When is invariant learning rational?

Abhimanyu Pallavi Sudhir

Supervisor: Jeroen Lamb

Abstract

We observe that invariant learning is sometimes effective in a supervised learning context even when the "nuisance factors" truly provide information on the target variable, and find sufficient conditions, in terms of information theory, for this to occur. Our work is the preliminary set-up for a significant line of research that could be opened into invariant learning in tasks where there are direct correlations between nuisance factors and the target variable.

Contents

1 Introduction

Invariant learning is a problem in machine learning, wherein one seeks to ensure the learned function (or distribution, etc.) is unaffected by some set of transformations of the input space. A simple example of such a set of transformations is the group of rigid transformations (rotations and translations) acting on an input space of images.

There is a significant corpus of literature detailing algorithms for learning invariantly to such particular transformation groups – the chief approaches among these are:

- **Manifestly invariant architectures**, i.e. that can only learn functions that are invariant to those transformations – e.g. group-convolutional neural networks [1] (of which the familiar convolutional neural networks are a special case). A particularly significant result (from [2]) is that any neural network architecture that is manifestly invariant to a compact group of transformations is a group-convolutional neural network.

- **Data augmentation**, i.e. adding transformations of the data points to "teach" the algorithm that these transformed inputs must be assigned the same label – see e.g. [3, 4].

We will highlight two specific line of research in the literature to motivate the problem we study in our project.

1. **The pathological 6 and 9 problem** (mentioned in e.g. [5, 6]) – If

you apply a rotation-invariant classification algorithm to distinguishing "6"s and "9"s, it will perform rather poorly, as a rotated "6" *is* in fact, very likely a "9", and should be read as such. [7] argues that the task can still be considered to be invariant, except to a different set (in particular a partial semigroup) of transformations including only rotations up to a certain angle; however, it remains true that naive rotation-invariant learning will not be effective for this task.

2. **Fairness in machine learning** (reviewed in e.g. [8]) – Consider a practical machine learning application used by say, a moneylender, to assess the probability of a loan applicant defaulting based on various variables (income, credit rating, etc.) Writers of machine learning ethics often argue [9] that some factors (typically "sensitive" variables like race and religion) ought to not be considered when predicting the output variable, as they only affect the output variable "through" their relationship with other input variables – and furthermore, that even if sensitive variables *do* provide non-redundant information (information not contained in the other input variables) about the output variable, they should still not be considered, for ethical reasons.

(In the latter problem, invariance is understood to mean "ignoring some variable (X_2) in favour of another (X_1) while learning the target (Y)". It might not immediately be clear how this is related to invariant learning as we introduced it – this is the *statistical* notion of invariance, which we will

motivate later in this section, and formally define in Sec 2.1, but it should be clear how this generalizes our earlier notion. The group orbits from before are our X_1, while X_2 determines the place of a data point on that orbit.)

In the first motivating problem, we ask: What property of the distribution (of hand-drawn digits) is it, mathematically, that tells us we should not expect invariant learning to be effective in this case? In the second problem, we ask: what can we say about the effectiveness of learning invariant to sensitive variables even when these variables provide non-redundant information about our output variable? Both problems have to do with the fundamental question, which we seek to approach in this project:

Question 1.1. *Under what circumstances will invariant learning (with respect to a particular set of transformations, group or otherwise) be effective or rational?*

The closest work to our own is [4], which we shall now briefly review.

[4] specifically looks at the effectiveness of data augmentation based approaches to invariant learning. Let the data $\{x_i\}_{i=1}^n$ be sampled from a random vector X taking values in some set $\mathcal{X}$; and let the task be to learn the distribution of X within some class of models parameterized by the parameter θ. Then whereas ordinary learning seeks to minimize $\sum_{i=1}^n L(\theta, x_i)$, the paper argues that augmented learning is equivalent to minimizing $\sum_{i=1}^n \bar{L}(\theta, x_i)$ where:

$$\bar{L}(\theta, x) = \int_G L(\theta, gX)\, d\mathbb{Q}(g)$$

And G is a compact topological group acting on X, $\mathbb{Q}(g)$ the Haar measure on G from which the augmentations are assumed to be sampled (we use the Haar measure because we want to think of the transformations as "uniformly distributed").

$\bar{L}(\theta, x)$ may then be viewed as the conditional expectation of $L(\theta, x)$ over the random variable $[x]$, the orbit of x under G. It then follows from the Rao-Blackwell theorem that if $[X]$ is sufficient for θ (which occurs if $\forall g \in G$, $X =_d gX$ ($=_d$ indicating equality in distribution)), then $\bar{L}$ has the same mean, and lower variance than L. Thus if $\forall g \in G$, $X =_d gX$ (what they call "exact invariance"), then invariant learning is effective.

However, our work differs from [4] in several key ways:

1. Their work only considers the effectiveness of data augmentation based approaches to learning; instead, we seek to abstract away the method by which invariance is achieved, and instead formulate Question 1.1 more generally: when is it effective to use an invariant learning approach, at all?

2. Our main result has to do with the case where exact invariance does *not* hold. While they also generalize their result to the case of "approximate invariance" $\forall g \in G$, $X \approx_d gX$ ($\approx_d$ indicating approximate equality in distribution, defined formally in terms of the Wasserstein metric), this

is quite different from the case of we consider, as we will see.

3. Their work does not readily generalize to supervised learning tasks – whereas we only study supervised learning tasks.

   In fact, we make here a correction to a claim made on p. 7 of the paper, it is claimed that for supervised learning applications where we seek to infer the conditional distribution $\mathrm{P}(Y \mid X)$, we can ask for exact invariance of (X, Y), i.e. $\forall g \in G,\ (X, Y) \approx_d g \cdot (X, Y)$, where the action of g on Y is trivial. They claim that this means "the probability of an image being a bird is [...] the same as the probability for a rotated image".

   However, this is not quite true – exact invariance of (X, Y) should actually be read as "the probability of finding an image and it being a bird, is the same as finding a rotated image and that being a bird". For example, if the rotated image is much less likely to be found in a real (non-augmented) dataset, then we don't have exact invariance of (X, Y), however this should not affect our decision to use invariant learning to infer $\mathrm{P}(Y \mid X)$.

   The result of [4] remains correct, as certainly exact invariance of (X, Y) implies the effectiveness of invariant learning – however, this condition is simply stronger than necessary for supervised learning tasks, and we will later formulate a more appropriate notion of "exact invariance" for conditional distributions in Sec 5.

We will mention that the second motivation we presented – fairness in machine learning – suggests a fundamental link between this problem and *causal modeling*. We will mathematically formulate causal modeling in Sec 2.2, but for now we will say that the idea that X_2 provides only redundant information on Y once given X_1 is fundamentally a statement of conditional independence: we say Y and X_2 are conditionally independent given X_1, or $\mathrm{P}(Y \mid X_1, X_2) = \mathrm{P}(Y \mid X_1)$, or $Y \perp\!\!\!\perp X_2 \mid X_1$. As we will see, this is the same as saying that Y and X_2, while *correlated*, are not directly causally linked – performing an intervention on X_2 (i.e. changing X_2 while leaving X_1 unchanged) will not change Y.

(Therefore in some fundamental sense, invariant learning is about learning causations rather than mere correlations. One may say that invariant learning is to be performed in the *anticausal direction*, i.e. to learn Y from X when Y is believed to be a cause for X rather than the other way around. This fact has been understood in the literature at least informally – e.g. the standard Bible of causal inference [10] highlights that anticausal learning is performed when the mechanisms being learned are "disentangled". We will not dwell on this highly abstracted perspective on causal inference, but it serves to put our work in perspective, and to motivate the use of a statistical rather than deterministic definition of invariance.)

Certainly invariant learning – informally for our purposes, ignoring some variable X_2 while learning $\mathrm{P}(Y \mid X_1, X_2)$ – is effective when X_2 is redundant for Y given X, i.e. when $Y \perp\!\!\!\perp X_2 \mid X_1$. However, we often see that invariant

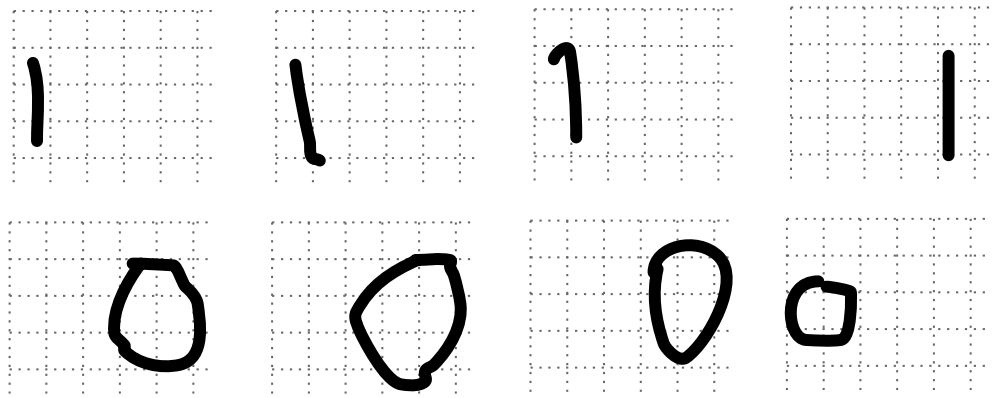


Figure 1.1: The target variable, the label, is directly correlated with the nuisance variable, the position of the image.

learning is sometimes still effective even when this is *not* the case.

In Fig 1.1, even though X_2 provides significant information on Y, we may still ignore it, because X_1 alone is also sufficient to determine Y with a great degree of accuracy. However, if we were classifying "6"s and "9"s, the orbit of a character under rotation would *not* carry sufficient information about Y – indeed, the orbit of a "6" image could indicate either a "6" or a "9", so ignoring the rotation and relying only on the orbit in learning the label would not be effective.

This very intuitive result is, in a nutshell, what we will aim to formally explain in this project.

An immediate implication of our result will be to fairness in machine learning – it will bound the error when applying "fair" (invariant) algorithms even when sensitive variables provide non-redundant information. For this reason and others, we suggest that further theoretical study of invariant

learning in situations with non-redundant information will be valuable, and outline directions for future research in this area.

1.1 Notation and conventions

Below is a list of notation we will use throughout this report without definition:

- *Probability theory* – Let Ω be a sample space, let $\mathcal{A}_i, \mathcal{B}_j, \mathcal{C}$ ($1 \leq i \leq m$, $1 \leq j \leq n$) be measure spaces: in particular, each $\mathcal{A}_i$ is a countable set with the counting measure α_i, and each $\mathcal{B}_j$ is a copy of $\mathcal{R}^{n_j}$ with the Lesbesgue measure β_j.
  - A measurable function $\Omega \to \mathcal{C}$ is called a "random quantity on $\mathcal{C}$". Hereon, we will omit mentioning the sample space Ω and assume that all random quantities are defined on the same sample space.
  - $\mathrm{P}\left(A_1, \dots A_n, B_1, \dots B_n\right)$ denotes, for A_i, B_j random quantities on $\mathcal{A}_i, \mathcal{B}_j$ respectively, the probability density function with respect to $\alpha_1 \otimes \cdots \otimes \alpha_m \otimes \beta_1 \otimes \cdots \otimes \beta_n$. Analogously for conditional probabilities.
  - $\mathbb{E}\left[\dots\right]$ is the expectation. If the term within the parentheses is a function of some random quantities $f(C_1, \dots C_k)$, we may write the expectation as $\mathbb{E}_{c_1,\dots c_k \sim C_1,\dots C_k}\left[f(c_1, \dots c_k)\right]$. Analogously for conditional expectation with respect to a random variable, we write $\mathbb{E}_{c_1,\dots c_k | c_{k+1}\dots c_l \sim C_1,\dots C_k | C_{k+1}\dots C_l}\left[f(c_1, \dots c_k) \mid C_{k+1} = c_{k+1} \dots C_l = c_l\right] =$

$\int f(c_1, \ldots c_k)\, d(\gamma_1 \otimes \cdots \otimes \gamma_k \mid \gamma_{k+1} \otimes \cdots \otimes \gamma_k)$ where γ_i denote the probability measures on the respective sets in which C_i take values, and $\mu \mid \nu$ denotes the regular conditional probability measure of μ against ν.

  - The notation $A \perp\!\!\!\perp B$ represents independence, and is to be read A and B are independent, i.e. $\mathrm{P}(A \mid B) = \mathrm{P}(A)$; $A \perp\!\!\!\perp B \mid C$ represents conditional independence, and is to be read "A and B are independent, conditional on C", i.e. $\mathrm{P}(A \mid B, C) = \mathrm{P}(A \mid C)$.

  - The notation $=_d$ reads "equal in distribution to", and the notation $\approx_d$ reads "approximately equal in distribution to".

- *Graph theory* – Let $\mathbf{V}$ be a graph consisting of vertices labelled $V_1, \ldots V_n$.

  - Parents(V_i) is the set of parents of V_i in the graph $\mathbf{V}$, i.e. of the nodes V_k with an arrow going $V_k \to V_i$.

  - Descendants(V_i) is the set of descendants of V_i in the graph $\mathbf{V}$, i.e. of the nodes V_k such that there exists some ordered list $(V_{i_1}, V_{i_2}, \ldots V_{i_m})$ $(m \geq 1)$ where $V_{i_1} = V_i$, $V_{i_m} = V_k$ and $\forall j < m$ there is an edge directed $V_{i_j} \to V_{i_{j+1}}$. Note that in particular $V_i \in$ Descendants(V_i).

- *Information theory* – We use the standard notations for entropies; three particular entropies that will be of relevance to us are listed below. All logarithms are base 2 unless stated otherwise.

– Let p and q be distributions on some discrete space $\mathcal{B}$ – then $\mathrm{H}[p,q] := -\sum_{b\in\mathcal{B}} p(b)\log q(b)$ denotes the cross-entropy of the distributions.

– Let A and B be random quantities on measurable spaces $\mathcal{A}$, $\mathcal{B}$ respectively; then $\mathrm{H}[B \mid A] := -\mathbb{E}_{a,b\sim A,B}[\log \mathrm{P}(b \mid a)]$ denotes the conditional entropy of B given A.

2 Prerequisites and literature

2.1 Invariance

We will detail a sequence of definitions of invariance from the literature, short of the generalized definitions we will give in Secs 3, 5.

Def 2.1 gives the "canonical" or most basic definitions of equivariance and invariance, as used in [2] (similar definitions are assumed in [1, 11–14]).

Definition 2.1 (Invariance – deterministic; group)**.** Where G is a group acting on $\mathcal{X}$ and $\mathcal{Y}$, a function $f : \mathcal{X} \to \mathcal{Y}$ is said to be G-equivariant if $\forall g \in G$ and $\forall x \in \mathcal{X}$:

$$f(g(x)) = g(f(x))$$

Of particular interest is the case where the action of G on $\mathcal{Y}$ is id – f is said to be G-invariant if $\forall g \in G$ and $\forall x \in \mathcal{X}$:

$$f(g(x)) = f(x)$$

Def 2.2 generalizes Def 2.1 to any partition on $\mathcal{X}$, not just the quotient space of the group action, $\mathcal{X}/G$. We mention this generalization, as this will closely connect to our later definitions of invariance.

Definition 2.2 (Invariance – deterministic; partition)**.** Where R is an equivalence relation on $\mathcal{X}$, a function $f : \mathcal{X} \to \mathcal{Y}$ is said to be R-invariant if it is

a function purely on the partition defined by R, i.e. $\forall x, x' \in \mathcal{X}$:

$$[x]_R = [x']_R \implies f(x) = f(x')$$

These definitions so far have been for a *deterministic function* f. In this project, we will instead treat learning as a statistical inference problem, i.e. of inferring the distribution $\mathrm{P}(Y \mid X)$ from data. We will state two statistical formulations of invariance and discuss their relevance (or lack thereof) to our problem, so as to better justify the definition we will make in Sec 3.

Definition 2.3 (Invariance – definition in [15])**.** Let $\mathcal{X}, \mathcal{Y}$ be sets ($\mathcal{Y}$ discrete), and X, Y are random quantities taking values in $\mathcal{X}, \mathcal{Y}$ (i.e. they are measurable functions $\Omega \to \mathcal{X}$, $\Omega \to \mathcal{Y}$ for sample space Ω). Now let:

- T be a *representation* of X: a random quantity on a set $\mathcal{T}$ such that $T \perp\!\!\!\perp Y \mid X$.
- N be a *nuisance* to Y: a random quantity on a set $\mathcal{N}$ such that $Y \perp\!\!\!\perp N$.

Then T is said to be an N-invariant representation if $T \perp\!\!\!\perp N$.

Def 2.3 is the formulation of invariance in Def [15] – it defines what it means for a representation (given by a random variable of its own) to be invariant to some nuisance variable.

This definition is somewhat hard to motivate, because the goal of [15] is different from ours – it is to give an explanation for the effectiveness of deep neural networks, arguing that this has to do with a certain propensity

for deep neural networks to learn invariant representations when the number of hidden layers is large. In particular, it has a crucial limitation for our purposes: it *presumes*, a priori, that $Y \perp\!\!\!\perp N$ – whereas our central result Thm 5.3 says essentially that invariant learning is rational when $Y \perp\!\!\!\perp N$, and that it can still be rational otherwise under certain conditions.

Definition 2.4 (Invariance – definition in [4])**.** Let $\mathcal{Z}$ be a set, Z be a random quantity taking values in $\mathcal{Z}$, and G be a group acting on $\mathcal{Z}$. Then Z is said to be G-invariant if for any $g \in G$, $gZ =_d Z$ (i.e. gZ has the same distribution as Z).

Def 2.4 is the formulation of invariance in Def [4] – it defines what it means for a distribution to be invariant. This is a fundamentally different idea than invariant representations – we're no longer talking about the invariance of an inferred distribution, but of the "true" distribution of the data. One may imagine that the relationship is closely relevant to Question 1.1 – indeed, if the "true" distribution is invariant (and we will there define this in a slightly more general way than below), then it is effective to learn invariantly.

2.2 Causation

As discussed in Sec 1, there is a close relationship between invariance and causal modeling. We will formally introduce causal modeling in this section; however, as causation is not the primary topic of this project, we will not dwell too long explaining the definitions, nor will we provide proofs for the

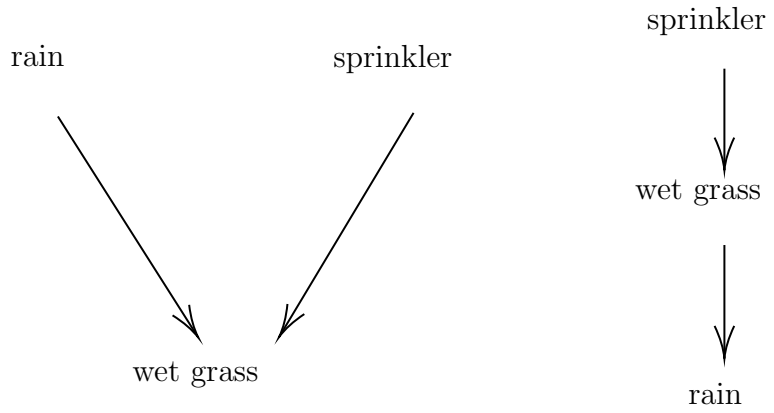


Figure 2.1: Distinct causal models

propositions – full treatments can be found in [10, 16].

At first, causation might seem like something outside the domain of serious statistics – indeed, for a long period of history, statisticians believed that causation could not be determined without knowing timestamps, or without performing direct interventions. However, this is no longer the case – indeed, causation must be a physically valid concept, because it is testable: for example, in Fig 2.1: "rain causes wet grass" can be distinguished from "wet grass causes rain" by turning on the sprinkler and checking if this results in rain.

We will shortly define formally what these arrows mean; but intuitively, a statement of causation can fundamentally only be expressed *relative* to some set of "control variables", namely in terms of independences conditional on these control variables.

Definition 2.5 (Causal model)**.** Let the random variables $X_1, \ldots X_n$ be sampled from some joint distribution $\mathrm{P}(X_1, \ldots, X_n)$. Have $\mathbf{X}$ be a Di-

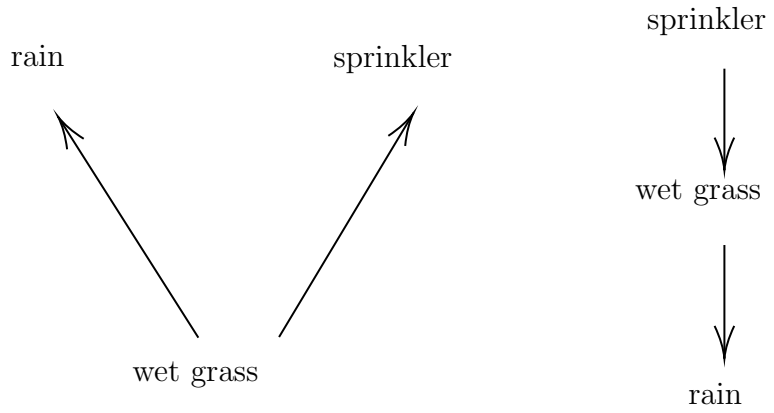


Figure 2.2: Equivalent causal models

rected Acyclic Graph (DAG) of $X_1, \ldots X_n$ (i.e. with these random variables uniquely represented as its nodes); then the ordered list

$$(\mathbf{X}, \mathrm{P}\left(X_1 \mid \mathrm{Parents}(X_1)\right), \ldots, \mathrm{P}\left(X_n \mid \mathrm{Parents}(X_n)\right))$$

is called a causal model for $\mathrm{P}\left(X_1, \ldots X_n\right)$.

Def 2.5 alone makes no non-trivial statement about the joint distribution $\mathrm{P}\left(X_1, \ldots, X_n\right)$ – indeed, any such causal model could be constructed for any joint distribution. What gives meaning to a causal model is the *Causal Markov condition* (Def 2.6), or equivalently Markov factorization (Prop 2.7), or equivalently *d*-Separation (Prop 2.9).

Briefly: the Causal Markov condition states that a causal model implies certain conditional independencies, the "hypothesis" entailed by the model; Markov factorization makes this hypothesis algebraically explicit – as the statement that specifying certain conditional distributions suffices to deter-

mine the joint distribution; d-Separation provides a rather intuitive picture for conditional independence as the absence of "unblocked paths" between two nodes for information to flow through.

Definition 2.6 (Causal Markov condition)**.** A causal model as in Def 2.5 is said to be *causally Markov* if $\forall X \in \mathbf{X}, X \perp\!\!\!\perp \text{Descendants}(X)^{\text{C}} \mid \text{Parents}(X)$

Proposition 2.7 (Markov factorization)**.** *A causal model as in Def 2.5 is causally Markov if and only if:*

$$\text{P}\left(X_1, \ldots, X_n\right) = \prod_i \text{P}\left(X_i \mid \text{Parents}(X_i)\right)$$

Definition 2.8 (d-Separation)**.** In any DAG $\mathbf{X}$, an ordered list of nodes $(X_{i_1}, \ldots X_{i_m})$ such that for all j, there is either an edge directed $X_{i_j} \to X_{i_{j+1}}$ or an edge directed $X_{i_j} \leftarrow X_{i_{j+1}}$ – is called a *path* and is denoted as $X_{i_1} \to X_{i_2} \leftarrow \cdots \leftarrow X_{i_m}$ depending on which edge exists between each consecutive pair of nodes. Each node $X_{i_2}, \ldots X_{i_{m-1}}$ is called a "junction" – a node X_{i_j} is classified into one of three types, as follows:

- "Chain" if there are edges directed $X_{i_{j-1}} \to X_{i_j} \to X_{i_{j+1}}$
- "Fork" if there are edges directed $X_{i_{j-1}} \leftarrow X_{i_j} \to X_{i_{j+1}}$
- "Collider" if there are edges directed $X_{i_{j-1}} \to X_{i_j} \leftarrow X_{i_{j+1}}$

Let $\mathbf{Y} \subseteq \mathbf{X}$. Then we say a junction X_{i_j} is "blocked when conditioned on $\mathbf{Y}$" iff one of the following holds (and unblocked ditto otherwise):

- It is a chain junction, and Descendants(X_{i_j}) $\cap \mathbf{Y} \neq \varnothing$.
- It is a fork junction, and Descendants(X_{i_j}) $\cap \mathbf{Y} \neq \varnothing$.
- It is a collider junction, and Descendants(X_{i_j}) $\cap \mathbf{Y} = \varnothing$.

The ordered list $(X_{i_1}, \ldots X_{i_m})$ is said to be "blocked when conditioned on $\mathbf{Y}$" iff any one of its junctions $X_{i_2}, \ldots X_{i_{m-1}}$ is blocked when conditioned on $\mathbf{Y}$.

Two nodes X_1 and X_2 are said to be "d-separated when conditioned on $\mathbf{Y}$" ($X_1 \nleftrightarrow X_2 \mid \mathbf{Y}$) if all paths between them are blocked when conditioned on $\mathbf{Y}$.

Two subsets $\mathbf{X}_1, \mathbf{X}_2 \subseteq \mathbf{X}$ are said to be "d-separated when conditioned on $\mathbf{Y}$" ($\mathbf{X}_1 \nleftrightarrow \mathbf{X}_2 \mid \mathbf{Y}$) if each pair $X_1 \in \mathbf{X}_1$, $X_2 \in \mathbf{X}_2$ are d-separated when conditioned on $\mathbf{Y}$.

Proposition 2.9 (d-Separation)**.** *A causal model as in Def 2.5 is causally Markov if and only if* $\forall \mathbf{X}_1, \mathbf{X}_2, \mathbf{Y} \subseteq \mathbf{X}$, $\mathbf{X}_1 \nleftrightarrow \mathbf{X}_2 \mid \mathbf{Y} \implies \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \mid \mathbf{Y}$.

Note that multiple DAGs may be consistent with the same hypothesis, i.e. lead to the same joint distribution – indeed, a joint distribution defines an equivalence class of DAGs (called *Markov equivalence*), rather than a unique DAG. Fig 2.1 shows an example of inequivalent causal models while Fig 2.2 shows an example of equivalent causal models.

The relationship between equivariance/invariance and causal modeling is alluded to in several areas in the literature: [8] describes that disentangled

representation (used as a synonym for equivariant representation in the reference) decomposes the data into generative factors that may be understood as causal parent variables, or in terms of conditional independence relations that represent invariances; [17] describes that disentangled representation is a consequence of the Markov factorization (referred to as "disentangled mechanisms") for particular causal graphs; [18–20] all introduce algorithms to obtain disentangled representations through Markov factorization; this relationship was an important theme of the NIPS 2017 workshop "Learning Disentangled Representations: from Perception to Control".

2.3 Miscellaneous mathematics

We will use the cross-entropy loss as our loss function throughout this report – because as we will see, it is most natural to think of our main result in terms of information theory. Lemma 2.10 will be useful to us in this regard.

Lemma 2.10 (Expected cross-entropy loss)**.** *Let $\mathcal{X}, \mathcal{Y}$ be sets ($\mathcal{Y}$ discrete), X, Y be random quantities taking values in $\mathcal{X}, \mathcal{Y}$ and $Q(y \mid x)$ be a distribution in y that is also a function in x – for some IID sample $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$ of (X, Y), we define the cross-entropy loss of Q on $\mathcal{D}$ as:*

$$L(Q, \mathcal{D}) = -\frac{1}{n} \sum_{(x,y) \in \mathcal{D}} Q(y \mid x)$$

Then the risk (expected loss) of Q is given by:

$$R(Q) = \mathbb{E}_{x\sim X}\left[\mathrm{H}\left[\mathrm{P}\left(y \mid x\right), Q(y \mid x)\right]\right]$$

Proof.

$$\begin{aligned}
R(Q) &= \mathbb{E}\left[L(Q, \mathcal{D})\right] \\
&= \frac{1}{n}\sum \mathbb{E}_{x,y\sim X,Y}\left[-\log Q(y \mid x)\right] \\
&= \mathbb{E}_{x,y\sim X,Y}\left[-\log Q(y \mid x)\right] \\
&= \mathbb{E}_{x\sim X}\left[\mathbb{E}_{y|x\sim Y|X}\left[-\log Q(y \mid x) \mid X = x\right]\right] \\
&= \mathbb{E}_{x\sim X}\left[\mathrm{H}\left[\mathrm{P}\left(y \mid x\right), Q(y \mid x)\right]\right]
\end{aligned}$$

□

Lemma 2.11 (Entropy cannot increase on conditioning)**.** *Let $A, B_1, \ldots B_n, B_{n+1}$ be random variables; then*

$$\mathrm{H}\left[A \mid B_1, \ldots B_n, B_{n+1}\right] \leq \mathrm{H}\left[A \mid B_1, \ldots B_n\right]$$

Proof.

$$\begin{aligned}
\mathrm{H}\left[A \mid B_1, \ldots B_n, B_{n+1}\right] &= \mathrm{H}\left[A, B_{n+1} \mid B_1, \ldots B_n\right] - \mathrm{H}\left[B_{n+1} \mid B_1, \ldots B_n\right] \\
&\leq \mathrm{H}\left[A \mid B_1, \ldots B_n\right] + \mathrm{H}\left[B_{n+1} \mid B_1, \ldots B_n\right] - \mathrm{H}\left[B_{n+1} \mid B_1, \ldots B_n\right] \\
&= \mathrm{H}\left[A \mid B_1, \ldots B_n\right]
\end{aligned}$$

□

We will also use a basic fact about group orbits to relate our formalism to group-theoretic formulations of invariance.

Lemma 2.12 (Product of group and set of orbits)**.** *Let $\mathcal{X}$ be a set and G be a group acting on this set such that all stabilizers are trivial (i.e. $\forall x \in \mathcal{X}$, $\{g : gx = x\} = \{1_G\}$). Then, where $\mathcal{X}/G$ is the set of orbits (i.e. its elements are the equivalence classes of $\mathcal{X}$ defined by the equivalence relation $x \sim y \iff \exists g, gx = y$), there is a bijection $\phi : \mathcal{X}/G \times G \to \mathcal{X}$.*

Proof. Let $\xi : \mathcal{X}/G \to X$ be some choice function, i.e. $\forall O \in \mathcal{X}/G, \xi(O) \in O$, i.e. $\xi(O)$ is some representative element in the orbit O. Then construct $\phi(O, g) := g\xi(O)$.

- Surjectivity: Let $x \in \mathcal{X}$. Where $[x]$ is the orbit of x, we have $\xi([x]) \in [x]$ thus $\exists g, g\xi([x]) = x$ (by definition of the equivalence relation that defines orbits) – hence $\phi([x], g) = g\xi([x]) = g$.

- Injectivity: suppose $\phi(O_1, g_1) = \phi(O_2, g_2)$, i.e. $g_1\xi(O_1) = g_2\xi(O_2) = x$: note that $\xi(O_1) \in O_1$, and since O_1 is an orbit, $g_1\xi(O_1) \in O_1$,

analogously $g_2\xi(O_2) \in O_2$. Thus $x \in O_1 \cap O_2$; but since the orbits are a partition, $O_1 \cap O_2 \neq \varnothing \implies O_1 = O_2$. Now (where $O = O_1 = O_2$), $g_1\xi(O) = g_2\xi(O) \implies g_2^{-1}g_1\xi(O) = \xi(O)$, but since all stabilizers are trivial, we must have $g_2^{-1}g_1 = 1_G$, thus $g_1 = g_2$. Thus we have $(O_1, g_1) = (O_2, g_2)$.

□

3 Formalism

We seek to precisely formulate and answer the question "When is invariant learning rational?" To be as general as possible, and to exploit the machinery of causal reasoning, we will be using a statistical generalization (Def 3.3) of our earlier definition (Def 2.2) of invariance. For reference throughout the paper, we define the following problem framework, and examples thereof.

Framework 3.1 (The learning problem)**.** Let $\mathcal{X}, \mathcal{Y}$ be sets ($\mathcal{Y}$ discrete), and X, Y are random quantities taking values in $\mathcal{X}, \mathcal{Y}$ (i.e. they are measurable functions $\Omega \to \mathcal{X}$, $\Omega \to \mathcal{Y}$ for sample space Ω).

We then seek to estimate, or learn the distribution $\mathrm{P}(Y \mid X)$. We do so by taking an IID sample $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$, and minimizing the loss:

$$L(Q, \mathcal{D}) = -\frac{1}{n} \sum_{(x,y) \in \mathcal{D}} \log Q(y \mid x)$$

By Lemma 2.10, $\mathbb{E}\left[L(Q, \mathcal{D})\right] = \mathrm{H}\left[\mathrm{P}\left(y \mid x\right), Q(y \mid x)\right]$, which we know is minimized by $\mathrm{P}(Y \mid X)$. We will call the distribution that minimizes $L(Q, \mathcal{D})$ the "minimum-loss estimator for $\mathrm{P}(Y \mid X)$".

Definition 3.2 (Invariant representation - general)**.** In Framework 3.1, where R is an equivalence relation on $\mathcal{X}$, we say that a distribution $Q(y \mid x)$ is R-invariant if $\forall x, x' \in \mathcal{X}$, $[x]_R = [x']_R \implies Q(y \mid x) = Q(y \mid x')$.

It is straightforward to see that this reduces to Def 2.2 when $Q(Y \mid X)$ has singleton support over y (i.e. when Y is merely a function of X). A special

case – which we shall use as it will be sufficient for most of our purposes – is given in Def3.3:

Definition 3.3 (Invariant representation)**.** In Framework 3.1, suppose that we can factor $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, writing each $X = (X_1, X_2)$. We say that $Q(y \mid x)$ is X_2-invariant if $\forall x_1, x_2, x_2'$, $Q(y|(x_1, x_2)) = Q(y|(x_1, x_2'))$.

Prop 3.4 describes the relationship between our Def 3.3 and that in [15] – one could imagine this as providing intuition for [15], as its original definition seemed rather un-motivated, whereas Def 3.3 is quite intuitive.

Proposition 3.4 (Def 2.3 vs. Def 3.3)**.** *Consider the set-up in Def 3.3, and suppose $Y \perp\!\!\!\perp X_2$ (i.e. X_2 is a nuisance to Y). Let T be any random variable with probability distribution given by $\mathrm{P}(T = t \mid X = x, Y = y) = Q(t \mid x)$. T is X_2-invariant (Def 2.3) iff $Q(y \mid x)$ is X_2-invariant (Def 3.3).*

Proof. Trivial:

$$\begin{aligned} & Q(t \mid (x_1, x_2)) \\ = \ & \mathrm{P}(T = t \mid X_1 = x_1, X_2 = x_2, Y = y) \\ = \ & \mathrm{P}(T = t \mid X_1 = x_1, X_2 = x_2', Y = y) \\ = \ & Q(t \mid (x_1, x_2')) \end{aligned}$$

And swap lines (12)(34) to show the converse. □

Definition 3.5 (Invariant learning)**.** In Framework 3.1, suppose that we can factor $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, writing each $X = (X_1, X_2)$. We once again seek to learn

the distribution $\mathrm{P}(Y \mid X)$ by taking an IID sample $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$, and minimizing $L(Q, \mathcal{D})$ – but this time, only among X_2-invariant distributions.

Note that the X_2-invariant distributions $Q(y \mid x)$ are in one-to-one correspondence with distributions in y that are functions in x_1: $\bar{Q}(y \mid x_1) := Q(y \mid (x_1, x_2)$ (which is well-defined because the right-hand-side is the same for any x_2); so this is equivalent to finding a $\bar{Q}(y \mid x_1)$ that minimizes the loss:

$$L(\bar{Q}, \mathcal{D}) = -\frac{1}{n} \sum_{(x,y) \in \mathcal{D}} \log \bar{Q}(y \mid x_1)$$

Since $\bar{Q}(y \mid x_1)$ is still a "distribution in y that is also a function of x", by Lemma 2.10, $L(\bar{Q}, \mathcal{D})$ is an unbiased estimator of $R(\bar{Q}) = \mathbb{E}_{x \sim X}\left[\mathrm{H}\left[\mathrm{P}(y \mid x), \bar{Q}(y \mid x_1)\right]\right]$. We will call the $\bar{Q}(y \mid x_1)$ that minimizes $L(\bar{Q}, \mathcal{D})$ the "minimum-loss estimator for $\mathrm{P}(Y \mid X_1)$"; we call the X_2-invariant $Q(y \mid x)$ that minimizes $L(Q, \mathcal{D})$ among X_2-invariant representations the "minimum-loss invariant estimator for $\mathrm{P}(Y \mid X)$".

(Note that because these loss functions are equal in value for corresponding $\bar{Q}, Q$, these estimators are also equal in value.)

4 Illustrative examples

We will now define several examples to motivate our research questions more thoroughly, and later to intuit the result in Thm 5.3. Ex 4.1 is the sort of task we'd actually like to study, while Ex 4.2,highlights toy examples that will represent our different cases where invariant learning might be suitable or effective, and will be very helpful in providing concrete interpretations of our causal diagrams.

Example 4.1 (Real example)**.** Let $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} = \{x : \mathbb{R}^2 \to \{0, 1\}\}$ be the set of images; define $\mathcal{X}_2 = \{T_v : \mathcal{X} \to \mathcal{X} : x \mapsto \lambda i.x(i - v) \mid v \in \mathbb{R}^2\} \cong \mathbb{R}^2$ to be the group of translations acting on $\mathcal{X}$; define $\mathcal{X}_1 = \mathcal{X}/\mathcal{X}_2$ to be the quotient by the group action (i.e. the orbits under translation). By Lemma 2.12, there is a bijection $\phi : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{X}$ given by $\phi(O, T_v) = T_v \xi(O)$ where $\xi(O)$ is some representative in O – in particular, one may choose the representative $\xi(O)$ to be the element x in O whose centroid $\mathrm{CM}\,[x] = \int_{i \in \mathbb{R}^2} x(i) i \, dA$ is 0 (it is easy to see that such an element exists in each orbit – pick any element and translate it by its centroid – and is unique in each orbit – any two elements differing by a translation differ in their centroid); then $\phi^{-1} : x \mapsto ([x], T_{\mathrm{CM}[x]})$. Composing this with the isomorphism $\mathcal{X}_2 \cong \mathbb{R}^2$ (given by $T_v \mapsto v$), we have $x \mapsto ([x], \mathrm{CM}\,[x])$. Thus the decomposition $\mathcal{X}_1 \times \mathcal{X}_2$ decomposes an image into its orbit under translations (its "shape") and its centroid (its "position").

This defines $\mathcal{X}, \mathcal{Y}$. Intuitively, one imagines $\mathrm{P}\,(X \mid Y)$ as representing the

process by which people draw "0"s and "1"s – every possible image has some probability of being drawn to show either a "0" or a "1". Any interesting distribution on $\mathcal{X}, \mathcal{Y}$ – one that could actually represent the generation of hand-drawn "0"s and "1"s in the real-world – would be too complicated to write down compactly, but we will qualitatively describe some possible distributions, for the purpose of illustration and understanding – see Fig 4.1 for a visualization.

(a) We randomly choose Y (whether to draw 0 or 1); then independently choose a position (X_2) to draw it at; then generate X_1 based on Y alone (we draw a shape based on the character it is supposed to represent); then proceed $X := (X_1, X_2)$ (the chosen shape is drawn at the chosen position).

(b) We randomly choose Y; then generate X_1 based on Y alone; then generate X_2 based on X_1 alone (e.g. if our wrist is twisted so that more rounded shapes are more likely to be drawn to the right and straighter shapes are more likely to be drawn to the left); then proceed $X := (X_1, X_2)$.

(c) We randomly choose Y, then generate X_1 from Y alone; then generate X_2 based on Y alone (e.g. for whatever reason, when we draw "0"s, we are more likely to draw them to the right, while when we draw "1"s, we are more likely to draw them to the left.

It might seem that (b) and (c) are identical (indeed, a sample from either

distribution is likely to look similar, like in Fig 4.2), but this is an artifact of Y being almost entirely determined by X_1. In (b), if we "mistakenly" draw a "0" like a "1", it will likely be to the left, while this is not so in (c).

Example 4.2 (Toy examples)**.** These toy examples will represent different cases where invariant learning might be suitable or effective. As we will see, the cases are distinguished essentially by their causal diagram.

(a) Let $\mathcal{X} = [0, 2)$, $\mathcal{Y} = \{0, 1\}$, and factor $\mathcal{X} = \{0, 1\} \times [0, 1)$ in the obvious way (by writing $x = (\lfloor x \rfloor, \{x\})$). Suppose that ($p$ is imagined close to 0):

$$Y \mid X \sim \begin{cases} \text{Bernoulli}(p) & \lfloor X \rfloor = 0 \\ \text{Bernoulli}(1-p) & \lfloor X \rfloor = 1 \end{cases}$$

And:

$$X \sim \text{Unif}(0, 2)$$

We have provided the distribution in this inverted form to visualize Y as a stochastic function of X (as in Fig 4.3); but we can also express it in a way that will be more useful for us later when we think of these terms causally:

$$Y \sim \text{Bernoulli}(1/2)$$

$$\mathrm{P}(X \mid Y) \quad = \quad \begin{array}{c|cc} & Y=0 & Y=1 \\ \hline 0 \le X < 1 & 1-p & p \\ 1 \le X < 2 & p & 1-p \end{array}$$

(b) Same as Ex 4.2 (a), but X is distributed as (q is imagined close to 0):

$$\mathrm{P}(X) = \begin{cases} 1-q & \frac{1}{2} \leq X < \frac{3}{2} \\ q & \text{otherwise} \end{cases}$$

Again, one may express this model as:

$$Y \sim \text{Bernoulli}(1/2)$$

$$\mathrm{P}(X \mid Y) \quad = \quad \begin{array}{r|rr} & Y=0 & Y=1 \\ \hline 0 \leq X < \frac{1}{2} & 2(1-p)q & 2pq \\ \frac{1}{2} \leq X < 1 & 2(1-p)(1-q) & 2p(1-q) \\ 1 \leq X < \frac{3}{2} & 2p(1-q) & 2(1-p)(1-q) \\ \frac{3}{2} \leq X < 2 & 2pq & 2(1-p)q \end{array}$$

(c) Let $\mathcal{X}, \mathcal{Y}$ be as before. Suppose that

$$Y \sim \text{Bernoulli}(1/2)$$

And that $\mathrm{P}(\lfloor X \rfloor, \{X\} \mid Y) = \mathrm{P}(\lfloor X \rfloor \mid Y)\,\mathrm{P}(\{X\} \mid Y)$ where:

$$\lfloor X \rfloor \mid Y \sim \begin{cases} \text{Bernoulli}(p) & Y = 0 \\ \text{Bernoulli}(1-p) & Y = 1 \end{cases}$$

$$\mathrm{P}\left(\{X\} \mid Y\right) \quad = \quad \begin{array}{c|cc} & Y=0 & Y=1 \\ \hline 0 \leq \{X\} < \frac{1}{2} & q & 1-q \\ \frac{1}{2} \leq \{X\} < 1 & 1-q & q \end{array}$$

Exs 4.2 (a), 4.2 (b), 4.2 (c) are illustrated in Fig 4.3 – on the left with a sample from the joint distribution of (X, Y), and on the right a causal diagram consistent with the joint distribution (i.e. which is a Markov factorization if the joint distribution). The correctness of the causal diagrams can easily be verified by hand, and they are motivated below.

Ex 4.2 (a) is the simplest example, in which $\{X\}$ is distributed (uniformly, in fact) independent of any other variables. This is the most elementary situation in which invariant learning is clearly desirable – we should *not* take $\{X\}$ into account while learning $\mathrm{P}\left(Y \mid X\right)$, as there is simply no flow of information between $\{X\}$ and Y.

Ex 4.2 (b) is more interesting. Y and $\{X\}$ are certainly *correlated* – if $Y = 0$, $\{X\} > 1/2$ with probability $1 - q$, and if $Y = 1$, $\{X\} > 1/2$ with probability q. However, the relationship is not a *causal* one – changing $\{X\}$ won't change the distribution of Y, changing Y won't change the distribution of $\{X\}$ (provided $\lfloor X \rfloor$ is the same). Indeed, the joint distribution can be factored as:

$$\mathrm{P}\left(\lfloor X \rfloor, \{X\}, Y\right) = \mathrm{P}\left(\{X\} \mid \lfloor X \rfloor\right) \mathrm{P}\left(\lfloor X \rfloor \mid Y\right) \mathrm{P}\left(Y\right)$$

Where:

$$\mathrm{P}(Y) = \begin{cases} 1/2 & Y = 0 \\ 1/2 & Y = 1 \end{cases}$$

$$\mathrm{P}(\lfloor X \rfloor \mid Y) \quad = \quad \begin{array}{c|cc} & Y=0 & Y=1 \\ \hline \lfloor X \rfloor = 0 & 1-p & p \\ \lfloor X \rfloor = 1 & p & 1-p \end{array}$$

$$\mathrm{P}(\{X\} \mid \lfloor X \rfloor) \quad = \quad \begin{array}{c|cc} & \lfloor X \rfloor = 0 & \lfloor X \rfloor = 1 \\ \hline 0 \leq \{X\} < \frac{1}{2} & q & 1-q \\ \frac{1}{2} \leq \{X\} < 1 & 1-q & q \end{array}$$

Thus $Y \longrightarrow \lfloor X \rfloor \longrightarrow \{X\}$ is a valid causal diagram for Ex 4.2 (b).

Should we, then, take $\{X\}$ into account while learning $\mathrm{P}(Y \mid X)$? Certainly not, as $\mathrm{P}(Y \mid X)$ does not depend on $\{X\}$ – we have $Y \perp\!\!\!\perp \{X\} \mid \lfloor X \rfloor$.

Ex 4.2 (c), however, models a distribution that – while it superficially looks similar to that in Ex 4.2 (b) (the left sides of Figs 4.3b, 4.3c are similar) – has a fundamentally different causal diagram. Even if we know $\lfloor X \rfloor$ – say, if we know that $\lfloor X \rfloor = 0$ – knowing $\{X\} > 1/2$ will indicate that Y is more likely to be 0. So really, we *should* take $\{X\}$ into account when learning $\mathrm{P}(Y \mid X)$.

However, in this case, even if we *don't* take $\{X\}$ into account – even if we were to just learn $\mathrm{P}(Y \mid \lfloor X \rfloor)$ and apply that to predict Y from X on our

data, we'd still get a low loss. Indeed – we can compute in this example:

$$\mathrm{P}\left(Y \mid X\right) \quad = \quad \begin{cases} \begin{array}{c|cc} & \lfloor X \rfloor = 0 & \lfloor X \rfloor = 1 \\ \hline 0 \leq \{X\} < \frac{1}{2} & \frac{(1-p)q}{(1-p)q+p(1-q)} & \frac{pq}{pq+(1-p)(1-q)} \\ \frac{1}{2} \leq \{X\} < 1 & \frac{(1-p)(1-q)}{pq+(1-p)(1-q)} & \frac{p(1-q)}{(1-p)q+p(1-q)} \end{array} & (Y=0) \\ \\ \begin{array}{c|cc} & \lfloor X \rfloor = 0 & \lfloor X \rfloor = 1 \\ \hline 0 \leq \{X\} < \frac{1}{2} & \frac{p(1-q)}{(1-p)q+p(1-q)} & \frac{(1-p)(1-q)}{pq+(1-p)(1-q)} \\ \frac{1}{2} \leq \{X\} < 1 & \frac{pq}{pq+(1-p)(1-q)} & \frac{(1-p)q}{(1-p)q+p(1-q)} \end{array} & (Y=1) \end{cases}$$

$$\mathrm{P}\left(Y \mid \lfloor X \rfloor\right) \quad = \quad \begin{cases} \begin{array}{cc} \lfloor X \rfloor = 0 & \lfloor X \rfloor = 1 \\ \hline 1-p & p \end{array} & (Y=0) \\ \\ \begin{array}{cc} \lfloor X \rfloor = 0 & \lfloor X \rfloor = 1 \\ \hline p & 1-p \end{array} & (Y=1) \end{cases}$$

Then for some sample $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$ of (X, Y), suppose we predict $\hat{y}^i$ with the distribution $\mathrm{P}\left(Y \mid \lfloor X \rfloor = \lfloor x_i \rfloor\right)$ (i.e. if our invariant learning algorithm computes $\mathrm{P}\left(Y \mid \lfloor X \rfloor\right)$ exactly) and compute our loss as:

$$L_{\text{inv}}(\mathcal{D}) = -\frac{1}{n} \sum_{(x,y)\in\mathcal{D}} \log \mathrm{P}\left(Y = y \mid \lfloor X \rfloor = \lfloor x \rfloor\right)$$

Then the expected loss is (as per Prop 2.10):

$$\mathbb{E}\left[L_{\text{inv}}\right] = \mathbb{E}_{x\sim X}\left[\mathrm{H}\left[\mathrm{P}\left(Y = y \mid X = x\right), \mathrm{P}\left(Y = y \mid \lfloor X \rfloor = \lfloor x \rfloor\right)\right]\right]$$

And we can compute it as:

$$\begin{aligned}
&\mathrm{H}\left[\mathrm{P}\left(Y = y \mid X = x\right), \mathrm{P}\left(Y = y \mid \lfloor X \rfloor = \lfloor x \rfloor\right)\right] \\
&= -\sum_{y\in\mathcal{Y}} \mathrm{P}\left(Y = y \mid X = x\right) \log \mathrm{P}\left(Y = y \mid \lfloor X \rfloor = \lfloor x \rfloor\right) \\
&= -\mathrm{P}\left(Y = 0 \mid X = x\right) \log \mathrm{P}\left(Y = 0 \mid \lfloor X \rfloor = \lfloor x \rfloor\right) \\
&\quad -\mathrm{P}\left(Y = 1 \mid X = x\right) \log \mathrm{P}\left(Y = 1 \mid \lfloor X \rfloor = \lfloor x \rfloor\right)
\end{aligned}$$

$=$

	$\lfloor x \rfloor = 0$	$\lfloor x \rfloor = 1$
$0 \le \{x\} < \frac{1}{2}$	$-\frac{(1-p)q}{(1-p)q+p(1-q)} \log(1-p)$ $-\frac{p(1-q)}{(1-p)q+p(1-q)} \log(p)$	$-\frac{pq}{pq+(1-p)(1-q)} \log(p)$ $-\frac{(1-p)(1-q)}{pq+(1-p)(1-q)} \log(1-p)$
$\frac{1}{2} \le \{x\} < 1$	$-\frac{(1-p)(1-q)}{pq+(1-p)(1-q)} \log(1-p)$ $-\frac{pq}{pq+(1-p)(1-q)} \log(p)$	$-\frac{p(1-q)}{(1-p)q+p(1-q)} \log(p)$ $-\frac{(1-p)q}{(1-p)q+p(1-q)} \log(1-p)$

Noting that (by simple application of law of total probability):

$$\mathrm{P}(X) \quad = \quad \begin{array}{c|cc} & \lfloor X \rfloor = 0 & \lfloor X \rfloor = 1 \\ \hline 0 \leq \{X\} < \frac{1}{2} & \frac{(1-p)q+p(1-q)}{2} & \frac{pq+(1-p)(1-q)}{2} \\ \frac{1}{2} \leq \{X\} < 1 & \frac{pq+(1-p)(1-q)}{2} & \frac{(1-p)q+p(1-q)}{2} \end{array}$$

The desired expectation simplifies as:

$$\mathbb{E}\left[L_{\text{inv}}\right] = \frac{(1-p)\log(1-p) + p\log(p)}{2}$$

Which tends to 0 as $p \to 0$.

This makes intuitive sense – even if $\{X\}$ genuinely provides new (independent of $\lfloor X \rfloor$) information on Y, we may still be safe to ignore it if $\lfloor X \rfloor$ provides enough information anyway (i.e. if p is close to 0). Crucially, the loss tends to 0 independent of whether $q \to 0$ – it does not matter how much information $\{X\}$ provides on Y, simply that $\lfloor X \rfloor$ provides a large amount of information on Y.

This will be the basic idea behind our main result, which we will prove in Sec 5: if Y is "almost determined" by X_1, then $\mathbb{E}_{x\sim X}\left[\mathrm{H}\left[\mathrm{P}\left(Y \mid X\right), \mathrm{P}\left(Y \mid X_1\right)\right]\right]$ is "almost 0".

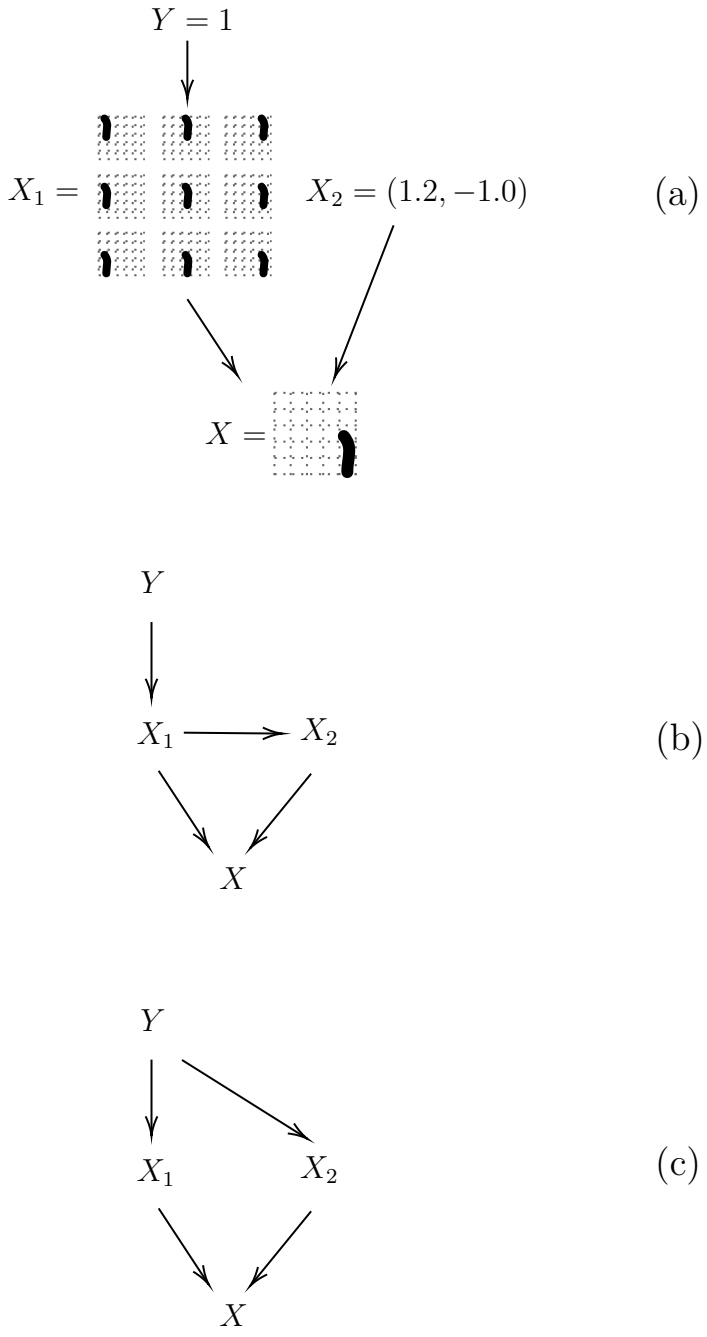


Figure 4.1: Causal models for Ex 4.1 (a), (b), (c), with size-1 sample in (a) for illustration purposes

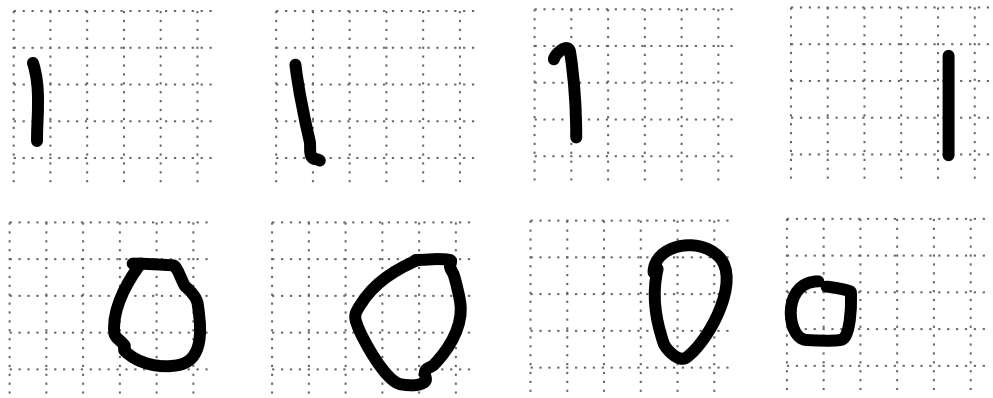


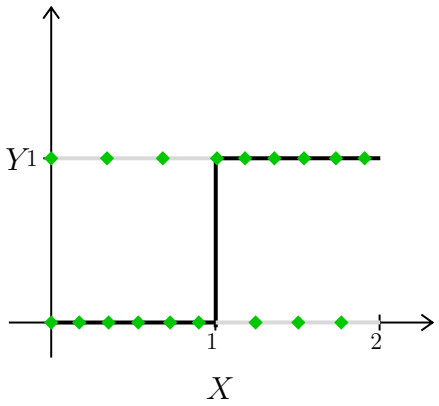
Figure 4.2: Sample from Ex 4.1 (c)

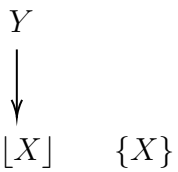
5 Main result

In Def 3.3, we defined what it meant for a representation – a learning algorithm, as thought of as an estimate $Q(y \mid x)$ for the true posterior $\mathrm{P}(Y = y \mid X = x)$. In Def 5.1, we will define what it means for the true posterior $\mathrm{P}(Y = y \mid X = x)$ to be invariant.

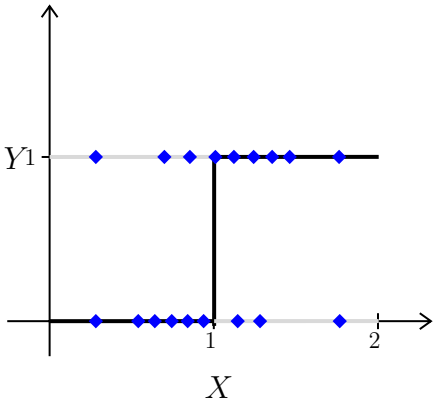
Definition 5.1 (Invariant posterior)**.** Let $\mathcal{X}, \mathcal{Y}$ be sets and X, Y are random quantities taking values in $\mathcal{X}, \mathcal{Y}$; suppose we can factor $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, writing each $X = (X_1, X_2)$. We say that the posterior $\mathrm{P}(Y \mid X)$ is X_2-invariant if $\mathrm{P}(Y \mid X_1, X_2) = \mathrm{P}(Y \mid X_1)$, i.e. $Y \perp\!\!\!\perp X_2 \mid X_1$.

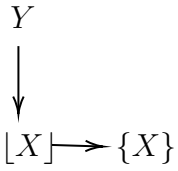
Note that Def 5.1 is similar in spirit to Def 2.4 – both talk about the invariance of a probability distribution, rather than of a representation. However, Def 5.1 speaks of invariance of the conditional distribution $\mathrm{P}(Y \mid X)$ while Def 2.4 speaks of the invariance of the joint distribution $\mathrm{P}(X, Y)$ – the former is more useful for supervised learning applications. We do, however have:



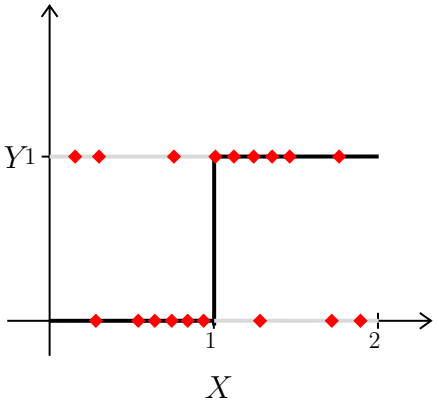


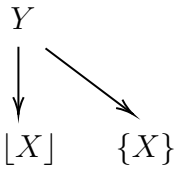
(a) Ex 4.2 (a)





(b) Ex 4.2 (b)





(c) Ex 4.2 (c)

Figure 4.3: Exs 4.2 (a), 4.2 (b), 4.2 (c) – (left) sample of joint distribution (right) causal diagram

Proposition 5.2 (Def 2.4 vs. Def 5.1)**.** *Consider the set-up in Def 2.4; suppose we can factor* $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ *so* $Z = (X, Y)$*, and that the action of* G *can be decomposed onto this factorization* $g(x, y) = (g(x), g(y))$ *with all its stabilizers on* $\mathcal{X}$ *trivial and its action on* $\mathcal{Y}$ *trivial (*id*). Further define* $\mathcal{X}_1 := \mathcal{X}/G$ *(the set of orbits) and* $\mathcal{X}_2 := G$ *– by Lemma 2.12, there is a bijection* $\phi : \mathcal{X}_1 \times \mathcal{X}_2 \simeq \mathcal{X}$*, so we can define the random variables* $(X_1, X_2) = \phi^{-1}(X)$ *taking values in* $\mathcal{X}_1$*,* $\mathcal{X}_2$*.*

If $\mathrm{P}(Y \mid X)$ *is* X_2*-invariant (Def 5.1) and* X *is* G*-invariant (Def 2.4), then* Z *is* G*-invariant (Def 2.4).*

Proof.

$$
\begin{aligned}
&\mathrm{P}(Y \mid X) \text{ is } X_2 - \text{invariant.} \\
&\iff \mathrm{P}(Y \mid X_1, X_2) = \mathrm{P}(Y \mid X_1) \\
&X \text{ is } G - \text{invariant.} \\
&\iff \forall g,\ X =_d gX \\
&\iff \forall x, g,\ \mathrm{P}(X = x) = \mathrm{P}(X = gx) \\
&\iff \forall x, g,\ \mathrm{P}\left((X_1, X_2) = \phi^{-1}(x)\right) = \mathrm{P}\left((X_1, X_2) = \phi^{-1}(gx)\right) \\
&\iff \forall x_1, x_2, x_2',\ \mathrm{P}(X_1 = x_1, X_2 = x_2) = \mathrm{P}(X_1 = x_1, X_2 = x_2') \\
&\iff \mathrm{P}(X_1, X_2) = \mathrm{P}(X_1)
\end{aligned}
$$

Z is $G-$invariant.

$\iff \forall g,\ Z =_d gZ$

$\iff \forall g,\ (X,Y) =_d (gX,Y)$

$\iff \forall x,y,g,\ \mathrm{P}\left(X=x,Y=y\right)=\mathrm{P}\left(X=gx,Y=y\right)$

$\iff \forall x,y,g,\ \mathrm{P}\left((X_1,X_2)=\phi^{-1}(x),Y=y\right)=\mathrm{P}\left((X_1,X_2)=\phi^{-1}(gx),Y=y\right)$

$\iff \forall x_1,x_2,x_2',y,\ \mathrm{P}\left(X_1=x_1,X_2=x_2,Y=y\right)=\mathrm{P}\left(X_1=x_1,X_2=x_2',Y=y\right)$

$\iff \mathrm{P}\left(X_1,X_2,Y\right)=\mathrm{P}\left(X_1,Y\right)$

Suppose the hypotheses. Then:

$\mathrm{P}\left(X_1,X_2,Y\right)=\mathrm{P}\left(Y\mid X_1,X_2\right)\mathrm{P}\left(X_1,X_2\right)=\mathrm{P}\left(Y\mid X_1\right)\mathrm{P}\left(X_1\right)=\mathrm{P}\left(X_1,Y\right).$

(This is the proof for discrete $\mathcal{Z}$ – for the continuous case, one may suitably replace probability mass functions by probability density functions.)

□

(This is in fact a correction to the paper [4] – the paper itself states that Def 2.4 means "the probability of an image being a bird is [...] the same as the probability for a rotated image", which as we discussed in Sec 1, is incorrect – in fact, this is Def 5.1, which is not equivalent. This is the reason we had to introduce Def 5.1, as it is what is relevant for supervised learning applications.)

We are finally ready to state our main result.

Theorem 5.3 (Condition for invariant learning to be suitable)**.** *In Framework 3.1, observe that* $\mathrm{P}\left(Y\mid X\right)$ *minimizes the risk among all distributions,*

and $\mathrm{P}(Y \mid X_1)$ *minimizes the risk among all invariant distributions. We claim that for any* $\varepsilon > 0$, $R(\mathrm{P}(Y \mid X_1)) - R(\mathrm{P}(Y \mid X)) < \varepsilon$ *holds if (not necessarily only if) either:*

- $\mathrm{P}(Y \mid X)$ *is* X_2*-invariant in the sense of Def 5.1, in which case* $R(\mathrm{P}(Y \mid X_1)) - R(\mathrm{P}(Y \mid X)) = 0$.
- $\mathrm{H}[Y \mid X_1] < \varepsilon$.

Proof.

$$
\begin{aligned}
&R(\mathrm{P}(Y \mid X)) \\
&= \mathbb{E}_X\left[\mathrm{H}\left[\mathrm{P}(Y \mid X), \mathrm{P}(Y \mid X)\right]\right] \\
&= \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathrm{P}(Y = y \mid X = x) \log \mathrm{P}(Y = y \mid X = x)\right] \\
&= \sum_{x, y \in \mathcal{X}, \mathcal{Y}} \mathrm{P}(X = x)\, \mathrm{P}(Y = y \mid X = x) \log \mathrm{P}(Y = y \mid X = x) \\
&= \sum_{x, y \in \mathcal{X}, \mathcal{Y}} \mathrm{P}(X = x, Y = y) \log \mathrm{P}(Y = y \mid X = x) \\
&= \mathrm{H}[Y \mid X] \\
&R(\mathrm{P}(Y \mid X_1) \\
&= \mathbb{E}_X\left[\mathrm{H}\left[\mathrm{P}(Y \mid X), \mathrm{P}(Y \mid X_1)\right]\right] \\
&= \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathrm{P}(Y = y \mid X = x) \log \mathrm{P}(Y = y \mid X_1 = x_1)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x,y \in \mathcal{X},\mathcal{Y}} \mathrm{P}\left(X = x\right) \mathrm{P}\left(Y = y \mid X = x\right) \log \mathrm{P}\left(Y = y \mid X_1 = x_1\right) \\
&= \sum_{x,y \in \mathcal{X},\mathcal{Y}} \mathrm{P}\left(X = x, Y = y\right) \log \mathrm{P}\left(Y = y \mid X_1 = x_1\right) \\
&= \sum_{x_1,x_2,y \in \mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}} \mathrm{P}\left(X_2 = x_2 \mid X_1 = x_1, Y = y\right) \mathrm{P}\left(X_1 = x_1, Y = y\right) \log \mathrm{P}\left(Y = y \mid X_1 = x_1\right) \\
&= \sum_{x_1,y \in \mathcal{X}_1,\mathcal{Y}} \left[\mathrm{P}\left(X_1 = x_1, Y = y\right) \log \mathrm{P}\left(Y = y \mid X_1 = x_1\right) \sum_{x_2 \in \mathcal{X}_2} \mathrm{P}\left(X_2 = x_2 \mid X_1 = x_1, Y = y\right) \right] \\
&= \sum_{x_1,y \in \mathcal{X}_1,\mathcal{Y}} \left[\mathrm{P}\left(X_1 = x_1, Y = y\right) \log \mathrm{P}\left(Y = y \mid X_1 = x_1\right) \cdot 1 \right] \\
&= \mathrm{H}\left[Y \mid X_1\right]
\end{aligned}
$$

If we have invariant posterior, then simply $\mathrm{P}\left(Y \mid X_1\right) = \mathrm{P}\left(Y \mid X\right)$. If $\mathrm{H}\left[Y \mid X_1\right] < \varepsilon$, observe that by Lemma 2.11, $\mathrm{H}\left[Y \mid X\right] = \mathrm{H}\left[Y \mid X_1, X_2\right] \leq \mathrm{H}\left[Y \mid X_1\right] < \varepsilon$. Thus $R(\mathrm{P}\left(Y \mid X_1\right)) - R(\mathrm{P}\left(Y \mid X\right)) = \mathrm{H}\left[Y \mid X_1\right] - \mathrm{H}\left[Y \mid X\right] < \varepsilon$. □

Thus in these situations, the "best invariant distribution" does no more than ε worse than the "best distribution" – our previous observation about invariance still making sense in the absence of an invariant posterior is therefore reduced to an information-theoretic result. The following corollary is a more direct answer to the question "When is invariant learning effective?".

Corollary 5.4. *Let Q and Q' be the minimum-loss estimator and the minimum-loss invariant estimator for $P(Y \mid X)$ respectively. Then for any $\varepsilon > 0$, $\mathbb{E}\left[L(Q') - L(Q)\right] < \varepsilon$ if (not necessarily only if) either:*

- $\mathrm{P}\,(Y \mid X)$ *is* X_2*-invariant in the sense of Def 5.1, in which case* $R(\mathrm{P}\,(Y \mid X_1)) - R(\mathrm{P}\,(Y \mid X)) = 0$.
- $\mathrm{H}\,[Y \mid X_1] < \varepsilon$.

6 Conclusion

We began our investigation with the following observation: even when some "nuisance factor" X_2 provides non-redundant information about the label Y in a supervised learning application, it is sometimes safe to ignore this factor (i.e. use a learning algorithm that was invariant to this factor), as it was in Fig 1.1. Certainly, the loss would be greater than if the factor was incorporated – but still low, so we could hope for a bound on it. Intuitively, we imagined that the reason that we could do so in Fig 1.1 was that the other factors X_1, namely the "shape" provides "sufficient information" on Y, so in some sense the information propagated from X_1 to Y "dominates" that propagated from X_2 to Y. Crucially, we saw in our analysis of Ex 4.2 (c) that this domination happens *independently* of the amount of information provided by X_2 on Y.

Our main result, Thm 5.3 formalizes this observation as an information theoretic result: specifically, we demonstrate that the expected loss $R(Q)$ for the distribution $P(Y \mid X)$ is precisely the conditional entropy $H(Y \mid X)$, and for the invariant distribution $P(Y \mid X_1)$ is precisely the conditional entropy $H(Y \mid X_1)$, which is a measure of the uncertainty that remains in Y after knowing X_1.

This result can be interpreted in the light of our motivating questions.

For applications to fair machine learning, Thm 5.3 says that the conditional mutual information $I(Y; X_2 \mid X_1) = H(Y \mid X_1) - H(Y \mid X_1, X_2)$

becomes precisely the bound on the increased risk from learning invariant to sensitive characteristics X_2.

Example 6.1 (6 and 9 problem in Framework 3.1)**.** Let $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} = \{x : \mathbb{R}^2 \to \{0, 1\}\}$ be the set of images; define $\mathcal{X}_2 = \{T_U : \mathcal{X} \to \mathcal{X} : x \mapsto \lambda i.x(U^{-1}i) \mid U \in SO(2)\} \cong SO(2)$ to be the group of rotations acting on $\mathcal{X}$; define $\mathcal{X}_1 = \mathcal{X}/\mathcal{X}_2$ to be the quotient by the group action (i.e. the orbits under rotation). By Lemma 2.12, there is a bijection $\phi : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{X}$ given by $\phi(O, T_v) = T_v\xi(O)$ where $\xi(O)$ is some representative in O; then $\phi^{-1} : x \mapsto ([x], T_{\theta(x)})$ where $T_{\theta(x)} \in \mathcal{X}_2$ is such that $T_{\theta(x)}\xi([x]) = x$, which exists as x is in the same orbit as $\xi([x])$. Composing this with the isomorphism $\mathcal{X}_2 \cong SO(2)$ (given by $T_U \mapsto U$), we have $x \mapsto ([x], \theta(x))$. Thus the decomposition $\mathcal{X}_1 \times \mathcal{X}_2$ decomposes an image into its orbit under rotations (its "shape") and its angle from some representative element in its orbit (its "angle").

Ex 6.1 formulates the 6 and 9 problem in the language of Framework 3.1 – in it, the posterior is not invariant: the angle $\theta(x)$ provides information on the label even knowing $[x]$, because a 6 and a 9 may be contained in the same orbit. And here invariant learning is not effective – but we don't expect it to be anyway, because $H(Y \mid X_1)$ is large; we still have very little certainty on what the label is after knowing the orbit, leaving "room" for X_2 to take away this uncertainty.

Our work is only some preliminary set-up for serious research into the problem of invariant learning in scenarios without invariant posteriors, which as we discussed in Sec 1 has many important implications.

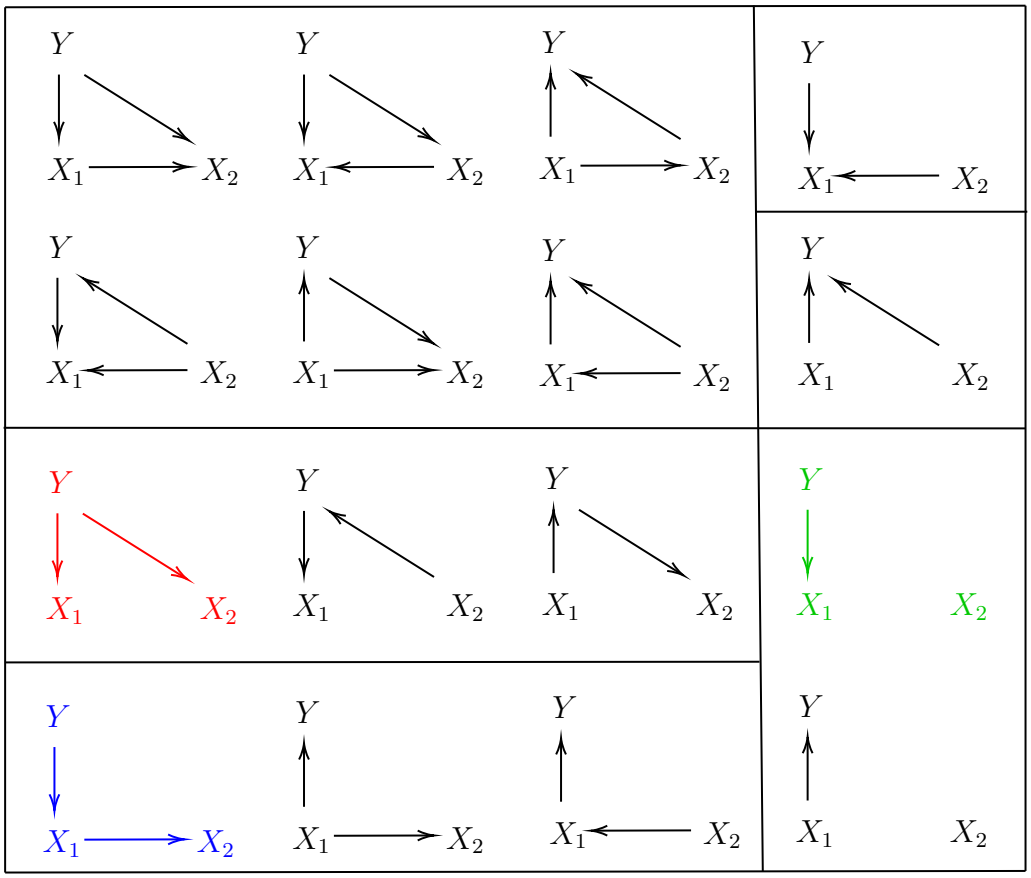


Figure 6.1: All possible causal diagrams that contain an edge between Y and X_1; the boxes are Markov equivalence classes; the causal structures for Exs 4.2 (a), 4.2 (b), 4.2 (c) are in green, blue and red respectively.

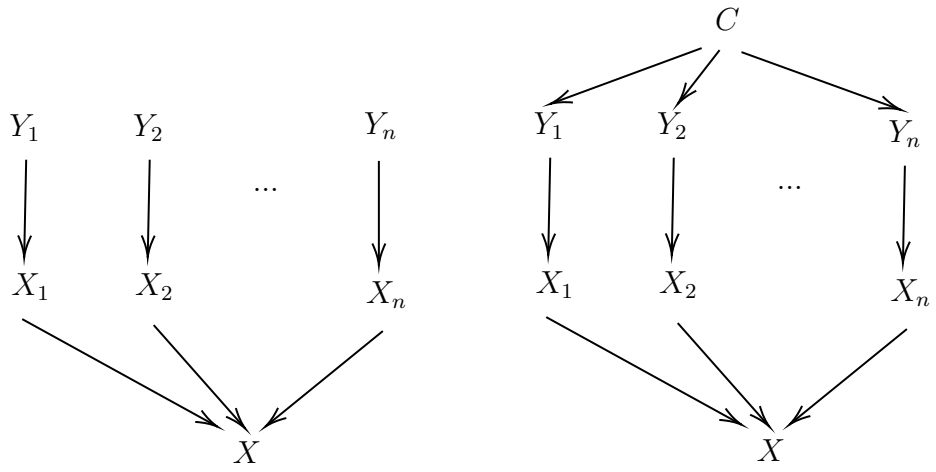


Figure 6.2: (left) Equivariant posterior (right) Non-equivariant posterior due to confounding variable, i.e. common cause introducing correlation

We now outline several directions for further research in this area.

1. Our result is specifically for the cross-entropy loss – while this is a natural loss function to take (especially for our motivating examples), and it is a rather beautiful result that the condition on this loss being bounded is also an information-theoretic one, we may be interested in other loss functions, and these would be bounded by corresponding measures of conditional uncertainty other than the conditional entropy.

   For example, I would conjecture that the corresponding bound for the mean-squared-error loss (for continuous Y) would be the conditional variance $\mathrm{Var}(Y \mid X_1)$. It would be of interest to see a general result, perhaps in the light of empirical risk minimization theory.

2. Our work stands in analogy to the results in [4], in the sense that the

two sufficient cases in Thm 5.3 are analogous to the results in [4] for "exact invariance" and "approximate invariance"; however, this is a rather vague analogy, in that our second case is unrelated to approximate invariance (while our first case is related to exact invariance by Prop 5.2).

More generally, Thm 5.3 only provides *sufficient* conditions for invariant learning to be suitable. The converse problem is significantly harder, as the loss from invariant learning may also be bounded if $P(Y \mid X_1)$ is "approximately equal" in distribution to $P(Y \mid X)$. In [4], "approximately equal in distribution" is defined in terms of the Wasserstein metric.

3. In Sec 1, we briefly hinted that invariant learning is thought to be suitable precisely in the anti-causal direction. While the second sufficient case in Thm 5.3 does not make any direct reference to the causal structure, this is relevant to the converse problem. Fig 6.1 is an exhaustive list of all the relevant possible causal structures between variables Y, X_1, X_2 – we ought to formally investigate and justify which causal structures is invariant learning rational for.

   More generally, one may consider similar problems for generalized learning algorithms on arbitrary causal diagrams – this would be relevant to generalizing our work to *equivariant* learning. We have not provided a formal definition of equivariance beyond the basic definition in Def 2.1;

however, I would suggest that the causal diagram for an "equivariant posterior" would look like Fig 6.2 (left).

4. While Thm 5.3 provides conditions for invariant learning to be no worse than (or rather "no more than ε worse than") non-invariant learning, it does not give reasons as to why one *should* adopt invariant learning in the first place – what makes it *better*, even in the second sufficient case where the loss is in fact greater. This is in contrast to the work of [4], which demonstrates that an estimator learned through invariant learning has lower variance, by providing a decomposition of the covariance of the non-invariantly learned estimator into the covariance of the invariantly learned estimator and another positive-definite matrix.

   Our representation $Q(y \mid x)$ is a random function, so we conjecture that the *covariance kernel* of the non-invariantly learned $Q(y \mid x)$ into the covariance kernel of the invariantly learned $Q(y \mid x)$ and another positive-definite kernel. Studying the covariance kernel of $Q(y \mid x)$ would also help us prove Corollary 5.4, perhaps through an application of the multidimensional Chebyshev inequality.

5. In contrast to algorithms that learn invariantly to a given set of symmetries, there are various algorithms in the literature that aim to learn the group of symmetries from the dataset itself, e.g. [7, 21, 22]. This is a goal rather similar to ours – from our perspective, a group G is a subgroup of the symmetry group for a task if $H(Y \mid X_1) - H(Y \mid X)$

is small (where X_1 are the G-orbits), and we can estimate $H(Y \mid X_1)$ from the data[1].

We may then ask two relevant questions thereof: (1) Would an algorithm like in [7, 21, 22] discover an invariance in the second sufficient case, i.e. when the posterior is not truly invariant? (2) Can we use this to produce an algorithm that *learns* the symmetries of a task – rather than simply check it? – i.e. an algorithm to find a factorization of $\mathcal{X}$ that minimizes $H(Y \mid X_1)$.

7 Acknowledgements

I would like to thank my supervisor Prof Jeroen Lamb and PhD student Victoria Klein for their valuable comments on earlier drafts.

And I would like to thank Rev Srikesh Maadhapuzhi Swaminathan Iyer Hegde – for *nothing*. For all the pressure I've put on him, what I've gotten in return? Nothing! Doesn't even rise to greet me.

[1]Estimating entropy from data is not, in general, a simple task, but there are various estimators available, e.g. [23–25]

References

[1] Taco S. Cohen and Max Welling. *Group Equivariant Convolutional Networks*. 2016. arXiv: 1602.07576 [cs.LG].

[2] Risi Kondor and Shubhendu Trivedi. *On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups*. 2018. arXiv: 1802.03690 [stat.ML].

[3] Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. *Selecting Data Augmentation for Simulating Interventions*. 2020. arXiv: 2005.01856 [stat.ML].

[4] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. *A Group-Theoretic Framework for Data Augmentation*. 2020. arXiv: 1907.10905 [stat.ML].

[5] Manik Varma and Debajyoti Ray. "Learning The Discriminative Power-Invariance Trade-Off". In: *2007 IEEE 11th International Conference on Computer Vision*. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4408875.

[6] Julia Ling, Reese Jones, and Jeremy Templeton. "Machine learning strategies for systems with invariance properties". In: *Journal of Computational Physics* 318 (2016), pp. 22–35. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2016.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0021999116301309.

[7] Gregory Benton et al. *Learning Invariances in Neural Networks*. 2020. DOI: 10.48550/ARXIV.2010.11882. URL: https://arxiv.org/abs/2010.11882.

[8] Francesco Locatello et al. *On the Fairness of Disentangled Representations*. 2019. arXiv: 1905.13662 [cs.LG].

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[10] J. Peters, D. Janzing, and B. Scholkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. MIT Press, 2017. ISBN: 9780262037310. URL: https://books.google.co.uk/books?id=XPpFDwAAQBAJ.

[11] Taco S. Cohen, Mario Geiger, and Maurice Weiler. *Intertwiners between Induced Representations (with Applications to the Theory of Equivariant Neural Networks)*. 2018. arXiv: 1803.10743 [cs.LG].

[12] Carlos Esteves. *Theoretical Aspects of Group Equivariant Neural Networks*. 2020. arXiv: 2004.05154 [cs.LG].

[13] Jan E. Gerken et al. *Geometric Deep Learning and Equivariant Neural Networks*. 2021. arXiv: 2105.13926 [cs.LG].

[14] Marco Reisert and Hans Burkhardt. "Learning Equivariant Functions with Matrix Valued Kernels". In: *Journal of Machine Learning Re-*

search 8.15 (2007), pp. 385–408. URL: http://jmlr.org/papers/v8/reisert07a.html.

[15] Alessandro Achille and Stefano Soatto. *Emergence of Invariance and Disentanglement in Deep Representations*. 2018. arXiv: 1706.01350 [cs.LG].

[16] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Penguin Books Limited, 2018. ISBN: 9780241242643. URL: https://books.google.co.uk/books?id=EmY8DwAAQBAJ.

[17] Bernhard Schölkopf et al. *Towards Causal Representation Learning*. 2021. arXiv: 2102.11107 [cs.LG].

[18] Hyunjik Kim and Andriy Mnih. *Disentangling by Factorising*. 2019. arXiv: 1802.05983 [stat.ML].

[19] Irina Higgins et al. *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*. 5th International Conference on Learning Representations. 2017.

[20] Xiao Liu et al. *Learning Disentangled Representations in the Imaging Domain*. 2021. arXiv: 2108.12043 [cs.CV].

[21] Ekin D. Cubuk et al. *AutoAugment: Learning Augmentation Policies from Data*. 2018. DOI: 10.48550/ARXIV.1805.09501. URL: https://arxiv.org/abs/1805.09501.

[22] Mark van der Wilk et al. *Learning Invariances using the Marginal Likelihood*. 2018. DOI: 10.48550/ARXIV.1808.05563. URL: https://arxiv.org/abs/1808.05563.

[23] Liam Paninski. "Estimation of Entropy and Mutual Information". In: *Neural Comput.* 15.6 (June 2003), pp. 1191–1253. ISSN: 0899-7667. DOI: 10.1162/089976603321780272. URL: https://doi.org/10.1162/089976603321780272.

[24] Paul Valiant and Gregory Valiant. "Estimating the Unseen: Improved Estimators for Entropy and other Properties". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/53c04118df112c13a8c34b38343b9c10-Paper.pdf.

[25] Evan Archer, Il Memming Park, and Jonathan Pillow. *Bayesian Entropy Estimation for Countable Discrete Distributions*. 2013. DOI: 10.48550/ARXIV.1302.0328. URL: https://arxiv.org/abs/1302.0328.