Some Fairly-Incomprehensive Background on Some Fairly-Related Things

Contents

1 Preamble

In this chapter we aim to provide background on the topics encompassed by the works in this thesis, both individually, and holistically. While we will on occasion point to exemplar methods, generally we will eschew delving deep into specific methodologies – of which there are many and many more being proposed by the day – in favour of keeping broader perspectives regarding the motivations of, assumptions made by, and interconnections between, the considered learning paradigms. This is to say, this chapter does not aspire to be a comprehensive survey of Domain Adaptation (DA), Domain Generalisation (DG), Fair Machine Learning (FairML), and the other germane subfields touched on herein; producing such for any one of these subfields is in itself a considerable undertaking given the breadth and depth the Machine Learning literature, the field having grown precipitously over the last decade since the onset of the deep-learning revolution heralded by Krizhevsky et al. [2012]. The aspirations of this chapter, on the contrary, are much more humble, simply being to provide the requisite (high-level) background for, and unified and alternative perspectives of, the problems and methodologies featured in Chapters 3, 4 and 5. Indeed, each of said chapters contain their own background sections drawing direct comparisons to related work and we wish to avoid repetition in this respect.

The main themes of this thesis, as indicated by the title, are Semi-Supervised Learning (SemiSL) and Distributional Robustness (DR), the latter in the context of FairML (Chapters 3 and 4) and DG (Chapter 5), specifically. We will cover these topics directly, but to properly contextualise and motivate them requires visiting both foundational and adjacent areas of ML.

With the above in mind, we begin our discussion of the classical supervised learning setup and how standard empirical-risk minimisation (ERM) is ill-suited to long-tailedness and distribution shifts, both pervasive phenomena in real-world applications. *Distributions shift* as a term is highly polysemous, meaning very different things, and demanding commensurately different solutions, depending on the underlying mechanisms and the direction of causality. We will give a brief taxonomy of the different kinds of distribution shifts in terms of how the marginal and conditional distributions are affected, and what may cause them.

Spurious correlations (SCs), or (statistical) shortcuts (we will use the terms interchangeably throughout), give rise to a particularly aggressive form of distribution shift as a result of features in the training data being highly (conceivably to the degree of a one-to-one correspondence) correlated with the target but not in a way that is causally consistent, and thus in a way that should not be expected to hold consistently at test time. The idea of SCs is central to both Chapters 3 and 4 (manifested in different ways) and the idea of Shortcut Learning (SCL) has close ties to DG [Arjovsky et al., 2019], the focus of Chapter 5; in light of this, we afford dedicated discussion of the SCL problem and what conditions are needed to engender it.

Before branching off into discussion of the specific learning paradigms, we first reframe the problem of supervised learning and distribution shift through the lens of causality, yielding a unified perspective of the latter and when SemiSL (and by extension *Self-Supervised Learning* (SelfSL)) should be expected to work. While this thesis does not directly tangle with questions of causal inference, the field of causality [Pearl, 2009] provides, through Causal Bayesian graphs (CBGs) and interventions thereof, the means of expressing different distribution shifts and the desired/undesired variances/invariances using a single, formalised calculus. Equipped with this calculus, we conclude this section – as alluded to above – with a discussion of the specific subfields and learning paradigms of relevance to the three papers constituting this thesis, namely: DA, DG, FairML, SemiSL, SelfSL, Adversarial Learning (AL), and Normalising Flows/Invertible Neural Networks (NFs/INNs).

2 Some notes on notation

We describe here some of the general notation schemes used throughout this background chapter, leave the concrete notation to be defined contextually, both to allow overloading (to allow for reuse and restrictedness of the alphabet) and to minimise cognitive overhead for the reader.

First, we denote random variables using upper-case (non-calligraphic) letters and their associated observed/deterministic/realised variables with the corresponding lower-case letters. Following convention, we consistently denote by X and Y the input (covariate) and target (response) variables, respectively; by S some auxiliary variable on which we want to condition (for evaluation and/or optimisation), such as the domain (in domain adaptation/generalisation) or sensitive attribute (in algorithmic fairness); by Z the latent space, representations, encodings, or embeddings (all synonymously) of some model. Second, calligraphic letters are used to denote (but not exclusively) the domain of a variable, e.g. $x \in \mathcal{X}$. Under this scheme, we would have for the random variable, $X : \Omega \rightarrow \mathbb{R}^d$, realisations $x \in \mathcal{X} \subset \mathbb{R}^d$ defined on a subset of the d-dimensional space of real numbers. We then use $P(\cdot)$ to denote probability distributions with conditioning indicated as $P(X = x)$ – continuing the foregoing example – and use $\mathcal{D}$ to denote *datasets* that correspond to the empirical distributions of variables; for instance $\mathcal{D} \triangleq \{x_i\}_{i=1}^N$ denotes a dataset made up of N observations of X. We will often augment this notation with super- and subscripts to indicate a variety of concepts including, inter alia, association with a particular subset of the data or concept, optimality, observability, and approximation. Some representative examples include $\mathcal{D}^{tr}$ and $\mathcal{D}^{te}$ to denote the training and test sets, respectively, f^* to denote the optimal function w.r.t. some optimisation problem, and $\hat{y}$ to denote a prediction made by some estimator (of $P(Y|X)$).

Finally, to simplify exposition, we abuse notation by allowing functions of the form $f : X \rightarrow Y$ to accept random and observed variables interchangeably; we assume that the derived function classes are Borel Measurable and as such that a function of a random variable is also a random variable. f to operate on random variables X. Thus, pedantically speaking, $f(X)$ should be read as shorthand for $f \circ X(\omega)$, for some event ω drawn from sample space, Ω, while $f(x)$

should be read in the standard fashion, with deterministic inputs and outputs.

3 Supervised learning, empirical risk minimisation, and its pitfalls

Traditional learning algorithms usually assume that (or are only optimal when) the training and test samples are *both* variables identically-and-independently distributed (i.i.d.) random variables, such that one has $P^{tr}(X, Y) \approx P^{te}(X, Y)$. Here, $P^{tr}(X, Y)$ and $P^{te}(X, Y)$ denote the (joint) training and test distributions, respectively. Based on this assumption, the method of Risk Minimisation seeks the hypothesis $f^* \in \mathcal{F}$ that is the minimiser of the *risk*, $\mathcal{R}$, defined according to some statistical distance, of *loss*, $\mathcal{L} : \mathbb{R}^{\Omega} \times \mathcal{Y}^{\Omega} \to \mathbb{R}$ between the predicted, $\hat{Y} \triangleq f(X)$, and ground-truth labels, Y over the training distribution, $P^{tr}(X, Y)$. A canonical example of such a distance for classifications tasks is the *cross-entropy loss*; in information-theoretic terms, this can be couched as the amount of information (here, in nats) required to identify a sample from the true distribution given a coding scheme optimised for the predictive distribution and takes the form

$$H(Y, f(X)) \triangleq \mathbb{E}_{P(Y)}[\log P(f(X))]. \tag{1}$$

An alternative, and perhaps more natural, way of viewing this function is by its decomposition into the sum of the Kullback-Leibler (KL) Divergence (also known as the *relative entropy*), $D_{KL}(P(f(X))||P(Y))$, and entropy of the marginal target distribution, $H(Y)$. When the latter carries no dependence on the learned parameters – as is generally the case, save for certain cases of model-distillation, consistency-regularisation etc. – the term vanishes from the gradient, leaving just the KL term. Returning from this brief aside, we can formally define the (population or true) risk as

$$\mathcal{R}(f) \triangleq \mathbb{E}_{(X,Y) \sim P^{tr}(X,Y)}[D(f(X, Y))] \tag{2}$$

In practice, of course, one does not have access to the true generative distribution, but only a finite set of realizations of it that together form a *dataset*, $\mathcal{D}^{tr}$, consisting of observed input-target pairs (x, y). Thus, we are instead doing *empirical* risk minimisation (ERM; Vapnik [1991]), defined as the risk is instead defined over the empirical distribution, a finite set of observations, rather than over the underlying distribution from which those observations were drawn. To accommodate this discrepancy, two things need to be accounted for. **First** we need to substitute $P^{tr}(X, Y)$ with its empirical counterpart, $\mathcal{D}^{tr} \triangleq \{(x_i, y_i)\}_{i=1}^{N^{tr}}$; since we are now operating over a finite set of N^{tr} tuples, the expectation can be replaced with a finite sum (with normalisation). **Second**, given the variables are deterministic rather than random, one can no longer frame the optimisation objective in terms of a statistical distance explicitly. Instead one measures – and uses as feedback to drive model-optimisation – the discrepancy between the predicted and observed targets using an empirical *loss* function, $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. In standard classification settings in which the targets are given by single (one-hot encoded) labels, Y is represented by a degenerate (delta) distribution, wherein for each instance we simply have an binary indicator of which class said instance belongs to, rather than a distribution over the probability simplex, allowing for capturing of intrinsic uncertainty in the task. With the foregoing adjustments in mind, one can then define the standard ERM objective as

$$\mathcal{R}(\hat{f})) \triangleq \frac{1}{|\mathcal{D}^{tr}|} \sum_{(x,y)\in\mathcal{D}^{tr}} \ell(f(x), y).$$

This version of the objective is simply a uniform (unweighted) average over all training pairs; it does not take into account the distribution of the inputs or targets. This is mentionable as many real-world datasets exhibit significant class class-imbalance [Zhu et al., 2014, Van Horn and Perona, 2017], or, more generally, 'long-tailedness', which is to say that the marginal distribution $P(Y)$ is not uniform over its support. Such motivates replacing the unweighted (or, more accurately, 'uniformly weighted') objective given by Eq. 3, with an importance *weighted* variant wherein the loss is weighted by $P^{tr}(Y)^{-1}$, or, in the empirical case, by the inverse frequencies of the targets, in the discrete (classification) case, or by the empirical density of the target (as given by kernel density estimate (KDE), for instance) in the continuous (regression) case. Here, we have assumed no foreknowledge of $P^{te}(Y)$ – this typically being the case in practice – with the choice of an uninformative, uniform distribution over the domain leading to its elimination from the importance weighting term that in general takes the form $\frac{P^{te}(Y)}{P^{tr}(Y)}$ (or the empirical equivalent).

Using $w \in \mathbb{R}^{+}$ to denote the weight assigned to instance x in $\mathcal{D}^{tr}$, we can then generalise Eq. 3 to

$$\hat{\mathcal{R}}(f) \triangleq \sum_{(x,y)\in\mathcal{D}^{tr}} w \cdot \ell(f(x), y),$$

with '$\cdot$' denoting regular scalar multiplication (over $\mathcal{R}$), noting that this form subsumes the unweighted form, which itself can be recovered by simply fixing w to $|\mathcal{D}^{tr}|^{-1}$ for all instances. It is also worth noting that these weights can be adaptive; they can be iteratively adjusted over the course of training according to some parametric or non-parametric function [Wang et al., 2021]. Instead of weighting the instance-losses, one can instead use the weights to adjust the sampling, which has several practical advantages when training with stochastic gradient descent (SGD), particularly: 1) The procedure is non-invasive: no modification to the data-loading nor the computation of the loss is required; 2) Highly-weighted samples appear in batches commensurately often; when weighting the loss, samples belonging to the long-tail will appear in batches rarely, resulting in forgetting and poor diversity as said samples are effectively duplicated. One can achieve a similar effect by under-sampling the majority classes, groups, or their intersections, such that they are equifrequent, and i.i.d. sampling from that subset $\mathcal{D}^{tr}_{\text{US}} \subset \mathcal{D}^{tr}$ (or, conversely by duplicating instances from the minority classes to the same end). Under- and over-sampling (US and OS, respectively) have long been used as a remedy for class imbalance [Chawla et al., 2002] but the former has recently been shown to be effective – matching or exceeding in performance more sophisticated algorithms – for group robustness and spurious-correlation problems [Sagawa et al., 2020, Idrissi et al., 2022] in part due to its early-stopping effect.

Despite its intuitiveness and long history, with roots in early statistical modelling, the practical usefulness of importance-weighted ERM in the context of modern deep learning has recently been impugned [Byrd and Lipton, 2019, Zhai et al., 2022]. Byrd and Lipton [2019] demonstrate that for *over-parametrised* models the effects of importance-weighting diminish over the course of training; these effects can be partially recovered when used in conjunction regularisation such as dropout, early-stopping and standard L_2 weight decay but without such interventions the converged-upon solution is identical for both IW-ERM and vanilla ERM. Evidence for this was also provided by Sagawa et al. [2019], who stress the importance of combining aggressive regularisation with (a dynamic form of) importance-weighting for strong worst-group generalisation. These empirical results were recently supported theoretically by Zhai et al. [2022], who prove that the implicit biases of these algorithms and standard ERM are indeed equivalent. Summarily, while importance-weighting may be intuitive, it provably does not alter the solution to the optimisation problem defined by the training set, which is to say, solutions that attain zero-loss are invariant under reweighting. This understanding has motivated other approaches, such as those based on polynomially-tailed losses (for binary classification; Wang et al. [2021]) and logit-adjustment [Menon et al., 2020].

As statistical models are only required model correlations in the data to satisfy the loss function, they ultimately only capture a superficial representation of the true physical processes involved. In the discriminative case (that this thesis is concerned with), for a given X and Y we are interested in approximating the conditional distribution $P(Y|X)$; this corresponds to tasks like predicting the probability that a given image contains a dog (image classification), or the probability that a given chest X-ray indicates a pulmonary infiltration, or some other thoracic condition. Indeed, the task of accurately estimating $P(Y|X)$ can be provably solved by observing a sufficient amount of i.i.d. data drawn from the joint distribution $P(X, Y)$, yet this only solves the problem from the aforementioned statistical perspective, and we will see that this perspective is not always aligned with the causal one, which can lead to problems in generalisation under certain conditions that crop up disconcertingly often in real-world applications, including those that safety-critical. This is to say, the predictions of a statistical model should only be trusted when the conditions of the training and test distributions are sufficiently similar, and, in short, arbitrary shifts (interventions on the data-generating distribution) can give rise to arbitrarily bad predictions [Pearl, 2009, Schölkopf et al., 2012].

Since the true causal relationships between independent and dependent variables is, generally, not *identifiable* given the training data alone, owing to confounding variables, additional information, as provided by interventions, or *envi-*

ronments [Peters et al., 2016], is needed to resolve the statistical ambiguity; this the tack is popular within the domain generalisation literature, wherein domain can be viewed as a different intervention on the true distribution. By 'confounding variable', or *confounder*, we mean some variable that is that is the causal parent of two or more other variables and explains the statistical dependency between them despite those variables not being causally related themselves; in the trivariate case this corresponds to the fork $X \leftarrow S \rightarrow Y$, wherein there exists a spurious (acausal) correlation between X and Y that is eliminated by conditioning on the confounder S. In the shortcut learning problems addressed in Chapters 3 and 4, we will see statistical learning breakdown in a similar way yet for essentially the opposite reason. Namely, instead of having latent variable that explains the statistical dependency X between Y in the absence of a causal dependency, we instead of have some spurious variable, S on which Y is strongly statistically, but not causally, dependent, with $X \rightarrow Y$ assumed to be the true causal mechanism. We will delve more deeply into what Shortcut Learning is and how the mechanisms that give rise to it in §5. For now, however, we will move onto discussing different types of distribution shift that statistical learning has to contend with.

4 A (brief) taxonomy of distribution shifts

In this section, we provide a brief taxonomy of the types of distribution shift that arise in the statistical-learning literature and discuss how and in what contexts they might practically emerge. To this end, we draw heavily upon the works of Moreno-Torres et al. [2012] and Castro et al. [2020] in our definitions, noting that the ML literature is not of a single mind regarding the terminology and its semantics. In §6 we will reframe these distribution shifts in causal terms by introduction of an exogenous variable – allowing for an elegant formulation of the distribution shift problem and its relation to invariance – but we leave that aside for now and seek to present them in more general terms.

4.1 Covariate shift

The most well-studied of the shifts, simply put, covariate shift refers to a change in the marginal distribution of the inputs, that is to say we have $P^{tr}(X) \not\approx P^{te}(X)$ while the conditional distribution remains (effectively) unchanged, i.e. $P^{tr}(Y|X) \approx P^{te}(Y|X)$. Departing from Moreno-Torres et al. [2012], we do not restrict its definition to problems of a causal ($X \rightarrow Y$) nature and do away with the distinction between covariate shift and its anticausal ($Y \rightarrow X$) analogue in *prior shift* to simplify exposition. Changes in the distribution of the target variable, Y, will be referred to as *target shift*, as explained in the subsection below. This is not to say that we disregard the importance of distinguishing between the two causal directions; in the context of DA and SemiSL, we will discuss at some length the dependence of these paradigms on this characteristic of the problem. Indeed, a common assumption in DA is that the source and target domains are separated by covariate shift [David et al., 2010], however this assumption breaks down when the problem is anticausal (when we have what Moreno-Torres et al. [2012] term prior shift).

4.2 Target shift

Diametric to the above, target shift describes, as the name suggests, a shift in the marginal distribution of the targets, Y, i.e. $P^{tr}(Y) \not\approx P^{te}(Y)$. In the classification setting, this means that classes do not appear equifrequently in the training and test data; many real-world datasets used for training exhibit long tails, w.r.t. the classes (or targets generally), in which the most-frequent class can appear orders-of-magnitude more frequently than the least-frequent class, while the test data has more even coverage. As discussed in §3, a classic approach to rectifying this kind of shift, in the case of the discriminative models we are usually concerned with, is to importance-weight the instance losses or, near-equivalently, the sampling mechanism, using the ratio $\frac{P^{te}(Y)}{P^{tr}(Y)}$, or simply by the denominator should $P^{te}(Y)$ not be reliably estimable (as is often the case).

4.3 Concept shift

To complete the triad of bivariate distribution shifts (we will later revisit distribution shift under the influence of an exogenous domain or environment variable) we have *concept shift*, referring to changes in the conditional distributions, $P(Y|X)$ or $P(X|Y)$, while the respective ($P(X)$ and $P(Y)$) marginal distributions are preserved. Thus, in the classification setting, concept shift corresponds to a change in the mechanism used to annotate the data; this might entail, for example, changes in the class definitions, differences in annotation protocol or grading scales between sites, or different proclivities/standards in the annotators in the case of human-driven annotation should the task possess an element of subjectivity (AI-alignment via RLHF [Bai et al., 2022] being a prime and topical example of such a task). In addition to the shifts discussed, one can naturally also consider their composition, giving rise to *compound shifts*, in which both the marginal and conditional distributions are subject to change. Such shifts, however. are unusual in the literature, and, perhaps more pertinently, impossible to solve unless one can draw upon strict assumptions, due to the need to decouple the constituent shifts (a problem of identifiability).

4.4 Sampling bias

Sampling (which we use synonymously with *selection* and *representational*) bias refers to distribution shifts that arise due to systematic flaws in the data-collection process that cause training samples to be selected in a non-uniform fashion from the general population being modelled. That is to say, the data is not missing at random but rather conditionally, and most notably when the conditioning is on the target or some other characteristic, such as a particular demographic. Thus, sampling bias is not a type of distribution shift in itself, but rather a mechanism by which the above-described distribution shifts can emerge, and it is particularly germane to Chapters 3 and 4 of this thesis in which we consider extreme cases of it in which certain demographics, or outcomes for certain demographics, are omitted from the training data, promoting spurious correlations between said demographics and the outcome. To give an example, in conducting a local survey there will invariably be subsets of the general population which are under-represented, or altogether excluded, from data-collection due to availability, willingness, and applicability to the research being conducted; if the locale in question were a university, then we would expect the population to be significantly younger and more liberal than on average. Indeed, this a problem is particularly well-noted in experimental psychology, in which cohorts overwhelmingly consist of a very narrow band of individuals from the so-called WEIRD (White, Educated, Industrialized, Rich, Democratic)[Henrich et al., 2010] group. The experimental data obtained from such homogeneous cohorts has been used by numerous high-profile journal papers to support broad claims about the general population, despite obvious issues with its representativity.

A prominent yet more subtle, mechanistically, example of sampling bias can be found in the credit scoring literature, in which no feedback is obtained from previously rejected candidates; this leads to bias amplification (as the model's past decisions directly shape the training data at future iterations) and in the context of fairness, demographic biases incurred due to such feedback models have been studied under the guises of Delayed Impact [Liu et al., 2018] and Residual Unfairness [Kallus and Zhou, 2018]. Indeed, the systematic censoring problem posed by Kallus and Zhou [2018] served as a prime motivator for the setup considered in Chapter 3, such that in the case of a binary decision system – one designed for automated hiring, for instance – and a population comprised of two subgroups, only positive outcomes are observed for the advantaged subgroup, while only negative outcomes are observed for the disadvantaged subgroup.

5 Shortcut learning

While the notoriety of shortcut-learning (SCL) in ML sphere is relatively recent, the phenomenon underpinning it is a fundamental one in statistics, one that may be summed up with the age-old aphorism *correlation does not imply causation.* DNNs define deeply expressive function classes, yet the solutions encoded by their parameter space need not be commensurately complex; in fact, it is well established that these models – in the absence of an countervailing (inductive) bias – exhibit a *simplicity bias* (SB; Valle-Perez et al. [2018]), that is, the tendency to favour simpler solutions, should those solutions serve sufficiently well for the task (as defined by the training set and loss function) at hand. While the spate of failures following the deep-learning revolution were surprising – and at the very least mildly-disenchanting to those with lofty hopes for ML – it is, given thought, *not* surprising that SB should exist and beget SCL, for while SB alone is not alone a precondition for SCL, the second precondition of SCL is a problem (on the data side) that has long challenged statistical modelling: *sampling bias*. It is the combination of simplicity bias and acausal or spurious, correlations generated by sampling bias that give rise to SCL, but sampling bias is not something trivially redressed, even if the seemingly-straightforward recourse of 'collect more data' does exist, which it often doesn't due to physical constraints (e.g. the data may only have been available within a given period of time) or limited (human or monetary) resources. Although the problem may stem from the data-collection side, one is not without recourse on the modelling side, so long as certain assumptions or criteria can be met; indeed, both DG and FairML are active – more so than ever – subfields of ML contending with different flavours of the problem and have successfully developed mitigation strategies for them.

The now-canonical example of shortcut-learning in the DG literature – which we will also invoke here for its simplicity – is due to Beery et al. [2018], wherein the task is one of distinguishing between cows and camels (binary classification). Since camels preponderate on sandy backgrounds, while, by contrast, cows preponderate on grassy backgrounds – owing to their natural habitats – the background is a viable shortcut solution based on which examples from the training set can be reliably predicted while taking the path of least resistance, something the model can hardly be blamed for in absence of the requisite inductive bias to disentangle the true and spurious features. While the brittleness of the shortcut solution will not be exposed if the test set consistently suffers the same sampling bias as the training set, it is perfectly conceivable that a cow could appear on a beach – a common sight on the island of Corsica, for example – and our model would mispredict in such a case because it does not grasp what the concept of a cow truly is – to it, 'grassy' and 'cow' are synonymous. This is a relatively benign example, but there are many real-world cases where this behaviour could lead to life-endangering failures, perhaps most obviously in the medical data domain where one could have a pneumonia classifier that has learned to predict pneumonia from X-ray images with near-perfect accuracy based solely on a hospital-specific token and the hospitals' pneumonia-prevalence rate, as observed by Zech et al. [2018]. There are also obvious ethical concerns that arise when the spurious features in question correspond to protected characteristics like 'race' and 'gender' [Buolamwini and Gebru, 2018, Wang et al., 2019], regardless of aggregate downstream performance. The landmark study by Buolamwini and Gebru [2018], for example, revealed significant disparities in the performance of face analysis algorithms on individuals from marginalised (dark skin, female) vs. non-marginalised (light skin, male).

There are two sides to shortcut-learning that impair generalisation. The obvious one, which we have already belaboured, is 'variance' ' to spurious features – features that are statistically but not causally related to the target; the second one, however, is more subtle and a consequence of the first one, that being *feature suppression*, in that the model is not simply variant to the 'wrong' features but invariant to the 'right' ones – it is not simply a matter of a difference in importance but, in reality, a more pernicious matter of inclusion/exclusion. This is to say, if a shortcut solution is robust enough to achieve near-zero loss on the training set, then there is little incentive – owing to gradient starvation [Pezeshki et al., 2021] and the provably-flatter minima of shortcut solutions [Scimeca et al., 2021] – for the model to learn alternative 'views' (collections of features; Allen-Zhu and Li [2020]). For instance, if texture is a reliable classification cue given the training data Geirhos et al. [2018], a model can latch onto that cue and ignore (be invariant to) other higher-level semantics, like shape and global structure, that human judgements are much more strongly attributed to. High-frequency cues, such as colour and texture, are readily modulated by (unstable under) changes in lighting, for instance, making them less reliably cues for object classification in a dynamic environment; we are not wont, for example, to classify an object in the shape of a cat as an elephant simply because the texture of the latter has been transplanted, *ceteris paribus*, to the former, a failure mode (in)famously shown by Geirhos et al. [2018] to apply to DNNs trained on ImageNet.

With the above in mind, it is obvious why more traditional approaches to improving group- and adversarial-robustness fail. The power of ensembles, for instance, resides in their combining of different views of the data – engendered by stochasticity in the weights and optimisation procedure – yet shortcut solutions create such a strong (easy-to-learn and potent) and stable attractor that all ensemble members simply converge onto that one corresponding view. Domain adversarial learning – popularised by Ganin et al. [2016] and a mainstay throughout the DA, DG, and FairML literature alike – on the other hand suffers from the problem that for the features of the model to be statistically independent of the spurious feature, so must it be statistically independent of the target since the target and spurious feature are themselves strongly correlated, as defined by the SCL problem.

6 Through the lens of causality

We now introduce a causal formalism of the distribution-shift problem, a formalism which has been frequently exploited in the DG and FairML literature as it provides a simple calculus with which to reason about desired (and undesired) variances. It should be noted in advance that we only draw upon this formalism in order to provide a unified formulation of the distribution-shift problems considered in this thesis; we do not operate on the domain of causal graphs nor attempt to perform causal inference. The background on causality is thus commensurably light and we refer the reader to Pearl [2009] for full exposition of the topic.

While the term 'domain' typically refers to the observed distributions as a whole in both DA and DG alike (i.e. 'source' versus 'target'), such terminology is somewhat rigid, as it fails to capture that the distributions share an underlying structure and how and which variables are shifted. It is more arguably flexible then, consistent with [Mooij et al., 2020], to think of the domain as some exogenous latent variable, which, by its conditioning, gives rise to the different observed distributions – or subgraphs in the discrete case – and explains how one is transformed ('shifted') into the other. We will denote said variable as E (for '**E**nvironment', as it is commonly termed in the DG literature [Arjovsky et al., 2019]), which need satisfy only the loose requirement that it belong to some Borel space (and thus may in theory be continuous or discrete). Most simply, in the case of DA, E is simply a binary random variable, such that we have $E : \Omega \to \{\text{source}, \text{target}\}$, with Ω being the sample space.

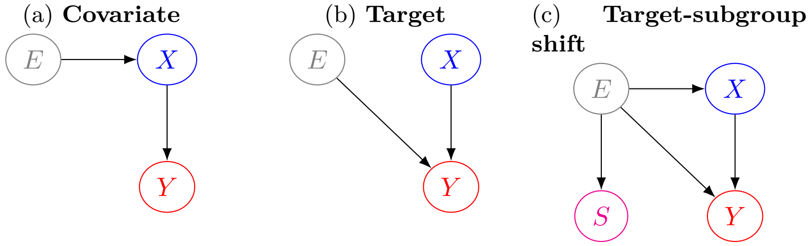
We view then view variables in our prediction task as constituting the nodes $\mathcal{V}$ in a Causal Bayesian Network (CBN; Pearl [1995]) where the direction of arrows (directed edges) between nodes indicate the direction of causality (e.g. $\mathbf{A} \to \mathbf{B}$ means that $\mathbf{A}$ causes (is a parent of) $\mathbf{B}$) while the absence of an edge between two nodes $\mathbf{A}$ and $\mathbf{B}$ indicates independence between them when conditioned on their parents, i.e. $\mathbf{A}|\mathrm{Pa}(\mathbf{A}) \perp \mathbf{B}|\mathrm{Pa}(\mathbf{B})$, where $\mathrm{Pa}(\cdot)$ denotes the causal parents of its argument node. Formally, a CBG is a kind of Directed Acyclic Graph (DAG), $\mathbf{G} \triangleq \langle \mathcal{V}, \xi \rangle$ with node-set (variables), $\mathcal{V}$, and (directed) edge-set, ξ consisting of tuples (ij) meaning $i \to j$, or 'node i is a parent of node j'. Each node in $\mathbf{G}$ then defines a probability distribution, conditional on its parents, such that the joint distribution of $\mathcal{V}$, $P(V)$, factorises as $P(V) = \prod_{v \in \mathcal{V}} P(v|\mathrm{Pa}(v))$ where we can now define $\mathrm{Pa}(\cdot)$ as a function that returns all nodes in ξ that form a pair with v as the second element, i.e. $\{i | i, j \in \xi, j = v\}$.

Without loss of generality, for the prediction task with inputs, X, and targets, Y, we may introduce the aforementioned variable E to convert the joint distribution $P(X, Y)$ into the conditional joint distribution $P(X, Y|E)$; the structure of the underlying CBN determines the factorisation of this distribution and thus the nature of the distribution shift in question. One can, for example characterise the case of covariate-shift with causal $f^\star$, as having edges $E \to X$ and $X \to Y$, giving rise to the factorisation $P(X, Y|E)|P(E) = P(Y|X)P(X|E)P(E)$. In Chapters 3 and 4 we go beyond the bivariate (excepting E) and covariate case and consider label-shift problems with an additional auxiliary label S – corresponding to an identifier of some subgroup or spurious feature we wish to be invariant to in the name of fairness or generalisation – in which E influences the joint distribution $P(S, Y)$ but not the marginal distribution $P(X)$, giving rise to representation bias and, from it, spurious correlations. We illustrate in Fig. 1 CBGs corresponding to different distribution shifts for a causal prediction task.

We will see that such a formulation is particularly useful when we come to discuss domain generalisation which involves a multitude of source distributions, going beyond the dyadic source-target setup characterising DA.

One common [Arjovsky et al., 2019, Krueger et al., 2021, Sagawa et al., 2019] and intuitive way of formulating the robust-prediction problem is as one of bilevel optimisation, where the inner loop entails computing the empirical risk over each domain and the outer loop corresponds to finding the function that minimises the maximum of said risks. We can then define, accordingly, the optimal predictor as

$$f^*_{\text{robust}} = \arg\min_{f \in \mathcal{F}} \max_{e \in \mathcal{E}} \mathbb{E}_{P^{tr}(X,Y|E=e)}[D(f(X), Y)]. \tag{3}$$


(a) **Covariate**

(b) **Target**

(c) **Target-subgroup shift**

Figure 1: Causal Bayesian Graphs (CBGs) corresponding to different distribution shifts, induced by exogenous variable, E, for a causal prediction ($X \to Y$) task, where X and Y correspond to the covariates and response variable, respectively. (c) introduces an additional, auxiliary label, S, which forms the basis of the problems tackled in Chapters 3 an 4.

As discussed at length before, domains/environments can be modelled as deriving from different interventions of the causal factorisation of $P(X, Y)$. It follows that, for Eq.3 to engender successful generalisation to arbitrary domains, $\mathcal{E}^\dagger$ outside $\mathcal{E}$ ($\mathcal{E}^\dagger \cap \mathcal{E} = \varnothing$; i.e. the goal of DG), $\mathcal{E}$ must be a representative (well-covering) set of samples from the generating distribution, $P(E)$, such that smooth interpolation along the underlying manifold is possible. When $\mathcal{E}$ is a finite set, as above, one can think of it as *perturbation set*, accordant with the robust optimisation literature [Ben-Tal et al., 2009]. While generalisation to arbitrary perturbations is provably hard (or impossible), in general [David et al., 2010], when $\mathcal{E}$ encodes prior information about the kinds of perturbations one expects to encounter at test-time then incorporating it into the optimisation process can be fruitful, both for allowing interpolation within the convex hull defined by the set and to an extrapolated region outside of it [Krueger et al., 2021]. Indeed, it stands to reason that by allowing the model to glean which features are and are not stable across environments would allow it to better approximate the true causal structure of the prediction task. This idea has been explored extensively in recent years in both the causal discovery [Peters et al., 2016, Bengio et al., 2019] and domain generalisation [Arjovsky et al., 2019, Ahuja et al., 2020, Creager et al., 2021] literature, with the caveat that a degree of inductive bias or additional information is necessary to provably identify the proper causal relations based on it [Lin et al., 2022].

7 Domain adaptation

Domain adaptation is a subfield of machine learning that deals with the problem of adapting a model trained on one distribution (the *source domain*) to a different but related distribution (the *target domain*), in such a way that the relevant shared structure is exploited while nuisance factors that are domain-specific (and not relevant to the prediction task) ignored. The downstream performance of the model is thus naturally dependent on both the performance on the source domain and by the degree of relatedness between the source and target domains. To proffer a real-world example, in building a spam detector, one might have annotated data (emails) available for training a model sourced from a previous group of users and wish to deploy (adapt) the detector to a new group of users in such a way that is robust to the temporal distribution shift.

In the classical DA setting, one assumes the distribution shift is *covariate* [David et al., 2010] in nature, that is, localised to the marginal distribution $P(X)$, with both the conditional, $P(Y|X)$ (corresponding to changes in the ground-truth labelling mechanism, $f^\star : \mathcal{X} \rightarrow \mathcal{Y}$), and label, $P(Y)$, distributions consistent across domains. This is not to say that there is not a substantial body of work that addresses other types of distribution shift [Zhao et al., 2019], and the covariate-shift assumption is perhaps stricter than one might initially presume. Indeed, it turns out that the covariate-shift assumption is only tenable in cases where $f^\star$ is *causal* $(X \rightarrow Y)$; practically, there are many cases for which the converse in fact holds true true, that the relationship between X and Y is *anticausal* $(Y \rightarrow X)$. Anticausal prediction tasks naturally crop up in the medical-imaging domain for instance, where Y is some gold-standard indicator of the presence of the disease and it is the disease that gives rise to aberrations in the input images signalling to a classifier a positive instance. For the task of melanoma-prediction, for example, one may be interested in training a classifier to diagnose patients based only on dermoscopic images using labels derived from (expensive and time-consuming but reliable) histopathological analysis [Castro et al., 2020]. The distinction between causal and anticausal tasks is an important one in ML generally, and we will revisit the idea on several occasions throughout the remainder of this chapter; for SemiSL, said distinction is particularly important as the efficacy of the paradigm hinges on $P(X)$ carrying information about f^*, and thus the task being an anticausal one.

8 Domain generalisation

While closely-related to DA, Domain Generalisation (DG) is distinct in the respect that the task is fundamentally, as the name suggests, one of o.o.d. generalisation rather than one of adaptation. By this we mean that while in (U)DA one is given a labelled dataset, $\mathcal{D}^{src} \triangleq \{x_i, y_i\}_{i=1}^{N^{src}}$, belonging to the source domain, along with an unlabelled dataset $\mathcal{D}^{tgt} \triangleq \{x_i\}_{i=1}^{N^{tgt}}$ belonging to the target domain, and the goal is to train a classifier to generalise from the former to the latter (which entails a degree of invariance), DG is more general, in that one is instead given datasets from multiple domains and seeks to train a classifier that can generalise to previously unseen ones. That is, given the (empirical) meta distribution $\mathfrak{D} \triangleq \{\mathcal{D}_e\}_{e \in \mathcal{E}}$ consisting of $|\mathcal{E}|$ distributions drawn from different domains, or *environments*, denoted by the index set $\mathcal{E} \subset \mathbb{N}$, the goal is to train a classifier that will perform optimally, or with minimal degradation, when presented with distribution $\mathcal{D}_{e^{te}}$ from a novel domain $e^{te} \notin \mathcal{E}$.

While we would ideally have a model that could generalise to any arbitrary environment (assuming the task remains consistent), this is sadly impossible given finite data [David et al., 2010], and so our expectations must be tempered to being able to generalise within some region the training distribution. The justification of DG can then be viewed from two perspectives:

1. It stands to reason that we should be able to exploit information about known the known set of variances – due to the domain – in order to learn a predictor that can generalise within the convex set (affine combinations of those variances) they define as well as to those variances that are close by. The principle here is similar to that of vicinal risk minimisation [Chapelle et al., 2000] as in Zhang et al. [2017], wherein data augmentation fulfils the role of a perturbation set that the environments fulfil in DG.

2. Given a set of interventions on the underlying causal graph defined by the set of environments, recover the causal relationship between the input features and the target such that the predictive mechanism is unaffected by causally-independent changes (by interventions on variables not among the target variable's causal parents). This idea of treating environments as interventions and using them to perform explicit or implicit causal inference has notably been exploited in Peters et al. [2016] and in the foundational (to DG) work of Arjovsky et al. [2019]. Indeed, in the wake of Arjovsky et al. [2019], it has become common [Gulrajani and Lopez-Paz, 2020, Krueger et al., 2021, Mahajan et al., 2021, Lin et al., 2022] to express the problem setup of DG and its desiderata in causal terms, and we will do so ourselves in §6 in order to provide a more unified perspective of the distribution shift problems discussed thus far.

9 Semi-supervised learning

Given that this thesis references semi-supervised learning (SemiSL) [1] in its title, it is only appropriate that a part of this background section be devoted to the topic. However, we note that the methods introduced in this thesis are not tailored for the typical SemiSL regime wherein one hopes to draw upon a large corpus of unlabelled data to shore up the paucity of annotated data available for direct supervision, with the assumption being that the unlabelled and labelled data are drawn from the same distribution. Rather, the unifying theme across the constituent papers is how one can use unlabelled data to buttress against different types of distribution shift. The problem setups considered thus more closely align with those found in DG and DA.

We would refer the reader to Chapelle et al. [2009] for excellent (in both clarity and depth) exposition of the theoretical underpinnings of SemiSL and methods for it from the pre-deep-learning era (many of which are still fundamentally applicable today, however), a book we will reference extensively throughout this

[1] Since self-supervised learning also features prominently in this thesis (primarily in Chapter 5) we must depart from the typical initialism, SSL, so as to be able to differentiate the two learning paradigms.

brief overview of the topic. For a comprehensive and current survey of SemiSL methods in the post-deep-learning, on the other hand, we would refer the reader to Yang et al. [2022].

The premise of SemiSL is a simple one: given that in many real-world cases collecting annotated data is expensive (monetarily or temporally) or even prohibitive (for instance, if the data is tied to a transient phenomenon) but collecting data in general is not, how can one exploit the unannotated data to improve a model's predictive power? Thus, we can think of our dataset as having two partitions: one corresponding to the labelled dataset, as in §3, which we can use for standard supervised learning and which we will override here with the notation $\mathcal{D}_l^{tr} \triangleq \{x_i\}_{i=1}^{N_l^{tr}}$ for clarity's sake, and a second partition corresponding to the unannotated data, which we will denote by $\mathcal{D}_u^{tr} \triangleq \{x_i\}_{i=1}^{N_u^{tr}}$. One is generally motivated to employ some form of SemiSL in cases when $N_l^{tr} \ll N_u^{tr}$, though this not need be the case and it may be that the unlabelled data can be useful beyond simply providing more data from the same distribution (as the labelled data), as epitomised by UDA and as explored in other contexts throughput this thesis. There exist many different branches and perspectives of SemiSL, as a learning paradigm with deep roots reaching as far back as the 60s [Scudder, 1965, Fralick, 1967]. On the perspective side, one can, for instance, view SemiSL as unsupervised learning subject to constraints – which is especially pertinent in the case of semi-supervised clustering [Bair, 2013] – though it is usually more natural to frame it from the opposite perspective, namely, as supervised learning with additional information [Chapelle et al., 2009].

Transduction

Closely related to SemiSL is the idea of *Transductive Learning* (TL), as pioneered by Vapnik [Gammerman et al., 1998]. With TL, rather than pursuing the lofty goal of learning a predictor that can generalise across the entire input domain, $\mathcal{X}$ – reflecting the inductive process of extrapolation – one instead focuses on predicting well on a restricted domain defined by the test points – reflecting the transductive process of transferring rules between specific cases. That is, the optimisation problem is reduced from finding the (loss-function) minimiser over $f \in \mathcal{F}$ to the considerably more tractable problem of finding the minimiser over $f|_{\mathcal{X}^{te}} \in \mathcal{F}$. Intuitively, it makes sense to optimise the predictor for the subset of the domain of interest, given that one has the access to said subset and has the necessary time/resources, rather than to take the more circuitous approach of learning general rules and applying them to specific cases (the process of *deduction*). To couch TL in SemiSL terms then simply requires equating $\mathcal{D}_u^{tr}$ with the test set $\mathcal{D}^{te}$. We can view this distinction between TL and (inductive) SemiSL as analogous to the distinction between domain adaptation and domain generalisation, in the sense that former has the transductive goal of generalising between two specific domains – with the target domain given at training time – while the latter has the inductive goal of generalising to all possible domains. Niceness of this inter-field parallel aside, we afford TL particular mention here due to its pertinence in Chapter 4, wherein we consider the possibility of using the test set itself as a reference dataset for the proposed matching procedure.

9.1 Justifying SemiSL

While SemiSL is a tantalising prospect whenever one has a large corpus of unlabelled data, and relatively sparing labelled data, it is unfortunately not the case that whenever there is unlabelled data available that one can mine it for additional information about the given task. In fact, in some cases – those in which there is an element of distribution shift – one might find a degradation in performance when using SemiSL, relative to the supervised baseline trained on a small fraction of the samples. SemiSL theory requires that certain assumptions about the data-generating process be met in order for the learning paradigm to bear fruit (in sense of reduced generalisation-error/improved sample efficiency), whatever the chosen method, though this is not to say that there one can't observe practical benefits – such as improved convergence-rates or stability – detached from those assumptions. We will broach the importance of the direction of causality – that is how the data-generating process factorises – later in this section; to begin with, following [Chapelle et al., 2009], we summarise the justifying assumptions for SemiSL. These assumptions are not complementary, in the sense that they can, or need be, simultaneously satisfied; rather they provide three different perspectives leading to different classes of algorithms. The cluster and low-density-separation (LDS) assumptions most obviously form a dual-view of the same fundamental principle, understanding 'cluster' to mean 'high-density-connectedness'.

Cluster Assumption

To elaborate, the *cluster assumption* posits that data-points that are connected by a path through density regions should belong to the same class. This is precisely the assumption that drives many traditional (density-based) clustering algorithms aiming to separate the data into groups of samples, or 'clusters', using density (estimable, for instance, with kernel methods or neighbourhood graphs) as a surrogate for ground-truth labels. Within SemiSL itself, the cluster assumption is well encapsulated by the method of *label propagation* [Szummer and Jaakkola, 2001] which (with great simplification) involves 1) building a neighbourhood graph with the labelled and unlabelled samples as its vertices and the edges weighted according to local correlation strength; 2) propagating the label distributions from the labelled samples to the adjacent unlabelled samples in a Markovian fashion.

Low-density-separation Assumption

If one flips the cluster assumption, such that we have instead have the axiom 'data-points that are not connected by paths through high-density regions belong to different clusters', then one obtains the LDS assumption, though this is more commonly expressed in terms of the decision boundary, namely that the plane separating any two classes should carve out a region of low density. Despite the equivalence, the two afford very different perspectives, from an optimisation standpoint. Indeed, the aforementioned density-based clustering algorithms, of which DBSCAN [Ester et al., 1996] is the paradigmatic example, focus on the data-points themselves – grouping together those that are sufficiently close (dense) – rather than on the space between them – that is, the problem of choosing the set of separating planes with sufficiently-low path integrals. A classical example of a SemiSL method derived from this principle is

Transductive Support Vector Machines (TSVMs; Joachims et al. [1999]), which share inductive Support Vector Machines (SVMs) aim of maximising the margin between the decision boundary and nearest data-points (the *support vectors*) – yielding the *maximum margin hyperplane* – yet consider both the (labelled) training and (unlabelled) test data during this procedure.

Smoothness Assumption

The final member of the triad, the *smoothness assumption*, can be viewed as imposing a kind of local K-Lipschitz or ϵ-isometric constraint on our function class, $\mathcal{F}$, local in the sense of applying only to high-density regions of the input space (whereas a global smoothness assumption would require even sparse regions of the input space to obey the constraint); i.e. for a pair of inputs x and x', we have

$$|d_{\mathcal{Y}}(f(x), f(x')) - d_{\mathcal{X}}(x, x')| \leq \epsilon, \tag{4}$$

with $d_{\mathcal{Y}}$ and $d_{\mathcal{X}}$ being the metrics (distances) associated with metric spaces $\mathcal{Y}$ (the output space) and $\mathcal{X}$ (the input space), respectively. Plainly speaking, the assumption embodies the desire to have similar inputs map to similar outputs (that distance should be preserved up to some relaxation factor) and from this assumption one naturally obtains the class of *consistency-regularised* methods. Equally, it should be possible to smoothly interpolate between the images of x and x' without straying into low-density regions (which, per the LDS assumption, define separating planes) and with this perspective, we are granted a reformulation of the cluster assumption. More generally, we can simply determine any two samples to be similar (e.g. based on a neighbourhood graph) and then seek to minimise the distance between them in $\mathcal{Y}$, thereby partially discretising the problem but allowing for increased flexibility. This is capitalised on in Chapter 4 we propose a consistency-regularised method for DG where pairs are determined by a cross-domain causal-matching algorithm and consistency is enforced between the members of those pairs.

9.2 Causal connections: when should(n't) SemiSL work?

Following Schölkopf et al. [2021], start by supposing that our prediction task follows the causal factorisation $X \to Y$, i.e. it is a causal, rather than an anticausal, one. As discussed before, the ICM principle states that modules in a joint distribution's causal decomposition do not inform or influence one another, i.e. $\mathcal{I}(X, Y) = 0$; this implies that in the when X is the causal parent to Y, as in the supposed case, a better estimate of $P(X)$ does not yield a better estimate of $P(Y|X)$ and it is the former that SemiSL compasses to learn using unlabelled data. However, that is not to say that SemiSL as in its totality is misguided, it merely requires the right condition to be met, that condition (which applies to a wide-range of real-world problems) being the contrary factorisation, $Y \to X$, which is to say that that the task under consideration is anticausal. In this case, X can contain information about the labelling mechanism, as X is now the effect, and Y is now the cause, opening up the possibility of exploiting dependencies in the marginal distribution to better estimate the conditional distribution; in Schölkopf et al. [2012] the authors corroborate this hypothesis.

While perhaps often overlooked, this requirement, in fact, well-aligned with the motivating arguments for SemiSL that we discussed at the beginning of this section. The *cluster assumption* predicates that points belonging to the same cluster in $P(X)$ abide by the same labelling mechanism; the LDS assumption predicates that the region in which $P(Y|X)$ is maximally entropic (defining the decision boundary) should have low $P(X)$, or, by invocation of Bayes' Theorem, low $\frac{P(X|Y)P(Y)}{P(Y|X)}$; the *smoothness assumption* predicates that if two inputs in a high-density region are close, then their respective images (under the model) also should be; notice that in all three of these cases the causal factorisation is implied to be $Y \to X$. The celebrated co-training theorem [Blum and Mitchell, 1998] similarly respects this precondition of anticausality in assuming that the co-trained predictors are conditionally independent given the label, as one would have if the label were the cause (the causal parent in the bivariate causal graph).

10 Fair machine learning

Research into algorithmic fairness, FairML, has flourished in recent years, the subfield growing from what was once an arguably niche one to one of great prominence, borne by ML's proliferation – and thus increased capacity to affect real lives consequentially – in all sectors of society. Indeed, a multitude of high-profile cases/studies have highlighted the discriminatory (unfair) nature of unchecked ML systems – Kasperkevic [2015], Angwin et al. [2016], Dastin [2018], and Buolamwini and Gebru [2018] – further spurring research to develop better-aligned algorithms and methods for validating them, and thereby regain public trust.

There are many strands of FairML following different criteria for what it means for a predictor to be *fair*; in this thesis we focus on those corresponding to group definitions of fairness [Barocas et al., 2019], where 'group' refers to some demographic group, such as gender or ethnicity, that is considered *sensitive* or *protected* (these two terms are often used interchangeably throughout the literature; we will favour the former). We will denote group membership using the random variable $S : \Omega \to \mathcal{S}$ – with realisation s – and will assume that this variable along with the target variable are discrete (and in most cases binary); this is the most common setup considered in the FairML literature – and toward which many standard metrics of fairness are geared [Feldman et al., 2015, Hardt et al., 2016, Woodworth et al., 2017] – though there are works that extend notions of – and methods for enforcing – fairness to settings with categorical and continuous (S and Y) attributes. Though the focus is on group fairness, at the tail-end of the section, we will touch briefly on the premise of individual fairness [Dwork et al., 2012] to take the opportunity to draw a parallel between it and the idea of consistency/smoothness in SemiSL.

Particularly, we will consider two particular kinds of group-oriented notions of fairness: 1) the family of notions predicated on equalised rates, imposing constraints on the predictive distribution; 2) the notion of *minimax fairness* predicated on maximising the worst-group performance, and which imposes no constraints on the relative performance between groups. It is interesting to note that, from an optimisation perspective, the latter partially subsumes the former: by solving its entailed problem, one can then readily solve the former by artificially inflating the error on the advantaged group (systematically flipping the predicted labels until the relevant constraint on the predictive rates is met).

Simply put, the unified goal of FairML is to learn some predictor, $\Gamma : \mathcal{X} \to \mathcal{Y}$ that is, according to some definition, non-discriminatory – that does not deprive a given individual of opportunities by dint of belonging to a particular sensitive group. A frequently-invoked example is that of financial institution employing an automated decision-making system to determine which loan applicants should/should not have their applications approved. In addition to financial information (e.g. credit-score history) the input features to Γ, $\mathcal{X}$, may encode, explicitly or implicitly, sensitive information such as an individual's gender or ethnicity, with the general assumption being that such information is acausal to the task at hand (one's gender should have no impact on one's eligibility for a loan). By 'implicitly', we mean that such information may be inferable (or predictable with above random accuracy) from other features, such as one's location, housing history, or level of education, such that solving problems of fairness is not so straightforward as simply excising the elements of $\mathcal{X}$ that directly correspond to the $\mathcal{S}$ – an approach referred to as *Fairness Through Unawareness* (FTU).

10.1 Equalised rates

A long-standing and vigorously-debated problem in FairML spheres is how exactly one should define the concept of 'fairness'; indeed what constitutes a 'fair' decision depends both on the (individual, social, or institutional) value system and the context in question. Much of the fairness literature has focussed on notions of fairness based on enforcing the predictor to output equal rates (e.g. of positive or correct predictions) across the sensitive groups. Here, we briefly discuss and formulate *Demographic Parity*, *Equal of Opportunity*, and *Equalised Odds* as the standard triad of metrics based on this tenet of equalisation of inter-group rates, and do so for the binary-classification regime to which they are most commonly applied, even though generalisations exist [Woodworth et al., 2017]. These metrics naturally share similarity of form, only differing in their conditioning and bear direct relevance to Chapter 3, the constituent paper of which being fairness-oriented and adopting these metrics as measures of invariance, though the method is applicable to spurious correlation problems in general, decontextualised from fairness. Chapter 4 adopts a perspective in line with the domain generalisation literature Sagawa et al. [2019], a perspective which is known under the guise of *minimax fairness* in the FairML literature. While perhaps obvious, it is relevant to note that that a predictive distribution provably cannot simultaneously satisfy all three notions of fairness dictated by DP, EqOp and EO [Kleinberg et al., 2016]. It is also relevant to note that these metrics presuppose scenarios in which the budget, the pool of allocatable resources, is limited (a cap on the number of grantable loans or the number of people that can be employed, for example), such that any resources allocated to an individual of one group are concomitantly being withheld from the other group(s) – fairness in this context thus corresponds to a type of resource allocation problem, from an econometric perspective.

Demographic Parity

The simplest of the triad, *Demographic parity* (DP; Zemel et al. [2013], Feldman et al. [2015]) – known also as statistical parity, group fairness, disparate impact, inter alia – demands that the probability of a positive prediction (positive rate) be uniform (at parity) across all sensitive groups. That is

$$\forall s \in \mathcal{S} : P(\hat{Y} = 1|S = s)\overset{!}{=}\text{constant}, \tag{5}$$

where $\hat{Y}$ denotes the random variable corresponding to the predictions of a given predictor, f, and ‘constant‘ denotes some placeholder constant value, noting that this could be any arbitrary value and does not take into account utility, such that a majority or randomly classifier would degenerately satisfy the condition. This above constraint is equivalent to requiring, according to the standard notion of statistical independence, namely the equality of the conditional and marginal distributions, i.e.

$$P(\hat{Y} = 1|S)\overset{!}{=}P(\hat{Y} = 1). \tag{6}$$

Since requiring that this condition be satisfied exactly is generally overly strict when doing constrained optimisation, it is common to introduce some relaxation factor, ϵ, that expands the constraint to a feasible region so that with some minor rearrangement we may instead write:

$$\forall s \in \mathcal{S} : P(\hat{Y} = 1|S = s)P(S = s) \in [1 - \epsilon, 1 + \epsilon]. \tag{7}$$

We can measure fairness (one notion of it, at least) of a predictor by evaluating by how much it violates this condition (or any of the conditions in this section), either in terms of differences or ratios which in non-binary cases can be computed pairwise and then optionally summarised by taking the maximum over the resulting set.

Mutual-information minimisation

This idea of statistical independence, upon which DP hinges, can be expressed generally in terms of the mutual information (MI) between Y and S, itself expressible as the KL divergence between the joint and product of the marginal distributions:

$$\mathcal{I}(Y; S) \triangleq D_{KL}\Big(P(Y, S)\|P(Y) \otimes P(S)\Big). \tag{8}$$

Iff $\mathcal{I}(Y; S)$, $\mathcal{I}(Y; S) = 0$ can the random variables said to be statistical independent. MI admits various decompositions into sums of marginal and joint/conditional entropies that make it particularly amenable for optimisation purposes. In invariant representation learning (encompassing fair representation learning, domain adaptation, and domain generalisation), for example, a common method for imparting independence, $Z \perp S$ – which is sufficient for $Y \perp S$, given a predictor head, $c : \mathcal{Z} \to \mathcal{Y}$ – between the representations learned by encoder $g : \mathcal{X} \to \mathcal{Z}$ and the sensitive attribute to train g to maximise the conditional entropy $H(\hat{S}|Z)$ generated by an adversarial predictor, $a : \mathcal{Z} \to \triangle^{|\mathcal{S}|}$ (itself trained via MLE).

Equality of Opportunity

Equality of Opportunity (EO) relaxes this desideratum of unconditional statistical independence, $\hat{Y} \perp S$, to one of *separation* (or conditional independence), dictating that $\hat{Y}$ and S need only be independent conditioned on the ground-truth label, Y (albeit only in the positive case). To phrase this conversely: whenever the outputs of our predictor are dependent on S, such must be justified by a dependence on Y for the predictor to be a fair one. Thus, EO can be written, as above, as

$$\forall s \in \mathcal{S} : P(\hat{Y} = 1|Y = 1, S = s) = P(\hat{Y} = 1|Y = 1), \tag{9}$$

Equalised Odds

By making the above symmetric w.r.t. Y, such that not only do we demand parity of the TPRs but also false-positive rates (FPRs) we obtain the final of the ERs-based metrics, *Equalised Odds* (EqOd; Hardt et al. [2016]) – also known as disparate mistreatment:

$$\forall ys \in \mathcal{Y}, \forall s \in \mathcal{S} : P(\hat{Y} = 1|Y = y, S = s) = P(\hat{Y} = 1|Y = y) \tag{10}$$

Fairness-accuracy trade-off

It has long been understood that there exists an inherent trade-off between the utility, as measured by aggregate performance and fairness under conceptions of fairness based on ERs [Kaplow and Shavell, 1999]; with accuracy being the principal measure of said utility in the context of classification, this trade-off is commonly referred to as the *accuracy-fairness trade-off*, though as we will see in the next subsection, such a trade-off does not apply to fairness in general, as in the case for notions of fairness defined by *minimax fairness* where one seeks to maximise worst-group utility rather than group parity. Given the existence of such a trade-off, the problem of learning a useful classifier subject to ER constraints induces not one optimal solution (or one equivalence class of optimal solutions, more accurately) – as one would have when focussed on only the utility – but rather a set of *Pareto optimal* solutions, the discovery of which is the remit of multi-objective optimisation (MOO; Sawaragi et al. [1985], Deb and Deb [2013]). MOO has appeared both implicitly and explicitly throughout the FairML literature, the latter only relatively recently (e.g. in Navon et al. [2020]). Indeed, examples of the former include the methods of Louizos et al. [2015] and Madras et al. [2018] which entail learning a *linear scalarised* solution [Boyd et al., 2004], with position of this solution on the *Pareto frontier* controlled by a linear weighting of loss terms optimising for utility and fairness separately.

10.2 Going beyond the fairness-accuracy trade-off with minimax group fairness

The foregoing notions are as intuitive as they are well-studied, and there is a wide range of applications to which they may be reasonably used as a lodestar for fairness. However, the fact that they require degrading performance of the advantaged groups is problematic for applications where the quality of service

– the utility of the model – is cardinal, or – phrased econometrically – scenarios not characterised by limited resources and thus to which the 'Robin Hood' principle of 'stealing from the rich to give to the poor' is inapplicable. Healthcare defines a whole host of applications to which this consideration applies, where accurately detecting positive cases can be a matter of life-or-death, and any marked degradation in this respect in the name of fairness is unacceptable. Rather than satisfying constraints on predictive parity, a more reasonable tact in such scenarios is instead to aim to maximise fairness while incurring minimal (ideally, no) degradation in the performance on any given group [Ustun et al., 2019]. This is the remit of *minimax fairness*, as originally formulated by [Martinez et al., 2020] in Pareto-optimal terms, though the same idea has long existed in distributionally robust optimisation (DRO), and an idea which has been inherited by domain generalisation as an instantiation of DRO.

The 'distributionally robust' part of DRO corresponds to the desire to find a solution that works well not on only a single distribution, or particular instantiation of a problem, but that works well over a range of proximal distributions/problems (also referred to as a *perturbation set* in some texts [Ben-Tal et al., 2009]). This desire naturally arises in any regime which naturally contends with some kind of meta distribution – a distribution over distributions – such as meta-learning [Collins et al., 2020], domain generalisation (as already noted; Sagawa et al. [2019]), or, indeed, fairness. Given a meta distribution $\mathfrak{P}(X, Y)$, from which we sample the bivariate distributions $P(X, Y)$, the minimax (ERM-based) DRO objective can be expressed as

$$\inf_{\Gamma} \sup_{P(X,Y)\sim\mathfrak{P}(X,Y)} \mathbb{E}_{(X,Y)\sim P(X,Y)}[\mathcal{L}(\Gamma(X), Y)], \tag{11}$$

recalling that we use $\mathcal{L}(\cdot, \cdot)$ to denote a loss function of random variables. Thus, in contrast to standard ERM, which would optimise over the 'flattened' meta distribution, the DRO objective defines a bilevel optimisation problem in which only the supremum of the loss over $\mathfrak{P}(X, Y)$ contributes to the overall objective to be minimised by our predictor, Γ. One can view this as a sparse form of IW-ERM where the weighting function is the indicator function returning 1 if a sample belongs to the 'worst' (most divergent) distribution and 0 otherwise.

For evaluation of classification tasks, one can align the standard notion of accuracy with minimax fairness by conditioning on the sensitive attributes and taking the minimum (worst) over the resulting set of $|\mathcal{S}|$ values. In the domain-generalisation/group-robustness literature this quantity is commonly known as *Robust Accuracy* (RobAcc). For convenience and clarity, we first restate standard (non-conditional) accuracy as

$$\mathrm{Acc}(f, \mathcal{D}^{eval}) \triangleq \mathbb{E}_{(x,y)\in\mathcal{D}^{eval}}[\delta_{f(x)y}], \tag{12}$$

with $\mathcal{D}^{eval}$ denoting the dataset over which the metric is being computed, δ Kronecker delta function that evaluates to 1 under equality of $f(x)$ and y (and 0 otherwise). The 'robust' version (RobAcc) is then simply the minimum accuracy computed over all subsets of $\mathcal{D}^{eval}$ created by conditioning on each $s \in \mathcal{S}$

$$\mathrm{RobAcc}(f, \mathcal{D}^{eval}) \triangleq \min_{s \in \mathcal{S}} \mathrm{Acc}(f, \mathcal{D}^{eval}_{S=s}), \tag{13}$$

with $\mathcal{D}^{eval}_{S=s}$ denoting the sth one of such subsets.

10.3 Individual fairness

While notions of group fairness consider fairness at the level of demographic groups, *individual fairness* focusses – as expected of the name – on fairness at the individual level, and may be pithily summed up with the apophthegm 'similar individuals should receive similar treatments'. As alluded to before, the premise of individual fairness greatly resembles the smoothness assumption from SemiSL. We mentioned in its associated section that the smoothness assumption can be seen as a K-Lipschitz constraint on our function class and this is the manner in which the *Fairness Through Awareness* (FTA) – the most general concept of individual fairness – is couched in Dwork et al. [2012], the name standing in contrast to FTU. The central question implied by FTA then is what constitutes an appropriate measure of similarity – in the input and output spaces separately – for the population and task under consideration.

10.4 Bias propagation and systematic censoring

We conclude this section with a brief discussion of the residual unfairness, as termed by Kallus and Zhou [2018], and the problem setting giving rise to it, owing to its pertinence to both Chapter 3 and Chapter 4. Residual unfairness refers to lingering inter-group disparities, stemming from sampling bias – in the fairness-contextualised sense of resulting from prejudiced historical policies engendered by limited initial data, heterogeneous decision-makers, or statistically discriminatory rules – after attempts to correct for those disparities, due to mismatches between the training and test populations. When data collected subject to such a mechanism is used to inform a decision policy – automated or otherwise – that then informs future policies, the enforcer is liable to creating positive (self-reinforcing) feedback loops that lead to the progressive amplification of already-grievous systemic biases, to the extent of *systematically censoring* certain outcomes for certain demographics.

To ground this, consider a recruitment policy giving preferential treatment to male software developers (such that men are significantly more likely to be hired than women) by virtue of men historically preponderating over women in the technology sector. Crucially, the nature of the process means that one only observes outcomes – performance metrics – for those individuals that are hired, i.e. we do not have access to counterfactual (what would have been had A happened instead of B) information that would provide an avenue for natural equilibration. Over several recruitment cycles with 'dynamics' governed by the aforementioned policy – and potentially updates to that policy incorporating data from new hirees – one would expect to end up with a training population remote from that of the ‘true’ population, potentially to the point of women vanishing – being censored – from the pool of hirees, while the rejected pool consists of a mixture of men and women. Kallus and Zhou [2018] show that in such cases, enforcing fairness on the training distribution provably does not guarantee fairness on the true population.

11 Adversarial Learning

Minimax

Adversarial learning (AdvL) is a general, multi-field-spanning learning paradigm, characterising – in game-theoretical parlance – non-cooperative (competitive) systems, or ‘games’ in which two or more ‘players’ compete over shared or interdependent *utility* [Fudenberg and Tirole, 1991]. In two-player cases – accounting for the majority of AdvL setups (notable exceptions to this include those based on self-play in which one computes the best response against a mixture of adversarial policies [Silver et al., 2017, Vinyals et al., 2019]) and those that we shall discuss here – the game can be formulated as a (zero-sum) minimax problem in which one player, μ_{max}, plays the role of the ‘maximiser’, the other player, μ_{min}, the role of the minimiser; given that in ML we optimise some parametric model to minimise a loss function (via gradient descent), it is natural to view μ_{min} as the model of interest, or ‘learner’, and μ_{max} as the ‘adversary‘ which frustrates μ_{min} in order to improve its own payoffs (but with the end goal of the game ultimately being to improve μ_{min} in some respect, such as robustness or fairness).

Bilevel optimisation

Nash equilibria

Such a problem can alternatively be viewed as an alternating (turn-based) bilevel optimisation problem, where at the tth iterate the learner selects (via some optimisation procedure) from Σ_{min}, the strategy that gives the best response, $\sigma^*_{min,t}$, to the best response of the adversary at the previous iterate, i.e. $\sigma^*_{max}|_{\sigma^*_{min,t-1}}$. The best response is determined by each players respective utility, as measured by the player-specific function $\pi : \Sigma \rightarrow \mathbb{R}^+$, with Σ denoting the space of strategic profiles, a strategic profile itself referring to tuple of chosen strategies characterising a game state. A hallmark of the adversarial regime is that the strategy of maximiser is dependent on the state of the minimiser and if the state of the minimiser changes so does the best-response, unless in a state of *Nash Equilibrium* (NE; or at least a local one). The concept of a NE is fundamental to game theory, referring to strategic profiles, $\sigma^* \in \Sigma^*$, from which no player can unilaterally deviate and achieve greater payoff, or, mathematically (treating μ_{min} as the reference player)

$$\forall \sigma_{min} \in \Sigma_{min} : \pi_{min}(\sigma^*_{min}, \sigma^*_{max}) \geq \pi_{min}(\sigma_{min}, \sigma^*_{max}). \tag{14}$$

where σ^*_{min} and σ^*_{max} are NE strategies for μ_{min} and μ_{max}, that by their pairing make up some NE strategic profile, σ^*. By this non-strict definition the set of NE solutions can be non-singleton – by virtue of the inclusive inequality – such that for any given game one may have a set of one or more NE strategic profiles. The notion of a NE is closely related to the idea of Pareto optimality in MOO, noting, however, that a NE strategic profile need not be a Pareto optimal one, PO being defined w.r.t. the maximum theoretically-achievable utility for each player (given the utilities of every other player), NE solely w.r.t. the relative utilities of the players and the resulting fixed points.

The minimax formulations we saw in the context of minimax fairness and (worst-group) distributional robustness comply – perhaps subtly – with this definition as the strategy of the maximiser can be interpreted as a weighting function whose best-response is the one under which only the loss of the highest-loss group contributes to the overall loss. Practically, computing the best response for each player over the entire dataset for each iterate is computationally intractable for most non-trivial cases, especially so if said players are deep neural networks and many iterates are required for convergence, and so some degree of approximation is required. This usually translates to performing only a fixed budget of updates in the minimising/minimising direction over random subsets of the data. In fact, Ganin et al. [2016] demonstrated that, in practice, one can ignore best-response dynamics altogether and obtain approximately domain-invariant representations with minimal overhead through concurrent (single-step) updates to the players' strategies.

Adversarial infomin

Notable concrete applications of adversarial learning include artificial curiosity [Schmidhuber, 1992], Generative Adversarial Networks (GANs; Goodfellow et al. [2014]), self-play [Silver et al., 2018], adversarial robustness [Szegedy et al., 2013], and, most germanely to this thesis, domain-invariant [Ganin et al., 2016, Zhao et al., 2019] and fair-representation learning [Edwards and Storkey, 2015, Madras et al., 2018]. In the latter applications, adversarial learning is frequently leveraged as an engine for (mutual) information-minimisation – or *infomin* – where the information to be minimised is that related to the domain or sensitive group in DA/DG and FairML, respectively. For this, both players take the form of an NN with strategies defined by their parameters – $\theta_{min} \in \Theta_{min}$ and $\theta_{min} \in \Theta_{max}$ for the learner and adversary, respectively – which they play according to their respective architectures, together constituting the actions $a_{max} : \mathcal{X} \times \Theta \rightarrow \mathcal{Z}$ and $a_{min} : \mathcal{Z} \rightarrow \mathbb{R}$. We note, incidentally, that due to the continuity and non-convexity of NNs, it is unfortunately not possible to guarantee the existence of Nash Equilibria for the resulting games, only ones that are locally defined [Unterthiner et al., 2018,]. The game in question can then be couched in terms of the following countervailing objectives:

- **adversary**: maximise the likelihood of a correct determination of the true value of s associated with given a input x, while having access to only the censored version (representation) of the input produced by f_i.
- **learner**: create a censored version of the input, z, that maximally minimises the amount of information about s determinable from it.

This game then gives rise to the following minimax objective function, for some dataset $\mathcal{D}$ made up of pairs of inputs, x, and attributes to be censored, s

$$\min_{\theta_{min} \in \Theta_{min}} \max_{\theta_{max} \in \Theta_{max}} -\mathbb{E}_{(x,s)\sim\mathcal{D}}[\ell(a_{max}(a_{min}(x, \theta), \psi), s)], \tag{15}$$

where we have negated the expectation (converting the loss into utility) to remain consistent with the idea of the adversary being the maximiser and defined the optima over the parameter spaces to make clear the idea that the parameters define the chosen strategies. When s is discrete, ℓ is typically taken to be the standard cross-entropy loss; in this case the fixed point of the objective function is attained when the outputs of a_{max}, with codomain the appropriate probability simplex, are maximally entropic (the derived predictions no better than random), which one hopes holds for all $\sigma_{max} \in \Sigma_{max}$ and connotes the invariance $Z \perp S$. Having the learner play this game without any additional objectives (maximising for utility w.r.t. the task of interest) is, of course, inauspicious if the goal is to have a representation that is useful for some task (other than foiling the adversary) – a trivial solution on the part of μ_{min} would, for instance, be to simply ignore the input and output a constant representation. In many cases, however – especially those arising in FairML and DA – there is competition not only between the adversary and the learner, but between the objectives themselves w.r.t. the latter alone.

Practical challenges

Assume, for instance, that the task of interest is a classification one, such that we have on top of the infomin objective defined Eq. 15 an *infomax* objective w.r.t. the ground-truth label y. Only if the condition $\mathcal{I}(S; Y) \approx 0$ (i.e. the target labels are uniformly distributed across $\mathcal{S}$) holds can the infomin and infomax objectives be simultaneously satisfied, which is to say, in optimisation terms, that the inner product of the gradients of the two respective losses (w.r.t. θ_{min}) is consistently non-negative and there is no trade-off, governed by the preference direction, leading to a Pareto front and suggesting treatment with MOO methods. More generally, unconditional infomin is problematic whenever there is conditional shift between the training (source) and test (target) sets, as anatomised by Zhao et al. [2019] in the context of UDA. Given full observability of y and s and consistent support of their joint distribution, i.e. $\text{supp}(\hat{P}^{tr}(S, Y)) = \text{supp}(\hat{P}^{te}(S, Y))$, one can realign the objective by computing the infomin component class-conditionally (practically, importance weighting based on the empirical distribution $\hat{P}^{tr}(Y)$). When this observability does not hold, generally or for a subset of the data (the target domain in the case of UDA), then matters are complicated, however, and approximations are required (e.g by clustering). In Chapter 4 we consider a problem of this nature and recast the problem as one of aligning the supports of the training – which is systematically missing certain combinations of s and y – and deployment – which is presumed unlabelled, as in UDA – sets.

We have already alluded to some of the challenges entailed in adversarial infomin approaches; here, we summarise the two most salient ones:

Approximation errors

1. The strength of information-minimisation is proportional to the strength adversary used to drive it – a fixed point attained by one adversary is not guaranteed to hold for any other adversary (differing in architecture, optimisation scheme, etc.) unless the fixed point corresponds to the desired invariance. Theory dictates that one computes the best response of each player at each iteration, however this is generally infeasible when

working with models with many parameters and datasets with many samples. Thus, approximations are required – e.g. by limiting the number of updates per iteration and by bootstrapping the best response from the previous one – but these can lead to unstable training dynamics or entrapment in bad optima. Indeed, a number of recent studies have shown that many adversarial approaches fail to faithfully produce infomin representations when probed [Moyer et al., 2018, Feng et al., 2019, Balunović et al., 2021].

Cyclic dynamics

2. Adversarial setups are generally susceptible to cycles in strategy space, e.g. where two players repeatedly switch between two non-NE strategies because doing so is mutually locally optimal. In reinforcement learning, for instance, this has led to the development of fictitious self-play algorithms [Brown, 1951, Heinrich et al., 2015, Vinyals et al., 2019] wherein each player has its best response computed against a uniform mixture of past opposing strategies (this mixture provably converging to a NE); while this allows for avoiding the aforementioned cycles it comes at a significant computational cost. In Chapter 3 we consider this pitfall and propose a middle ground of training against a stochastic ensemble of adversaries.

12 Invertible neural networks

Bijectivity

Chapter 3 of this thesis explores the application of invertible neural networks (INNs) to fair-representation learning and so will afford some brief discussion to their basic workings here. An INN [Kobyzev et al., 2020], as the name suggests, refers to any neural network for which both the usual forward mapping, $f(\cdot)$, and its inverse $f(\cdot)$ are defined, with the assumed property that both are differentiable and as such that the function belongs to the class of *diffeomorphisms*, $f \in \text{Diff}(\mathcal{X})$. Thus, we have a function that is an invertible bicontinuous map from input space, $\mathcal{X} \subset \mathbb{R}^d$ to latent space $\mathcal{Z} \subset \mathbb{R}^d$, noting that the domain and codomain are equidimensional, as presupposed by the function's bijectivity. It is obvious, but nonetheless worth stating, that for f is composed of subfunctions $f \triangleq f_L \circ \cdots \circ f_2 \circ f_1$ and each individual subfunction is diffeomorphic, then f in its totality, also satisfies this property, allowing us to build arbitrarily complex INNs by chaining together layers defining these subfunctions.

Normalising flows

The usual bailiwick of INNs is density estimation – and by complement, generative modelling – due to their hallmark diffeomorphic property that make it possible for densities under the models to be calculated *exactly*, in contrast to variational methods that only do so up to a lower bound (the so-called ELBO). This calculation is enabled by the change-of-variables theorem, allowing one to track how the density of the distribution changes as the INN warps a known (and tractable) base distribution into a complex, highly-multimodal, one. Like with variational auto-encoders [Kingma and Welling, 2014], the base density, $P(Z)$, is generally taken to be an Isotropic Gaussian distribution; the posterior density, $P(X)$, 'flows' through the network – in a manner reminiscent of a Galton Board

– into this normalised base density, earning this class of methods the name *Normalising Flows* (NFs; Rezende and Mohamed [2015], Kobyzev et al. [2020]). Practically, for a given sample x, its log-likelihood under the INN f, with base density $\mathcal{N}(\cdot; 0, \mathbb{I})$ the aforementioned Gaussian distribution, can be computed as

$$\log P(X = x) = \log P(Z = z) + \sum_{l=1}^{L} \log \left| \det \left(\frac{\mathrm{d} f_l}{f_{l-1}} \right) \right|,$$
$$P(Z = z) = \mathcal{N}(z; 0, \mathbb{I}),$$

and training the model simply amounts to maximising this quantity over the empirical training distribution in the usual fashion. As with GANs and VAEs, to sample from $P(X)$, one needs only draw a random sample from the corresponding base density, $z \sim P(Z)$, and push that sample through f.

Losslessness

In Chapter 3, however, it is not the foregoing density-estimation capabilities of INNs that we are interested in, rather the diffeomorphic property itself, insofar as it guarantees the learned representations are *lossless* w.r.t. the inputs, as well as a means to visualise the factors of said representations due to its having an exact inverse (whereas auto-encoders have only an approximate inverse (the decoder) that must be trained separately from the encoder). That is to say, while f may deform manifold $\mathcal{X}$ in arbitrarily non-linear ways, since each point is mapped uniquely from the domain to codomain only the form of the information contained in the input can change, not its extent. This is in contrast to conventional architectures that define *surjective* mappings that embed inputs into spaces much smaller than $\mathbb{R}^d$ (in line with the *Manifold Hypothesis* [Fefferman et al., 2016]). Other works have also capitalised on this information-preservation explicitly, e.g. both Hoogeboom et al. [2019] and Xie et al. [2021] explore the natural suitability of INNs for lossless image compression. Contrastingly, in work postdating that done in Chapter 3, normalising flows have been applied applied to FairML problems with the insight that one can leverage the exact-density computation to define define an optimal adversary [Balunović et al., 2021, Cerrato et al., 2022]. This allows for obtaining provably fair representations while also obviating the optimisation challenges that accompany (parametric) adversarial training, though at the cost of an independent INN for each of the sensitive groups.

Practical drawbacks

As discussed, INNs have a number of unique and compelling properties that would seem to make them the choice method for many generative purposes; INNs do have their share of practical shortcomings, however. Notably, bijectivity does not come at a cost; while there are some ways of mitigating it, such as factoring out parts of the representation at intermittent steps [Hoogeboom et al., 2019], one is constrained to having a latent space that is equidimensional to the input space and when the latter is large, as in the case of images, training an INN can be computationally challenging. Conventional architectures do not suffer this problem as they can make free use of coarsening (downsampling) operations throughout their extent. This drawback is further compounded by the fact that the layers making up an INN are necessarily less expressive than their invertible counterparts and thus more them are needed to achieve comparable levels of expressiveness in composition. The coupling layers that constitute the atomic building blocks in Dinh et al. [2014] restrict their non-linear, non-invertible functions, to only a subset of the input dimensions at a time so that the layer as a whole remains invertible, thus limiting the degree to which interdependencies between the input dimensions can be modelled. Finally, without proper parametric constraints (e.g. to be bidirectionally K-Lipschitz), the optimisation of INNs can be prone to instabilities that can render them *numerically* non-invertible, despite their design, and thus invalidate computations made according to Eq. 12 [Behrmann et al., 2021].

References

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2016.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Eric Bair. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5):349–361, 2013.

Mislav Balunović, Anian Ruoss, and Martin Vechev. Fair normalizing flows. *arXiv preprint arXiv:2106.05937*, 2021.

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

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Jörn-Henrik Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1800. PMLR, 2021.

A. Ben-Tal, L. El Ghaoui, and A.S. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

George W Brown. Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, 13(1):374, 1951.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR, 2019.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

Mattia Cerrato, Marius Köppel, Alexander Segner, and Stefan Kramer. Fair group-shared representations with normalizing flows. *arXiv e-prints*, pages arXiv–2201, 2022.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 33:18860–18871, 2020.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2018.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

Kalyanmoy Deb and Kalyanmoy Deb. Multi-objective optimization. In *Search methodologies: Introductory tutorials in optimization and decision support techniques*, pages 403–449. Springer, 2013.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, 1996.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049, 2016.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341*, 2019.

S Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 13(1):57–64, 1967.

Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

A Gammerman, V Vovk, and V Vapnik. Learning by transduction. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 148–155, 1998.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *International conference on machine learning*, pages 805–813. PMLR, 2015.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Emiel Hoogeboom, Jorn Peters, Rianne Van Den Berg, and Max Welling. Integer discrete flows and lossless compression. *Advances in Neural Information Processing Systems*, 32, 2019.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209, 1999.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2444–2453, 2018.

Louis Kaplow and Steven Shavell. The conflict between notions of fairness and the pareto principle. *American Law and Economics Review*, 1(1):63–77, 1999.

Jana Kasperkevic. Google says sorry for racist auto-tag in photo app. *The Guardian*, 1:2015, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent tradeoffs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Yong Lin, Shengyu Zhu, and Peng Cui. Zin: When and how to learn invariance by environment inference? *arXiv preprint arXiv:2203.05818*, 2022.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, 31, 2018.

Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the pareto front with hypernetworks. In *International Conference on Learning Representations*, 2020.

Judea Pearl. From bayesian networks to causal networks. *Mathematical models for handling partial knowledge in artificial intelligence*, pages 157–182, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34: 1256–1272, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

Yoshikazu Sawaragi, HIROTAKA NAKAYAMA, and TETSUZO TANINO. *Theory of multiobjective optimization*. Elsevier, 1985.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6):863–879, 1992.

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.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. *arXiv preprint arXiv:2110.03095*, 2021.

Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Martin Szummer and Tommi Jaakkola. Partially labeled classification with markov random walks. *Advances in neural information processing systems*, 14, 2001.

Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb gans: Provably optimal nash equilibria via potential fields. In *International Conference on Learning Representations*, 2018,.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.

Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Ke Alexander Wang, Niladri Shekhar Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? In *International Conference on Learning Representations*, 2021.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

Yueqi Xie, Ka Leong Cheng, and Qifeng Chen. Enhanced invertible encoding for learned image compression. In *Proceedings of the 29th ACM international conference on multimedia*, pages 162–170, 2021.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Understanding why generalized reweighting does not improve over erm. *arXiv e-prints*, pages arXiv–2201, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.

Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2014.