s.t.} \quad & \mathbb{E}[Y X_i] = \phi_i, \text{ with } i \in \{1,2\} \\ & \mathbb{E}[Y] = q \\ & \sum_y \mathrm{p}(y \mid \mathbf{x}) = 1, \quad \text{for each } \mathbf{x}. \end{aligned} \tag{4}$$

Where $\mathrm{p}(\mathbf{X})$ is the one we found by solving Equation (3).

Proposition 1 (Resulting predictor in the causal direction)**.** *Using the Lagrange multiplier formalism for the optimisation problems in Equations* (3) *and* (4) *we obtain: (i) a multivariate Gaussian distribution for $P(\mathbf{X})$, and (ii) the density of Y conditioned on $\mathbf{X}$ given by*

$$\mathrm{p}_\lambda(y \mid x_1, x_2) = \exp\left(\lambda_0 y + \lambda_1 y x_1 + \lambda_2 y x_2 + \alpha(x_1, x_2)\right) \tag{5}$$

$$\alpha(x_1, x_2) = \log \sum_y \exp\left(\lambda_0 y + \lambda_1 y x_1 + \lambda_2 y x_2\right), \tag{6}$$

where $\alpha(\mathbf{x})$ is a normalising constant.

The density can be written as

$$\mathrm{p}_\lambda(y = 1 \mid x_1, x_2) = \frac{1}{2}(1 + \tanh(\lambda_0 + \lambda_1 x_1 + \lambda_2 x_2)). \tag{7}$$

The proof of this result can be found in [19, Section 3.1 and 3.2].

Remark 2 Notice that Equation (7), our predictor of interest, is just a rescaled version of a sigmoid function. That is, we can estimate $\mathrm{p}(Y \mid \mathbf{X})$ with a logistic regression. A similar observation was done in [8] in the context of using an exponential loss for boosting. This relation was further explored in [24], where a more direct relation to maximum likelihood and exponential families was established. In Section 5 we discuss how these results in the statistical literature can be interpreted as making causal assumptions about the relation between the predictor and target variables.

3.2 The anticausal direction

Now consider the graph in Figure 1b. In this scenario, covariates of our predictor of interest are the effects of our target variable. Following the CMAXENT principle, we first find the density $\mathrm{p}(Y)$ with maximum entropy and is consistent with first moment of Y and then find the density $\mathrm{p}(\mathbf{X} \mid Y)$ with maximum conditional entropy consistent with the moments that involve $\mathbf{X}$ and $\mathrm{p}(Y)$ found in the previous step. After this two-step process, we are left with the joint density $\mathrm{p}(Y, \mathbf{X})$ from which we can derive a predictor of Y, $\mathrm{p}(Y \mid \mathbf{X})$ using Bayes' Theorem (Section 3.3). The whole procedure can be summarised with the following optimisation problems. For the cause, we have

$$\begin{aligned} \max_{\mathrm{p}(y)} \quad & H(Y) = -\sum_y \mathrm{p}(y) \log \mathrm{p}(y) \\ \text{s.t.} \quad & \mathbb{E}[Y] = q \\ & \sum_y \mathrm{p}(y) = 1. \end{aligned} \tag{8}$$

And for the effects,

$$\begin{aligned} \max_{\mathrm{p}(\mathbf{x}|y)} \quad & H(\mathbf{X} \mid Y) = -\int_{\mathbb{R}^2} \sum_y \mathrm{p}(\mathbf{x} \mid y)\mathrm{p}(y) \log \mathrm{p}(\mathbf{x} \mid y) d\mathbf{x} \\ \text{s.t.} \quad & \mathbb{E}[Y X_i] = \phi_i, \text{ with } i \in \{1,2\} \\ & \mathbb{E}[X_i] = \bar{x}_i, \text{ with } i \in \{1,2\} \\ & \mathbb{E}[X_i^2] = \bar{s}_i^2, \text{ with } i \in \{1,2\} \\ & \mathbb{E}[X_1 X_2] = \bar{s}_{1,2} \\ & \int_{\mathbb{R}^2} \mathrm{p}(\mathbf{x} \mid y) d\mathbf{x} = 1, \quad \text{for each } y. \end{aligned} \tag{9}$$

Proposition 3 (Resulting predictor in the anticausal direction)**.** *Using the Lagrange multiplier formalism for the optimisation problems in Equations* (8) *and* (9)*, we obtain a Bernoulli distribution for Y with $\mathrm{p}(y = 1) = q$, and $\mathrm{p}_\lambda(\mathbf{x} \mid y)$ given by*

$$\begin{aligned} \mathrm{p}_\lambda(\mathbf{x} \mid y) = \exp[&\lambda_1 y x_1 + \lambda_2 y x_2 + \lambda_3 x_1 + \lambda_4 x_2 \\ &+ \lambda_5 x_1^2 + \lambda_6 x_2^2 + \lambda_7 x_1 x_2 + \beta(y)] \end{aligned} \tag{10}$$

$$= \exp\left[\sum_k \lambda_k h_k(\mathbf{x}, y) + \beta(y)\right] \tag{11}$$

$$\beta(y) = \log \int_{\mathbb{R}^2} \exp\left[\sum_k \lambda_k h_k(\mathbf{x}, y)\right] d\mathbf{x}, \tag{12}$$

where h_k are the different functions for which we have the sample averages. The density $\mathrm{p}_\lambda(\mathbf{X} \mid Y)$ is a mixture of multivariate Gaussian distributions. Both components $\mathrm{p}_\lambda(\mathbf{X} \mid y = -1)$ and $\mathrm{p}_\lambda(\mathbf{X} \mid y = 1)$ have the same covariance matrix.

For the following sections, we introduce the following notation for the expectations of the mixture of Gaussians.

$$\mathbb{E}[\mathbf{X} \mid y] = \boldsymbol{\mu}_y = \begin{bmatrix} \mu_{y,1} \\ \mu_{y,2} \end{bmatrix}, \qquad \mathbb{E}[\mathbf{X}\mathbf{X}^\top \mid Y] = \boldsymbol{\Sigma}_{\mathbf{X}|Y} \tag{13}$$

In addition, we will include the subscripts "causal" and "anticausal" where it might be ambiguous (e.g., $\boldsymbol{\Sigma}_{\mathbf{X}|Y,\text{causal}}$ represents the conditional covariance in the causal scenario and $\boldsymbol{\Sigma}_{\mathbf{X}|Y,\text{anticausal}}$ in the anticausal scenario). As mentioned in Propostion 3, the conditional covariance $\Sigma_{\mathbf{X}|Y}$ is the same for both values of y. However, we keep the conditional notation to distinguish it from the *marginal* covariance of $\mathbf{X}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ introduced in Equation (2). In Appendix A we derive the conditional expectations in $Equation$ (13) and the marginal expectations used as constraints.

Remark 4 Even though the causal graph in the anticausal direction implies that the conditional covariance $\Sigma_{\mathbf{X}Y}$ is diagonal, the CMAXENT solution does not result in a diagonal conditional covariance. This is true because of the constraints in Equations (1) and (2) and the law of total covariance. Note that the CMAXENT distribution is not necessarily Markov relative to the given DAG. As shown in [17, Section 5], CMAXENT only provides the best guess and may therefore mix over different Markovian distributions such that the result is no longer Markovian.

This relation between the marginal and conditional expectations will be essential in the subsequent sections where we explore the difference in the decision boundaries of the two resulting predictors of Y.

3.3 The predictor of Y in the anticausal direction

Recall that our main goal is to produce a predictor of Y as a function of the covariates $\mathbf{X}$. In Section 3.1 we obtain the predictor of Y directly as a result of the CMAXENT principle, given that the predictor is already in the direction of the causal mechanism. On the other hand, in Section 3.2, we have to derive the predictor of Y using the found conditional distributions and Bayes' rule. The main result of this section is that with the constraints we have used, CMAXENT in the anticausal direction is equivalent to Linear Discriminant Analysis [14, Section 4.3]. Furthermore, we generalise this result to Quadratic Discriminant Analysis, and to an exponential family version of discriminant analysis.

Theorem 5 (Predictor of Y using Bayes' rule)**.** *Using the results from Propostion 3, the density $\mathrm{p}_\lambda(Y = y \mid \mathbf{X})$ is the ratio of the product of the Gaussian component with $\mathrm{p}_\lambda(Y = y)$ and the mixture of Gaussians resulting from Propostion 3. Minimising the expected 0-1 loss, the optimal decision rule arising from this density is equivalent to Linear Discriminant Analysis (LDA).*

Corollary 6 (Quadratic Discriminant Analysis (QDA))**.** *Quadratic Discriminant Analysis can be interpreted as CMAXENT in the anticausal direction. This is achieved by replacing*

$$\mathbb{E}[X_i^2] = \bar{s}_i^2 \text{ with } i \in \{1, 2\}, \text{ and } \mathbb{E}[X_1 X_2] = \bar{s}_{1,2}. \tag{14}$$

in Equation (9) *with the following constraints:*

$$\mathbb{E}[X_i^2 \mid y] = \bar{s}_{i,y}^2 \text{ with } i \in \{1, 2\}, \text{ and } \mathbb{E}[X_1 X_2 \mid y] = \bar{s}_{1,2,y}. \tag{15}$$

We will now extend this idea, where instead of modelling $\mathrm{p}(\mathbf{X})$ as a mixture of Multivariate Gaussians (with equal covariance in LDA or unequal covariance in QDA), $\mathrm{p}(\mathbf{X})$ now becomes a mixture of distributions, each coming from an exponential family of distributions corresponding to a more general set of constraints.

Corollary 7 (Exponential family discriminant analysis)**.** *Let f_i be an arbitrary measurable function and $\tilde{f}$ its corresponding sample average. In the general case where Y is a discrete variable and we have d covariates $\mathbf{X}$ in the anticausal direction, the CMAXENT problem with constraints of the form:*

$$\mathbb{E}[f_i(\mathbf{X}) \mid y] = \tilde{f}_{i,y}, \tag{16}$$

where $\tilde{f}_{i,y}$ are the sample averages of f_i for a specific y as in Section 2.2, results in $\mathrm{p}_\lambda(\mathbf{X} \mid Y)$ being a mixture of exponential family distributions which then can be inverted (using Bayes' rule) to a predictor of Y.

Remark 8 In the previous corollary, the functions f_i can be constant on any of the variables in $\mathbf{X}$.

This idea has been extended to use kernels as a way to map $\mathbf{X}$ into more complex feature spaces. The resulting algorithm is called Kernel Fisher discriminant analysis [26, 32, 9].

3.4 The geometry of the decision boundaries

Hastie et al. [14, Chapter 4.4.5] conclude that the log-posterior odds of the logistic regression and LDA are both linear in $\mathbf{x}$, but with different parameters defining the linear relation. In this section, we revisit these results in more detail and explore whether the CMAXENT solution in causal direction differs from the solution in anticausal direction. From a statistical decision theory perspective, the log-posterior odds correspond to the Maximum A Posteriori (MAP) rule, the optimal decision boundary of a classifier when minimising the expected 0-1 loss [3, Ch. 4.3.3].

Proposition 9 (Normal vector to the decision boundaries in causal and anticausal direction)**.** *Under the 0-1 loss, the normal vector to the decision boundary of the CMAXENT predictor is proportional to*

1. $\mathbf{\Sigma}^{-1}_{\mathbf{X},causal}\phi$ *in the causal direction.*

2. $\mathbf{\Sigma}^{-1}_{\mathbf{X}|Y,anticausal}\phi$ *in the anticausal direction.*

We will now prove that in the case where we know all the expectations in Equations (1) and (2), the slope of the decision boundaries in causal and anticausal direction are the same.

Theorem 10 (Slope of the decision boundary is the same in causal and anticausal direction)**.** *Using the constraints in Equations* (1) *and* (2)*, the slope of* $\mathrm{p}_\lambda(Y \mid \mathbf{X})$ *inferred using CMAXENT is the same in causal and anticausal direction.*

Although it might seem from the above result that there is no asymmetry between the logistic regression and LDA even when we include causal information, this is not entirely true. To begin with, the decision boundaries may be unequal although they are parallel, but more importantly learning the parameters of certain model might be easier. In the next section we explore the differences that persist even under the light of Theorem 10.

3.5 What are the differences?

In the previous sections we found that the slopes of the decision boundary of CMAXENT in both the causal and anticausal direction are linear and agree, whenever we have the first and second moments as in Equations (1) and (2). This implies that, if the test data will come from the same distribution as the training data, either algorithm will work equally well. Previous research has studied the advantages and disadvantages [14, 33, Chapter 4.4.5] of each method and their properties such as asymptotic relative efficiency [7], parameter bias [13], asymptotic error under label noise [4] and online learning performance [1]. All of these analyses base their results on the fact that the logistic regression does not make an assumption on how the covariates $\mathbf{X}$ are distributed, whereas LDA does.

An alternative way of viewing this distinction is through the lens of generative and discriminative models. LDA is a generative model since it models both covariates and target variable and logistic regression only models the target as a function of the input. Ng and Jordan [28] analyse the difference in efficiency between the Naive Bayes algorithm (a generative model similar to LDA) and logistic regression, and find that both models have regimes in which they perform better than the other. Using the same models, Blöbaum et al. [5] and data from [35], find empirically that generative models perform better in anticausal than in causal direction.

4 Partially known covariances

In this section we explore variations of the solution of the CMAXENT solution in causal and anticausal direction when some of the sample averages are not known. In Section 4.1 we explore the case where the covariance between a particular predictor and the target is not known, and in Section 4.2 the case where we do not know the covariance between the predictors. In both cases we will see that the models we can infer (that is, $\mathrm{p}(Y \mid \mathbf{X})$) with CMAXENT will depend on the underlying causal assumptions.

4.1 Unknown predictor-target covariance

Without loss of generality, suppose we do not have the sample covariance between X_2 and Y, that is, we do not know ϕ_2 in Equation (1).

In the **causal direction**, the CMAXENT solution of the distribution of the causes $\mathbf{X}$ will still be a multivariate normal distribution with expectations given by the constraints relating $\mathbf{X}$. The conditional density of the effects is the logistic-like regression of Equation (7), however, λ_2 will be 0, as this is the parameter corresponding to the covariance between X_2 and Y. In other words, X_2 becomes irrelevant in the estimation of our target predictor.

In the **anticausal direction**, the distribution of the cause Y is unchanged because $\mathrm{P}(Y)$ is determined by the constraints and thus does not depend on ϕ_2. However, using the fact that the Gaussian distribution maximises the entropy over all distributions with the same variance [37, Theorem 8.6.5.], we can derive a bound on ϕ_2. We use the entropy of the Gaussian distribution because we do not know a closed form expression for the conditional covariance of $\mathrm{p}(Y \mid \mathbf{X})$ as given by Theorem 5.

Proposition 11 (Bounds on unknown covariance between predictor and target)**.** *Assuming the causal graph in Figure 1b and we do not know ϕ_2, an upper bound for ϕ_2 is given by:*

$$\frac{q(1-q)\bar{s}_{1,2}\phi_1}{q(1-q)\bar{s}_1^2 - \phi_1^2}. \tag{17}$$

The bound in Propostion 11 is found by differentiating the determinant of the conditional covariance (to which the differential entropy of the multivariate Gaussian is proportional to) with respect to the unknown covariance, ϕ_2 in this case, and finding the value for ϕ_2 for which this derivative is 0.

The implication of the previous result is that even when we have not observed any joint data between X_2 and Y we can still build a model of Y that depends on X_2, as long as we can assume the data generation process is anticausal. We can consider this an instance of Out Of Variable (OOV) generalisation studied in [16, 12], where we can exploit causal information and partial data to make models including variables that were never observed jointly with the target.

4.2 Unknown predictor covariance

Now suppose we observe all the sample averages in Equations (1) and (2) but we do not observe $\bar{s}_{1,2}$.

In the **causal** direction this implies that the multivariate Gaussian distribution resulting from the MAXENT problem on $\mathbf{X}$ is diagonal, that is, $\mathbf{X}$ are marginally independent. The exponential form of $\mathrm{p}_\lambda(Y \mid X)$ does not change, as we still observed q and $\boldsymbol{\phi}$, nevertheless, the parameters of the exponential family do change, as the density of $\mathbf{X}$ changed so that the resulting $\mathrm{p}_\lambda(Y \mid \mathbf{X})$ needs to adapt in order to match $\boldsymbol{\phi}$.

Now we will explore the **anticausal** case. We will proceed as in Sections 3.1 and 3.2. First we find $\mathrm{p}(Y)$ by maximising the entropy subject to the empirical average of Y, which is trivial because $\mathrm{p}(Y)$ is already determined by its moments, and then we find $\mathrm{p}(\mathbf{X} \mid Y)$ subject to all the moments in Equations (1) and (2) with the exception of $\bar{s}_{1,2}$. We obtain the following result from solving the CMAXENT optimisation problem

Proposition 12 (Diagonal conditional covariance in the anticausal direction with unknown predictor covariance)**.** *The density $\mathrm{p}(\mathbf{X} \mid Y)$ that maximises the conditional entropy subject to the following constraints:*

$$\hat{\mathbb{E}}[\mathbf{X}Y] = \begin{bmatrix}\phi_1 \\ \phi_2\end{bmatrix}, \quad \hat{\mathbb{E}}[\mathbf{X}] = \begin{bmatrix}0 \\ 0\end{bmatrix}, \quad \hat{\mathbb{E}}[X_1^2] = \bar{s}_1^2, \quad \hat{\mathbb{E}}[X_2^2] = \bar{s}_2^2, \tag{18}$$

and $\mathrm{p}(Y)$ inferred on the first step of CMAXENT, is independent after choosing a value of y; that is, $\mathbf{X}$ is conditionally independent given Y.

Remark 13 This result is reassuring given that under these moment constraints, $\mathrm{p}(\mathbf{X} \mid y)$ turns out to be Markov relative to the DAG in the anticausal direction. Contrary to Remark 4, where we concluded that CMAXENT is not always Markov relative to a DAG.

In Appendix E we derive the slopes of the decision boundaries in the causal and anticausal direction when we do not know the covariance between the predictors. We also find necessary and sufficient conditions for which the slopes are the same. From this simple example, we have learned the following: in causal direction, our inductive bias tells us that the covariates are not correlated and hence, the decision boundary depends only on the marginal variance of each $\mathbf{X}_i$ and the covariance between Y and $\mathbf{X}$. In the anticausal direction, CMAXENT infers $\mathbf{X}_1$ and $\mathbf{X}_2$ to be marginally correlated because they need to be conditionally independent (this fact can proved using the law of total covariance). Hence, the marginal covariance of $\mathbf{X}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ is different in both scenarios. This is something we did not observe in the case with full information (Section 3).

In addition, we derive the expressions of the decision boundaries in the causal and anticausal direction (see Appendix E). That is, as proved in Propostion 9, we have that the decision boundaries of the predictors in causal and anticausal direction will differ *with the same moments, but different causal assumptions*. In Figure 2, we showcase this phenomenon with synthetic data.

5 Discussion

In this article we have studied the differences arising from merging of predictors in the causal and anticausal directions. In particular, we have studied a simple case with a binary target variable and

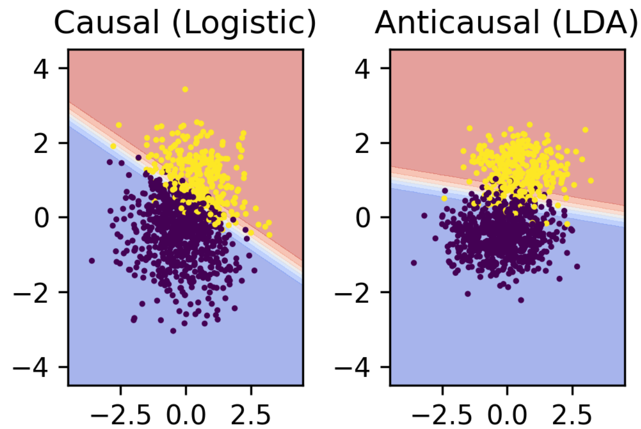

Figure 2: Decision boundaries of the solution of CMAXENT in the causal (left) and anticausal (right) direction when we do not have the covariance between the predictor variables $\bar{s}_{1,2}$.

two continuous variables. Although in this simple example we have already found connections with classical classification algorithms, and differences in the solutions in causal and anticausal direction, the example can be easily extended to more covariates (where the resulting distribution of the covariates would be a d-dimensional Gaussian instead of bivarate Gaussian), a discrete target variable (as in Corollary 7), and a causal graph that contains both parents and children as predictors.

As stated at the end of Section 3.1, the relation between merging of experts and logistic regression has been explored in the past. Friedman et al. [8] interpret the solution to the AdaBoost procedureas as additive logistic regression. They arrive at this interpretation starting from an exponential loss function. They then propose likelihood based estimator of the AdaBoost procedure. Thus, since our results align with those in Friedman et al. [8], we give yet another interpretation of AdaBoost as the solution of the merging of experts in causal direction using the CMAXENT principle.

Even though we have used CMAXENT as inductive bias to merge the predictors throughout the article, we believe that the asymmetries we found here (in particular, the geometry of the decision boundaries) would hold when using any other inductive bias that allows causal information to be included. Whatever method we use to merge predictors, the following asymmetry seems natural: In *anticausal* direction we try to *explain* correlations between X_1, X_2 as a result of Y influencing both components. In *causal* direction, correlations between X_1 and X_2 do not tell us anything about the relation between X and Y, following the principle of Independent Mechanisms (see [29] for an overview and [11] for a recent Bayesian view).

The previous observation is useful in straightforward scenarios where we are merging data from different sources for a supervised learning task, say datasets with overlapping variables, or datasets produced from different experimental conditions (also called environments); but also in cases where the merging of data is more subtle, for example, in federated learning where the notion of horizontal and vertical federated learning [42] coincides precisely with the data sources described above but where causality is underexplored.

Acknowledgments and Disclosure of Funding

We thank William R. Orchard and Yuchen Zhu for their valuable comments in a previous version of this article.

References

[1] A. Banerjee. An analysis of logistic models: Exponential family connections and online performance. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 204–215. SIAM, 2007.

[2] M. S. Bartlett. An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1):107–111, 1951.

[3] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[4] Y. Bi and D. R. Jeske. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. *Journal of Multivariate Analysis*, 101(7): 1622–1637, 2010.

[5] P. Blöbaum, S. Shimizu, and T. Washio. Discriminative and generative models in causal and anticausal settings. In *Advanced Methodologies for Bayesian Networks: Second International Workshop, AMBN 2015, Yokohama, Japan, November 16-18, 2015. Proceedings 2*, pages 209–221. Springer, 2015.

[6] L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[7] B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2): 337–407, 2000.

[9] B. Ghojogh, F. Karray, and M. Crowley. Fisher and kernel fisher discriminant analysis: Tutorial. *arXiv preprint arXiv:1906.09436*, 2019.

[10] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, pages 1367–1433, 2004.

[11] S. Guo, V. Toth, B. Schölkopf, and F. Huszar. Causal de Finetti: On the identification of invariant causal structure in exchangeable data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36463–36475. Curran Associates, Inc., 2023.

[12] S. Guo, J. Wildberger, and B. Schölkopf. Out-of-variable generalization for discriminative models. In *The Twelfth International Conference on Learning Representations (ICLR)*, May 2024.

[13] M. Halperin, W. C. Blackwelder, and J. I. Verter. Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *Journal of chronic diseases*, 24(2-3):125–158, 1971.

[14] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[15] T. K. Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[16] D. Janzing. Merging joint distributions via causal model classes with low VC dimension. *arXiv preprint arXiv:1804.03206*, 2018.

[17] D. Janzing. Causal versions of maximum entropy and principle of insufficient reason. *Journal of Causal Inference*, 9(1):285–301, 2021.

[18] D. Janzing and B. Schölkopf. Semi-supervised interpolation in an anticausal learning scenario. *Journal of Machine Learning Research*, 16:1923–1948, 2015. URL `http://jmlr.org/papers/v16/janzing15a.html`.

[19] D. Janzing, X. Sun, and B. Schölkopf. Distinguishing cause and effect via second order exponential models. *arXiv preprint arXiv:0910.5561*, 2009.

[20] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[21] Z. Jin, J. von Kügelgen, J. Ni, T. Vaidhya, A. Kaushal, M. Sachan, and B. Schoelkopf. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, 2021.

[22] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[23] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[24] G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. *Advances in neural information processing systems*, 14, 2001.

[25] W. B. Levy and H. Deliç. Maximum entropy aggregation of individual opinions. *IEEE transactions on systems, man, and cybernetics*, 24(4):606–613, 1994.

[26] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee, 1999.

[27] I. J. Myung, S. Ramamoorti, and A. D. Bailey Jr. Maximum entropy aggregation of expert predictions. *Management Science*, 42(10):1420–1436, 1996.

[28] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

[29] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[30] D. Poole and A. E. Raftery. Inference for deterministic simulation models: the bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255, 2000.

[31] A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[32] V. Roth and V. Steinhage. Nonlinear discriminant analysis using kernel functions. *Advances in neural information processing systems*, 12, 1999.

[33] Y. D. Rubinstein, T. Hastie, et al. Discriminative vs informative learning. In *KDD*, volume 5, pages 49–53, 1997.

[34] M. Saerens and F. Fouss. Yet another method for combining classifiers outputs: a maximum entropy approach. In *International Workshop on Multiple Classifier Systems*, pages 82–91. Springer, 2004.

[35] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[36] X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible markov kernels. In *Ninth International Symposium on Artificial Intelligence and Mathematics (AIMath 2006)*, pages 1–11, 2006.

[37] Thomas M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Nashville, TN, 2 edition, June 2006.

[38] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[39] S. Weichwald, B. Schölkopf, T. Ball, and M. Grosse-Wentrup. Causal and anti-causal learning in pattern recognition for neuroimaging. In *2014 International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4. IEEE, 2014.

[40] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[41] Y. Yao, L. M. Carvalho, and D. Mesquita. Locking and quacking: Stacking bayesian models predictions by log-pooling and superposition. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[42] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

A Relation between the expectations of the Mixture of Gaussians and the known marginal expectations

In Propostion 3 we proved that the distribution resulting from the constraints in Equations (1) and (2) and the anticausal optimisation problem in Equation (9) result in a mixture of Gaussian distributions. Now we are going to explore the relation between the moments of the resulting distribution and the constraints used in the MAXENT optimisation problem.

We have the following expectations under Gaussian mixture model

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}Y\right] =& q\boldsymbol{\mu}_1 - (1-q)\boldsymbol{\mu}_{-1} && (19)\\
\mathbb{E}\left[\mathbf{X}\right] =& q\boldsymbol{\mu}_1 + (1-q)\boldsymbol{\mu}_{-1} && (20)\\
\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right] =& \boldsymbol{\Sigma}_X && (21)\\
=& \mathbb{E}\left[\mathrm{Var}(\mathbf{X} \mid Y)\right] + \mathrm{Var}(\mathbb{E}\left[\mathbf{X} \mid Y\right]). && (22)
\end{aligned}$$

Where Equation (22) follows from the law of total covariance. We have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Var}(\mathbf{X} \mid Y)\right] =& q\boldsymbol{\Sigma}_{X|Y} + (1-q)\boldsymbol{\Sigma}_{X|Y} && (23)\\
=& \boldsymbol{\Sigma}_{X|Y}, && (24)\\
\mathrm{Var}(\mathbb{E}\left[\mathbf{X} \mid Y\right]) =& \mathbb{E}[\mathbb{E}[\mathbf{X} \mid Y]^2] - \mathbb{E}[\mathbb{E}[\mathbf{X} \mid Y]]^2 && (25)\\
\mathbb{E}[\mathbb{E}[\mathbf{X} \mid Y]^2] =& q\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + (1-q)\boldsymbol{\mu}_{-1}\boldsymbol{\mu}_{-1}^\top && (26)\\
\mathbb{E}[\mathbb{E}[\mathbf{X} \mid Y]]^2 =& (q\boldsymbol{\mu}_1 + (1-q)\boldsymbol{\mu}_{-1})(q\boldsymbol{\mu}_1 + (1-q)\boldsymbol{\mu}_{-1})^\top. && (27)
\end{aligned}$$

So that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right] =& \boldsymbol{\Sigma}_{X|Y} + q\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + (1-q)\boldsymbol{\mu}_{-1}\boldsymbol{\mu}_{-1}^\top \\
& - (q\boldsymbol{\mu}_1 + (1-q)\boldsymbol{\mu}_{-1})(q\boldsymbol{\mu}_1 + (1-q)\boldsymbol{\mu}_{-1})^\top && (28)\\
=& \boldsymbol{\Sigma}_{X|Y} + (1-q)q\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + (1-q)q\boldsymbol{\mu}_{-1}\boldsymbol{\mu}_{-1}^\top - (1-q)q\boldsymbol{\mu}_{-1}\boldsymbol{\mu}_1^\top - (1-q)q\boldsymbol{\mu}_1\boldsymbol{\mu}_{-1}^\top && (29)\\
=& \boldsymbol{\Sigma}_{X|Y} + (1-q)q[\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + \boldsymbol{\mu}_{-1}\boldsymbol{\mu}_{-1}^\top - \boldsymbol{\mu}_{-1}\boldsymbol{\mu}_1^\top - \boldsymbol{\mu}_1\boldsymbol{\mu}_{-1}^\top] && (30)\\
=& \boldsymbol{\Sigma}_{X|Y} + (1-q)q(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^\top. && (31)
\end{aligned}$$

Recall that the empirical averages used as constraints in the maximum entropy optimisation problem are coincide with the expectations under the resulting exponential family distribution. Then, using the equations above and the constraints, the means of the multivariate Gaussian distribution are

$$\begin{aligned}
\boldsymbol{\mu}_1 =& \frac{\bar{\mathbf{x}} + \boldsymbol{\phi}}{2q} && (32)\\
\boldsymbol{\mu}_{-1} =& \frac{\bar{\mathbf{x}} - \boldsymbol{\phi}}{2(1-q)}. && (33)
\end{aligned}$$

And the conditional covariance, which is the same for both components, is

$$
\begin{aligned}
\mathbf{\Sigma}_{X|Y} &= \begin{bmatrix} \bar{s}_1^2 & \bar{s}_{1,2} \\ \bar{s}_{1,2} & \bar{s}_2^2 \end{bmatrix} \\
&- q(1-q)\left[\frac{(\bar{\mathbf{x}}+\boldsymbol{\phi})(\bar{\mathbf{x}}+\boldsymbol{\phi})^\top}{2^2(1-q)^2} + \frac{(\bar{\mathbf{x}}-\boldsymbol{\phi})(\bar{\mathbf{x}}-\boldsymbol{\phi})^\top}{2^2 q^2}\right. \\
&- \left.\frac{(\bar{\mathbf{x}}+\boldsymbol{\phi})(\bar{\mathbf{x}}-\boldsymbol{\phi})^\top}{2^2 q(1-q)} - \frac{(\bar{\mathbf{x}}-\boldsymbol{\phi})(\bar{\mathbf{x}}+\boldsymbol{\phi})^\top}{2^2 q(1-q)}\right]
\end{aligned} \tag{34}
$$

$$
\begin{aligned}
\mathbf{\Sigma}_{X|Y} &= \begin{bmatrix} \bar{s}_1^2 & \bar{s}_{1,2} \\ \bar{s}_{1,2} & \bar{s}_2^2 \end{bmatrix} \\
&- q(1-q)\left[\frac{\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + \bar{\mathbf{x}}\boldsymbol{\phi}^\top + \boldsymbol{\phi}\bar{\mathbf{x}}^\top + \boldsymbol{\phi}\boldsymbol{\phi}^\top}{2^2(1-q)^2} + \frac{\bar{\mathbf{x}}\bar{\mathbf{x}}^\top - \bar{\mathbf{x}}\boldsymbol{\phi}^\top - \boldsymbol{\phi}\bar{\mathbf{x}}^\top + \boldsymbol{\phi}\boldsymbol{\phi}^\top}{2^2 q^2}\right. \\
&+ \left.\frac{-\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + \bar{\mathbf{x}}\boldsymbol{\phi}^\top - \boldsymbol{\phi}\bar{\mathbf{x}}^\top + \boldsymbol{\phi}\boldsymbol{\phi}^\top}{2^2 q(1-q)} + \frac{-\bar{\mathbf{x}}\bar{\mathbf{x}}^\top - \bar{\mathbf{x}}\boldsymbol{\phi}^\top + \boldsymbol{\phi}\bar{\mathbf{x}}^\top + \boldsymbol{\phi}\boldsymbol{\phi}^\top}{2^2 q(1-q)}\right]
\end{aligned} \tag{35}
$$

$$
\begin{aligned}
\mathbf{\Sigma}_{X|Y} &= \begin{bmatrix} \bar{s}_1^2 & \bar{s}_{1,2} \\ \bar{s}_{1,2} & \bar{s}_2^2 \end{bmatrix} \\
&- \frac{q(1-q)}{2^2 q^2 (1-q)^2}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top(q^2 + (1-q)^2 - q(1-q) - q(1-q)) \\
&+ \bar{\mathbf{x}}\boldsymbol{\phi}^\top(q^2 - (1-q)^2 + q(1-q) - q(1-q)) \\
&+ \boldsymbol{\phi}\bar{\mathbf{x}}^\top(q^2 - (1-q)^2 - q(1-q) + q(1-q)) \\
&+ \boldsymbol{\phi}\boldsymbol{\phi}^\top(q^2 + (1-q)^2 + q(1-q) + q(1-q))]
\end{aligned} \tag{36}
$$

$$
\begin{aligned}
\mathbf{\Sigma}_{X|Y} &= \begin{bmatrix} \bar{s}_1^2 & \bar{s}_{1,2} \\ \bar{s}_{1,2} & \bar{s}_2^2 \end{bmatrix} \\
&- \frac{1}{2^2 q(1-q)}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top(4q^2 - 4q + 1) + \bar{\mathbf{x}}\boldsymbol{\phi}^\top(2q-1) + \boldsymbol{\phi}\bar{\mathbf{x}}^\top(2q-1) + \boldsymbol{\phi}\boldsymbol{\phi}^\top]
\end{aligned} \tag{37}
$$

$$
\mathbf{\Sigma}_{X|Y} = \begin{bmatrix} \bar{s}_1^2 & \bar{s}_{1,2} \\ \bar{s}_{1,2} & \bar{s}_2^2 \end{bmatrix} - \frac{1}{2^2 q(1-q)}[(2q-1)\bar{\mathbf{x}} + \boldsymbol{\phi}][(2q-1)\bar{\mathbf{x}} + \boldsymbol{\phi}]^\top. \tag{38}
$$

We enumerate the individual elements:

$$
\mathbf{\Sigma}_{X|Y,\,(i,i)} = \frac{1}{2^2 q(1-q)}[\bar{s}_i^2 - (2q-1)^2 \bar{x}_i^2 - 2(2q-1)\bar{x}_i\phi_i - \phi_i^2] \tag{39}
$$

$$
\mathbf{\Sigma}_{X|Y,\,(1,2)} = \frac{1}{2^2 q(1-q)}[\bar{s}_{1,2} - (2q-1)^2 \bar{x}_1\bar{x}_2 - (2q-1)\bar{x}_1\phi_2 - (2q-1)\bar{x}_2\phi_1 - \phi_1\phi_2]. \tag{40}
$$

B Predictor in the anticausal direction

Theorem 5 (Predictor of Y using Bayes' rule)**.** *Using the results from Propostion 3, the density* $\mathrm{p}_\lambda(Y = y \mid \mathbf{X})$ *is the ratio of the product of the Gaussian component with* $\mathrm{p}_\lambda(Y = y)$ *and the mixture of Gaussians resulting from Propostion 3. Minimising the expected 0-1 loss, the optimal decision rule arising from this density is equivalent to Linear Discriminant Analysis (LDA).*

Proof. We prove this for the case $y = 1$. the case for $y = -1$ can be derived in an analogous way. The result follows from the application of Bayes' rule:

$$\mathrm{p}(y = 1 \mid \mathbf{x}) = \frac{\mathrm{p}(\mathbf{x} \mid y = 1)\mathrm{p}(y = 1)}{\mathrm{p}(\mathbf{x})} \tag{41}$$

$$= \frac{q \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right)}{q \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right) + (1-q)\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{-1})^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{-1})\right)} \tag{42}$$

$$= \frac{1}{1 + \frac{(1-q)}{q}\exp\left(-\frac{1}{2}(2\mathbf{x}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}) + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_{-1}\right)} \tag{43}$$

$$= \frac{1}{1 + \frac{(1-q)}{q}\exp\left((\mathbf{x} - \frac{\boldsymbol{\mu}_{-1}}{2})^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_{-1} - (\mathbf{x} - \frac{\boldsymbol{\mu}_1}{2})^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_1\right)} \tag{44}$$

□

C Derivation of the decision boundary

In the following two sections we give the proof of Propostion 9 for the causal and anticausal direction separately. First, we restate the proposition

Proposition 9 (Normal vector to the decision boundaries in causal and anticausal direction)**.** *Under the 0-1 loss, the normal vector to the decision boundary of the CMAXENT predictor is proportional to*

1. $\boldsymbol{\Sigma}_{\mathbf{X},causal}^{-1}\boldsymbol{\phi}$ *in the causal direction.*
2. $\boldsymbol{\Sigma}_{\mathbf{X}|Y,anticausal}^{-1}\boldsymbol{\phi}$ *in the anticausal direction.*

As mentioned on the proposition, we frame these results within the statistical decision theory framework [3], choosing a particular loss function $L(h(\mathbf{x}), y)$, where $h(\mathbf{x})$ is the predictor of y we want to evaluate. We consider the 0-1 loss function. That is, $L(h(\mathbf{x}), y) = 1$ if $h(\mathbf{x}) = y$, and 0 otherwise. The optimal decision rule for this loss is the well-known Maximum A Posteriori (MAP) rule from which we can derive our decision boundary.

C.1 Proof of Propostion 9 in the causal direction

In the causal direction, the Maximum A Posteriori (MAP) rule, results in the decision boundary given by the following equation

$$\mathrm{p}(y = 1 \mid \mathbf{x}) = \mathrm{p}(y = -1 \mid \mathbf{x}) \tag{45}$$

$$\frac{1}{2}(1 + \tanh(\lambda_0 + \lambda_1 x_1 + \lambda_2 x_2)) = \frac{1}{2}(1 + \tanh(-\lambda_0 - \lambda_1 x_1 - \lambda_2 x_2)) \tag{46}$$

$$\lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 = -\lambda_0 - \lambda_1 x_1 - \lambda_2 x_2 \tag{47}$$

$$\lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 = 0. \tag{48}$$

In words, the decision boundary in the causal direction is a linear function of the covariates. Using this result, we proceed to prove the relation between the marginal covariance matrix and the normal to the decision boundary as in Item 1 of Propostion 9

We want to prove $\boldsymbol{\lambda} \propto \Sigma_{\mathbf{X}}^{-1}\Sigma_{\mathbf{X},Y} = \Sigma_{\mathbf{X}}^{-1}\boldsymbol{\phi}$.

First, we define the random variable $Z := \lambda_1 X_1 + \lambda_2 X_2$. We can write $\mathrm{p}(y = 1|\mathbf{x})$ entirely as function of Z, thus $\mathbf{X} \perp\!\!\!\perp Y \,|Z$.

To continue with the proof, we consider the Hilbert space of centered random variables with basis given by $\mathrm{span}\,(\mathbf{X})$ and covariance as inner product. Following this geometric interpretation, we define $W_j := X_j - \alpha_j Z$, where $\alpha_j Z$ is the projection of X_j onto the span of Z. That is,

$\alpha_j = \text{Cov}[X_j, Z]\text{Var}(Z)^{-1}$. We have that $\mathbf{W} = \{W_1, W_2\} \in \text{span}(\mathbf{X})$, so that $\text{Cov}[Z, \mathbf{W}] = 0$. As a result, $\mathbf{W} \perp\!\!\!\perp Z$ because all variables in the span of $\mathbf{X}$ are Gaussian. Together with $\mathbf{W} \perp\!\!\!\perp Y \mid Z$, this implies via the semi-graphoid axioms [23] that $\mathbf{W} \perp\!\!\!\perp (Y, Z)$, so that $\mathbf{W} \perp\!\!\!\perp Y$ and thus $\text{Cov}[\mathbf{W}, Y] = 0$.

Taking the inner product with Y on both sides of $X_j = \alpha_j Z + W_j$ gives us $\text{Cov}[X_j, Y] = \alpha_j \text{Cov}[Z, Y] + \text{Cov}[W_j, Y] = \alpha_j \text{Cov}[Z, Y] = \text{Cov}[X_j, Z]\text{Var}(Z)^{-1}\text{Cov}[Z, Y]$. This is valid for $j = 1, 2$, hence $\Sigma_{\mathbf{X},Y} = \Sigma_{\mathbf{X},Z}\sigma_Z^{-2}\Sigma_{Z,Y}$. By definition, $\Sigma_{\mathbf{X},Z} = \boldsymbol{\lambda}\Sigma_{\mathbf{X}}$, so that $\Sigma_{\mathbf{X},Y} = \boldsymbol{\lambda}\Sigma_{\mathbf{X}}\sigma_Z^{-2}\sigma_{Z,Y}$ and finally $(\Sigma_{\mathbf{X}}\sigma_Z^{-2}\sigma_{Z,Y})^{-1}\Sigma_{\mathbf{X},Y} = \boldsymbol{\lambda}$, as required.

C.2 Proof of Propostion 9 in the anticausal direction

The normal to the decision boundary using the MAP rule in anticausal direction is derived in a similar way to Appendix C.1. In particular, the normal is given by the points of $\mathbf{x}$ where we are indifferent between choosing $y = 1$ and $y = -1$. To find such a vector, we solve for $\mathbf{x}$ in

$$\text{p}(y = 1 \mid \mathbf{x}) = \text{p}(y = -1 \mid \mathbf{x}) \tag{49}$$

$$\text{p}(\mathbf{x} \mid y = 1)\text{p}(y = 1) = \text{p}(\mathbf{x} \mid y = -1)\text{p}(y = -1). \tag{50}$$

In the second line of the above equation, we used $\text{p}(Y = y \mid \mathbf{x}) \propto \text{p}(\mathbf{x} \mid Y = y)P(Y = y)$.

We have

$$q \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right) = (1 - q) \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{-1})^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{-1})\right) \tag{51}$$

$$\log\left(\frac{q}{1 - q}\right) = -\frac{1}{2}(2\mathbf{x}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}) + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_{-1}). \tag{52}$$

The above equation is linear in $\mathbf{x}$, giving us a linear decision rule, and we would choose $Y = 1$ if

$$\mathbf{x}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}) > \frac{1}{2}\boldsymbol{\mu}_{-1}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_{-1} - \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\mu}_1 - \log\left(\frac{q}{1 - q}\right) \tag{53}$$

$$= \frac{1}{2}(\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}(\boldsymbol{\mu}_{-1} + \boldsymbol{\mu}_1) - \log\left(\frac{q}{1 - q}\right). \tag{54}$$

Remark 14 As mentioned in Theorem 5, the decision rule in Equation (53) is known as the Gaussian discriminant analysis [14]. This is a special case of Linear Discriminant Analysis. The family of LDA algorithms also contains Naive Bayes (if all the $\mathbf{x}$ are conditionally independent) and Quadratic Discriminant Analysis (QDA) (if the covariance matrices for each Y are not equal, giving a curved decision rule).

Now we will prove that if we have use all the moments in Equations (1) and (2) as constraints, the slope of the two decision boundaries are the same.

Theorem 10 (Slope of the decision boundary is the same in causal and anticausal direction)**.** *Using the constraints in Equations* (1) *and* (2)*, the slope of* $\text{p}_\lambda(Y \mid \mathbf{X})$ *inferred using CMAXENT is the same in causal and anticausal direction.*

Proof. In Propostion 9 we proved that in the causal direction, the normal vector to the decision boundary in the causal direction is $\Sigma_{\mathbf{X}}^{-1}\boldsymbol{\phi}$. Furthermore, using the law of total covariance (and the assumption that $\bar{x} = 0$), we can write $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{X}|Y} + c\boldsymbol{\phi}\boldsymbol{\phi}^\top$, where $c = 1/(2^2 q(1 - q))$ (see Equation (38)). Using the Sherman-Morrison formula [2], we can write

$$\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} = (\boldsymbol{\Sigma}_{\mathbf{X}|Y} + c\boldsymbol{\phi}\boldsymbol{\phi}^\top)^{-1} \tag{55}$$

$$= \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1} - \frac{c\boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}}{1 + c\boldsymbol{\phi}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi}}. \tag{56}$$

Applying this operator to $\boldsymbol{\phi}$, and noticing that $\boldsymbol{\phi}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi}$ is a scalar, we obtain

$$\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi} + k\boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi}, \tag{57}$$

where $k = (-c\boldsymbol{\phi}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi})/(1 + c\boldsymbol{\phi}^\top \boldsymbol{\Sigma}_{\mathbf{X}|Y}\boldsymbol{\phi})$. Thus $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\phi} \propto \boldsymbol{\Sigma}_{\mathbf{X}|Y}\boldsymbol{\phi}$, as required. □

D Missing covariance between the outcome variable and one of the covariates

Suppose we do not observe $\mathbb{E}[YX_2] = \phi_2$. Can CMAXENT say anything about $\mathrm{p}(Y \mid \mathbf{x})$? The answer is positive under the assumption that Y and X_2 are correlated. To see this, we will use the following result in information theory: The entropy of a distribution with given first and second moments is always less than the entropy of a multivariate Gaussian given the same first and second moments [37, Theorem 8.6.5]. Hence, we can analytically compute the maximum entropy solution of ϕ_2 via the entropy of the multivariate Gaussian as an upper bound on the entropy of $\mathbf{X}$ given Y.

To do this, we will use the results of Appendix A, where we found an expression of $\mathbf{\Sigma}_{\mathbf{X}|Y}$ as a function of ϕ_2 . Since ϕ_2 appears on several elements of the $\mathbf{\Sigma}_{\mathbf{X}|Y}$, we first compute the determinant of $\mathbf{\Sigma}_{\mathbf{X}|Y}$, differentiate with respect to ϕ_2 and equate to 0 to find the optimal ϕ_2.

For reference, the differential entropy of a multivariate Gaussian of k dimensions and covariance matrix $\mathbf{\Sigma}$ is

$$H(f) = \frac{k}{2} + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(\det(\mathbf{\Sigma})) \tag{58}$$

First we compute the determinant of $\mathbf{\Sigma}_{\mathbf{X}|Y}$:

$$\begin{aligned}
\det(\mathbf{\Sigma}_{X|Y}) =& \bar{s}_1^2\bar{s}_2^2 - \frac{\bar{s}_1^2(2q-1)^2\bar{x}_2^2 - \bar{s}_1^2 2(2q-1)\bar{x}_2\phi_2 - \bar{s}_1^2\phi_2^2}{2q(1-q)} \\
&- \frac{(2q-1)^2 x_1^2\bar{s}_2^2}{2q(1-q)} + \frac{(2q-1)^4\bar{x}_1^2\bar{x}_2^2 + 2(2q-1)^3\bar{x}_1^2\bar{x}_2\phi_2 + (2q-1)^2\bar{x}_1^2\phi_2^2}{(2q(1-q))^2} \\
&- \frac{2(2q-1)\bar{x}_1\phi_2\bar{s}_2^2}{2q(1-q)} + \frac{2(2q-1)^3\bar{x}_1\phi_1\bar{x}_2^2 + 2^2(2q-1)^2\bar{x}_1\bar{x}_2\phi_1\phi_2 + 2(2q-1)\bar{x}_1\phi_1\phi_2^2}{(2q(1-q))^2} \\
&- \frac{\phi_1^2\bar{s}_2^2}{2q(1-q)} + \frac{(2q-1)\bar{x}_2^2\phi_1^2 + 2(2q-1)\bar{x}_2\phi_1^2\phi_2 + \phi_1^2\phi_2^2}{(2q(1-q))^2} \\
&- \bar{s}_{1,2}^2 + \frac{\bar{s}_{1,2}(2q-1)^2\bar{x}_1\bar{x}_2 + \bar{s}_{1,2}(2q-1)\bar{x}_1\phi_2 + \bar{s}_{1,2}(2q-1)\bar{x}_2\phi_1 + \bar{s}_{1,2}\phi_1\phi_2}{2q(1-q)} \\
&+ \frac{(2q-1)^2\bar{x}_1\bar{x}_2\bar{s}_{1,2}}{2q(1-q)} - \frac{(2q-1)^4\bar{x}_1^2\bar{x}_2^2 - (2q-1)^3\bar{x}_1^2\bar{x}_2\phi_2 - (2q-1)^3\bar{x}_1\bar{x}_2^2\phi_1}{(2q(1-q))^2} \\
&- \frac{(2q-1)^2\bar{x}_1\bar{x}_2\phi_1\phi_2}{(2q(1-q))^2} + \frac{(2q-1)\bar{x}_1\phi_2\bar{s}_{1,2}}{2q(1-q)} - \frac{(2q-1)^3\bar{x}_1^2\phi_2\bar{x}_2 - (2q-1)^2\bar{x}_1^2\phi_2^2}{(2q(1-q))^2} \\
&- \frac{(2q-1)^2\bar{x}_1\bar{x}_2\phi_1\phi_2 - (2q-1)\bar{x}_1\phi_1\phi_2^2}{(2q(1-q))^2} + \frac{(2q-1)\bar{x}_2\phi_1\bar{s}_{1,2}}{2q(1-q)} \\
&- \frac{(2q-1)^3\bar{x}_1\bar{x}_2^2\phi_1 - (2q-1)^2\bar{x}_1\bar{x}_2\phi_1\phi_2}{(2q(1-q))^2} - \frac{(2q-1)^2\bar{x}_2^2\phi_1^2 - (2q-1)\bar{x}_2\phi_1^2\phi 2}{(2q(1-q))^2} \\
&+ \frac{\phi_1\phi_2\bar{s}_{1,2}}{2q(1-q)} - \frac{(2q-1)^2\bar{x}_1\bar{x}_2\phi_1\phi_2 - (2q-1)\bar{x}_1\phi_1\phi_2^2}{(2q(1-q))^2} - \frac{(2q-1)\bar{x}_2^2\phi_1^2\phi_2 - \phi_1^2\phi_2^2}{(2q(1-q))^2}.
\end{aligned} \tag{59}$$

Now we differentiate $\det\left(\mathbf{\Sigma}_{X|Y}\right)$ with respect to ϕ_2, equate to 0 and solve for ϕ_2

$$\begin{aligned}
\frac{\partial \det \mathbf{\Sigma}_{X|Y}}{\partial \phi_2} =& -\frac{\bar{s}_1^2 2(2q-1)\bar{x}_2}{2q(1-q)} - \frac{2\bar{s}_1^2\phi_2}{2q(1-q)} + \frac{2(2q-1)^3\bar{x}_1^2\bar{x}_2}{(2q(1-q))^2} + \frac{2(2q-1)^2\bar{x}_1^2\phi_2}{(2q(1-q))^2} \\
&+ \frac{2^2(2q-1)^2\bar{x}_1\bar{x}_2\phi_1}{(2q(1-q))^2} + \frac{2^2(2q-1)\bar{x}_1\phi_1\phi_2}{(2q(1-q))^2} + \frac{2(2q-1)\bar{x}_2\phi_1^2}{(2(2q-1))^2} + \frac{2\phi_1^2\phi_2}{(2(2q-1))^2} \\
&+ \frac{\bar{s}_{1,2}(2q-1)\bar{x}_1}{2q(1-q)} + \frac{\bar{s}_{1,2}\phi_1}{2q(1-q)} - \frac{(2q-1)^3\bar{x}_1^2\bar{x}_2}{(2(2q-1))^2} - \frac{(2q-1)^2\bar{x}_1\bar{x}_2\phi_1}{(2(2q-1))^2} \\
&+ \frac{(2q-1)\bar{x}_1\bar{s}_{1,2}}{2q(1-q)} - \frac{(2q-1)^3\bar{x}_1^2\bar{x}_2}{(2q(1-q))^2} - \frac{2(2q-1)^2\bar{x}_1^2\phi_2}{(2q(1-q))^2} - \frac{(2q-1)^2\bar{x}_1\bar{x}_2\phi_1}{(2q(1-q))^2} \\
&- \frac{2(2q-1)\bar{x}_1\phi_1\phi_2}{(2q(1-q))^2} - \frac{(2q-1)^2\bar{x}_1\bar{x}_2\phi_1}{(2q(1-q))^2} - \frac{(2q-1)\bar{x}_2\phi_1^2}{(2q(1-q))^2}.
\end{aligned} \tag{60}$$

Equating the above derivative to 0, we obtain

$$\begin{aligned}
&\phi_2[-2\bar{s}_1^2 q(1-q) + 2(2q-1)^2\bar{x}_1^2 + 2^2(2q-1)\bar{x}_1\phi_1 + 2\phi_1^2 - 2(2q-1)^2\bar{x}_1^2 - 2(2q-1)\bar{x}_1\phi_1] \\
&\quad = \bar{s}_1^2 2^2 2q(1-q)(2q-1)\bar{x}_2 - 2(2q-1)^3\bar{x}_1^2\bar{x}_2 - 2^2(2q-1)^2\bar{x}_1\bar{x}_2\phi_1 \\
&\quad - 2(2q-1)\bar{x}_2\phi_1^2 - \bar{s}_{1,2}2q(1-q)(2q-1)\bar{x}_1 - 2q(1-q)\bar{s}_{1,2}\phi_1 \\
&\quad + (2q-1)^3\bar{x}_1^2\bar{x}_2 + (2q-1)^2\bar{x}_1\bar{x}_2\phi_1 - (2q-1)\bar{x}_1\bar{s}_{1,2} \\
&\quad + (2q-1)^3\bar{x}_1^2\bar{x}_2 + (2q-1)^2\bar{x}_1\bar{x}_2\phi_1 + (2q-1)2\bar{x}_1\bar{x}_2\phi_1 + (2q-1)\bar{x}_2\phi_1^2.
\end{aligned} \tag{61}$$

Which can be simplified to

$$\begin{aligned}
\phi_2 =& \frac{1}{2\phi_1^2 + 2(2q-1)\bar{x}_1\phi_1 - 2\bar{s}_1^2 q(1-q)} \cdot \\
& [2^2 q(1-q)(2q-1)\bar{s}_1^2\bar{x}_2 - (2q-1)\bar{x}_2\phi_1^2 - 2q(1-q)(2q-1)\bar{s}_{1,2}\bar{x}_1 \\
& - 2q(1-q)\bar{s}_{1,2}\phi_1 - (2q-1)\bar{x}_1\bar{s}_{1,2} + (2q-1)^2 q(1-q)].
\end{aligned} \tag{62}$$

Although we have derived here the general case where the sample means are not zero, we will continue the analysis by coming back to such assumption. That is, $\bar{x}_1 = \bar{x}_2 = 0$, giving us the following expression for ϕ_2 which is easier to interpret

$$\phi_2 = \frac{q(1-q)\bar{s}_{1,2}\phi_1}{q(1-q)\bar{s}_1^2 - \phi_1^2}. \tag{63}$$

In the numerator, we have that as the covariance between X_1 and X_2 increases, the MAXENT covariance between X_2 and Y increases too. Furthermore, we see that the denominator is always greater than 1 by the Cauchy-Schwarz inequality of random variables, $\mathrm{Cov}(X_1, Y)^2 \leq \mathrm{Var}(X_1)\mathrm{Var}(Y)$, given that $q(1-q)$ is the variance of Y, $\bar{s}_1^2$ is the sample variance of X_1, and ϕ_1 is the sample covariance between X_1 and Y.

E Derivation of the decision boundary with unknown predictor covariance

Causal. In the causal case, we have that the distribution of the causes is a multivariate Gaussian with diagonal covariance matrix, and the conditional distribution of the target variable given the covariates is the same as in Equation (7).

As a result, we have that the decision boundary is still proportional to

$$\mathbf{\Sigma}_{\mathbf{X},causal}^{-1}\boldsymbol{\phi} = \begin{bmatrix} \bar{s}_1^{-2}\phi_1 \\ \bar{s}_2^{-2}\phi_2 \end{bmatrix} \tag{64}$$

as derived in section Appendix C.1.

Anticausal. In the anticausal direction, we have that the target variable follows a Bernoulli distribution, and the conditional distribution of the covariates given the target variable is again a mixture of Gaussians with diagonal conditional covariance matrix. We first prove Propostion 12.

Proposition 12 (Diagonal conditional covariance in the anticausal direction with unknown predictor covariance)**.** *The density* $\mathrm{p}(\mathbf{X} \mid Y)$ *that maximises the conditional entropy subject to the following constraints:*

$$\hat{\mathbb{E}}[\mathbf{X}Y] = \begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix}, \quad \hat{\mathbb{E}}[\mathbf{X}] = \begin{bmatrix}0\\ 0\end{bmatrix}, \quad \hat{\mathbb{E}}[X_1^2] = \bar{s}_1^2, \quad \hat{\mathbb{E}}[X_2^2] = \bar{s}_2^2, \tag{18}$$

and $\mathrm{p}(Y)$ *inferred on the first step of CMAXENT, is independent after choosing a value of* y*; that is,* $\mathbf{X}$ *is conditionally independent given* Y*.*

Proof. The solution to the constrained optimisation problem has the same form as in Equation (10), without the cross term:

$$\mathrm{p}_\lambda(\mathbf{x} \mid y) = \exp[\lambda_1 y x_1 + \lambda_2 y x_2 + \lambda_3 x_1 + \lambda_4 x_2 + \lambda_5 x_1^2 + \lambda_6 x_2^2 + \beta(y)]. \tag{65}$$

Conditioning on any specific value of Y, gives us an uncorrelated multivariate Gaussian, as required. □

Using Equation (31) we can express the conditional covariance as

$$\mathbf{\Sigma}_{\mathbf{X}|Y} = \mathbf{\Sigma}_{\mathbf{X}} - (1-q)q(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^\top \tag{66}$$

$$= \begin{bmatrix}\bar{s}_1^2 & \psi\\ \psi & \bar{s}_2^2\end{bmatrix} - q(1-q)\begin{bmatrix}(\mu_{1,1} - \mu_{-1,1})^2 & (\mu_{1,1} - \mu_{-1,1})(\mu_{1,2} - \mu_{-1,2})\\ (\mu_{1,2} - \mu_{-1,2})(\mu_{1,1} - \mu_{-1,1}) & (\mu_{1,2} - \mu_{-1,2})^2\end{bmatrix}. \tag{67}$$

Since we know that $\mathbf{\Sigma}_{\mathbf{X}|Y}$ is diagonal, then $\psi = q(1-q)(\mu_{1,1} - \mu_{-1,1})(\mu_{1,2} - \mu_{-1,2})$. From Equations (32) and (33), we can conclude that $q(1-q)(\mu_{1,i} - \mu_{-1,i})^2 \propto \phi_i^2$. Wit this, we find an expression of $\mathbf{\Sigma}_{\mathbf{X}|Y}$ as a function of the constraints

$$\mathbf{\Sigma}_{\mathbf{X}|Y} = \begin{bmatrix}\bar{s}_1^2 - \phi_1^2 & 0\\ 0 & \bar{s}_2^2 - \phi_2^2\end{bmatrix}. \tag{68}$$

On Appendix C.2 (see also Hastie et al. [14, Sec. 4.4.5]) we proved that the slope of the decision boundary in the anticausal direction is proportional to

$$\mathbf{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi}, \tag{69}$$

and using Equations (32) and (33), we have that the slope of the decision boundary is proportional to

$$\mathbf{\Sigma}_{\mathbf{X}|Y}^{-1}\boldsymbol{\phi} \propto \begin{bmatrix}(\bar{s}_1^2 - \phi_1^2)^{-1}\phi_1\\ (\bar{s}_2^2 - \phi_2^2)^{-1}\phi_2\end{bmatrix}. \tag{70}$$

A natural question arises: when are these slopes the same? in other words, when are Equations (64) and (70) linearly dependent? This question can be answered be equating

$$\frac{(\bar{s}_1^2 - \phi_1^2)^{-1}\phi_1 \bar{s}_2^2 \phi_2}{\bar{s}_1^2 \phi_1}, \tag{71}$$

to

$$(\bar{s}_2^2 - \phi_2^2)^{-1}\phi_2. \tag{72}$$

We have

$$\frac{(\bar{s}_1^2 - \phi_1^2)^{-1}\phi_1 \bar{s}_2^2 \phi_2}{\bar{s}_1^2 \phi_1} = (\bar{s}_2^2 - \phi_2^2)^{-1}\phi_2 \iff \tag{73}$$

$$(\bar{s}_2^2 - \phi_2^2)\phi_1 \bar{s}_2^2 \phi_2 = (\bar{s}_1^2 - \phi_1^2)\phi_2 \bar{s}_1^2 \phi_1 \iff \tag{74}$$

$$\frac{(\bar{s}_2^2 - \phi_2^2)\phi_1 \bar{s}_2^2 \phi_2}{(\bar{s}_1^2 - \phi_1^2)\phi_2 \bar{s}_1^2 \phi_1} = 1 \iff \tag{75}$$

$$\frac{(\bar{s}_2^2 - \phi_2^2)\bar{s}_2^2}{(\bar{s}_1^2 - \phi_1^2)\bar{s}_1^2} = 1. \tag{76}$$

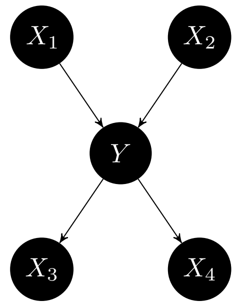

Figure 3: Graph in the causal and anticausal direction

F Derivation of the target predictor when merging predictors in causal and anticausal direction

In this section we explore the predictor resulting from merging predictors including causes and predictors including effects of the target variable. We will assume the causal graph in Figure 3. We will use the first and second moments of each variable, the covariance between each X_i and Y, the covariance between X_1 and X_2, and between X_3 and X_4, as constraints.

Using CMAXENT, we first find the density $\mathrm{p}(X_1, X_2)$ with maximum entropy subject to the moment constraints; then $\mathrm{p}(Y \mid X_1, X_2)$ with maximum entropy subject to the moment constraints (including $\mathrm{p}(X_1, X_2)$, found in the previous step); and finally the density $\mathrm{p}(X_3, X_4 \mid Y)$ that maximises the entropy subject to the moment constraints (again, including the found $\mathrm{p}(Y)$).

It is possible to see that these process will result in the same predictors as in Section 3. That is, we find that $\mathrm{p}(X_1, X_2)$ is a multivariate Gaussian, $\mathrm{p}(Y \mid X_1, X_2)$ a logistic-like regression, and $\mathrm{p}(X_3, X_4 \mid Y)$ a Mixture of Bivariate Gaussians.

These distributions provide us with enough information to find the joint distribution of all our variables $\mathrm{p}(Y, X_1, X_2, X_3, X_4)$, with which we can derive our predictor of interest. We have

$$\mathrm{p}(y \mid x_1, x_2, x_3, x_4) = \frac{\mathrm{p}(x_1, x_2, x_3, x_4 \mid y)\mathrm{p}(y)}{\mathrm{p}(x_1, x_2, x_3, x_4)} \tag{77}$$

$$= \frac{\mathrm{p}(x_1, x_2 \mid y)\mathrm{p}(x_3, x_4 \mid y)\mathrm{p}(y)}{\mathrm{p}(x_1, x_2, x_3, x_4)} \tag{78}$$

$$= \frac{\mathrm{p}(y \mid x_1, x_2)\mathrm{p}(x_1, x_2)\mathrm{p}(x_3, x_4 \mid y)\mathrm{p}(y)}{\mathrm{p}(y)\mathrm{p}(x_1, x_2, x_3, x_4)} \tag{79}$$

$$= \frac{\mathrm{p}(y \mid x_1, x_2)\mathrm{p}(x_1, x_2)\mathrm{p}(x_3, x_4 \mid y)}{\mathrm{p}(x_1, x_2, x_3, x_4)} \tag{80}$$

$$= \frac{\mathrm{p}(y \mid x_1, x_2)\mathrm{p}(x_1, x_2)\mathrm{p}(x_3, x_4 \mid y)}{\sum_y \mathrm{p}(x_1, x_2, x_3, x_4 \mid y)\mathrm{p}(y)} \tag{81}$$

$$= \frac{\mathrm{p}(y \mid x_1, x_2)\mathrm{p}(x_1, x_2)\mathrm{p}(x_3, x_4 \mid y)}{\sum_y \mathrm{p}(y \mid x_1, x_2)\mathrm{p}(x_1, x_2)\mathrm{p}(x_3, x_4 \mid y)}. \tag{82}$$

Where the second inequality follows from the conditional independence between $\{X_1, X_2\}$ and $\{X_3, X_4\}$ given Y, and the third inequality follows from Bayes' rule.

Equation (82) gives us the desired predictor. Notice that we found all of the elements needed to compute $p(y \mid \mathbf{x})$ on Section 3.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the theoretical claims in the abstract and introduction are part of the main article.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Since this is a theoretical paper, most of the limitations come from the assumptions made to obtain the theoretical results. In addition we discuss some potential computational problems, when applying these results in real world problems. These limitations are discussed in the main article, in any case.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This is the bulk of the paper. We provide clear assumptions and proofs of every result (except corollaries) on the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA] .

Justification: The paper does not contain any empirical results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA] .

   Justification: The paper does not include any data or code, besides a very small toy example.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: As above.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: As above.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: As above.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Due its theoretical nature, the paper adheres to NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We briefly discussed that practitioners should take some of these results into account but there is no broad impact in the potential application of the ideas outlined in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no major risks posed by the paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use any licensed assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.