SEMI-SUPERVISED METHODS FOR DISTRIBUTIONALLY-ROBUST LEARNING

MYLES BARTLETT

A thesis submitted for the degree of Doctor of Philosophy.

School of Engineering and Informatics
University of Sussex

April 2023

Dedicated to all the fabulists that shaped the story that I am, the stories I have told, and all the stories I hope yet to have and tell, and to you the reader, without whom this story would be made but not alive.

ABSTRACT

Over the last decade, deep learning has paved the way for machine-learning systems to achieve unprecedented levels of performance on myriad real-world tasks by learning complex decision rules directly from large quantities of data. However, the data-driven, and fundamentally statistical – rather than causal – nature of these systems is a double-edged sword that has led to catastrophic, difficult-to-diagnose, failures in various safety-critical and ethically-sensitive applications. Indeed, worryingly often they learn to exploit *shortcuts* that are extraneous w.r.t. the underlying task. This is said to be a problem of *shortcut-learning*; the first two papers constituting this thesis tackle different manifestations of such with appropriately different solutions, both predicated on leveraging additional data to obviate biases immanent in the training data. In the first, we consider a setup in which sampling bias induces a one-to-one correspondence between subgroup and target; since learning from the training data alone is ill-posed, we propose a two-stage, interpretable framework exploiting the unique properties of invertible models, founded on the assumption that more-diverse unlabelled data is often readily obtainable, e.g. from censuses. In the second, we relax the aforementioned problem such that subgroup-target combinations are missing in an asymmetric fashion, w.r.t. the target, and propose to solve this by matching the support of the training set with that of an unlabelled dataset representative of the test set. This is realised through a combination of clustering, adversarial training with a set discriminator, and a hierarchical-sampling strategy. The third and final paper contends with the broader problem of generalising to data from *domains* outside the training distribution and how unlabelled data from extra domains can effectively further this goal. Here, changes in domain induce natural distribution shifts corresponding, for example, to variations in lighting, perspective, and environs; to counteract them, we propose a simple, consistency-regularised approach based on causal-matching.

DECLARATION

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree. Except where indicated by specific stated in the text, this thesis was composed by myself and the work contained therein in my own.

Brighton, April 2023

Myles Bartlett

ACKNOWLEDGEMENTS

I would thank my family for being eternally-supportive of me and ever-tolerant of my heteroclitic ways, to all members of PAL – and among them most of all Thomas and Oliver, my once officemates and hopefully-forever friends (for what tempers friendship stronger than the burning together the midnight oil (or, more accurately, 'depleting the packet of custard creams') on the eve of a deadline), without whom I may not have made it through to where I am, or even begun the journey in the first place – and to my supervisor, Novi, for having faith in me when I had none in myself and putting up with me at times when I defied understanding. I would also give specials thanks to the European Research Council (ERC) for partial funding of my work through the BayesianGDPR project (Grant agreement ID: 851538).

CONTENTS

ACRONYMS

AE auto-encoder

AF algorithmic fairness

AdvL adversarial learning

CBN causal Bayesian network

CNN convolutional neural network

DAG directed acyclic graph

DA domain adaptation

DG domain generalisation

DL deep learning

DNN deep neural network

DP demographic parity

DRO distributionally-robust optimisation

DR distributional robustness

ELBO evidence lower bound

ERM empirical risk minimisation

ER equalised rate

EqOd equalised odds

EqOp equality of opportunity

FPR false positive rate

FRL fair-representation learning

GAN generative adversarial network

HGRMC Hirschfeld-Gebelein-Rényi maximal correlation

ID in-distribution

IF individual fairness

INN invertible neural network

IW importance weight

KL Kullback-Leibler

MI mutual information

MLE maximum likelihood estimation

MLP multi-layer perceptron

ML machine learning

MMD maximum mean discrepancy

MOO multi-objective optimisation

MS missing source

NF normalising flow

OOD out-of-distribution

PF Pareto frontier

PO Pareto optimal

RLAIF reinforcement learning from AI feedback

RLHF reinforcement learning from human feedback

RobAcc robust accuracy

SCL shortcut learning

SC spurious correlation

SGD stochastic gradient descent

SL supervised learning

SelfSL self-supervised learning

SemiSL semi-supervised learning

TNR true negative rate

TPR true positive rate

UDA unsupervised domain adaptation

UL unsupervised learning

VAE variational auto-encoder

cFlow conditional flow

cVAE conditional VAE

GLOSSARY

CelebA A popular benchmark computer-vision dataset comprising more than 200K celebrity head-shots annotated with various physical and affective attributes, such as 'Smiling', 'Gender', and 'Age'. The dataset has cemented itself as one of the principal benchmark datasets in the algorithmic fairness (AF) and distributionally-robust optimisation (DRO) literature due to the spurious correlations (SCs) consequent of its (conditional) label-imbalance ('Blond' individuals being predominantly female, for instance).

Domain The 'source' of a particular subset of the data, each domain assumed to embody a different sub-distribution of the collective data induced by, for instance, variations in recording equipment, category (in the context of sentiment analysis), environs, geography (in the context of remote sensing), and weather conditions. Thus, like with sensitive attribute, the domain imposes some secondary structure on the data and in the context of domain adaptation (DA) and domain generalisation (DG) I will talk of desiring invariance to this structure, a desideratum that aligns with that of AF (at least in certain senses). I would note that I also use, with some frequency, **domain** in the functional-analysis sense, referring to the space of expected input values (correlatively with codomain), however, it should hopefully always be clear from the context which sense I am invoking.

Environment See **domain**.

MNIST A dataset of handwritten digits (itself derived from a larger dataset known as the NIST Special Database) and a foundational computer-vision benchmark that continues to see use even to this day under a variety guises, despite its simplicity (this simplicity, making it well-suited for demonstrating feasibility). In this thesis, I will talk of a colourised version of the dataset, in both Chapters 3 and 4 – variants of which have seen widespread adoption in the DG literature, most notably in (Arjovsky et al., 2019) – where the colourisation is performed correlatively with the digit labels so as to induce a SC.

Protected attribute See **sensitive attribute**.

Sensitive attribute An attribute of the data that, for legal or ethical reasons, should not be factored into the predictions of a model, and w.r.t. fairness metrics are computed to judge the degree of this violation.

Subgroup The general term (void of any subfield-specific connotations) for sensitive attribute and domain (and, transitively, their synonyms), and thus referring to a given secondary group, partition, or sub-distribution of the data.

UCI Adult (Income) A popular tabular AF dataset derived from U.S. census data; the canonical task is to predict whether an individual earns $50K or more (positive class) or not (negative class) and the sensitive attribute, w.r.t. which fairness is computed, is typically taken to be 'Sex' 'Race', or 'Age'.

WILDS A suite of datasets (spanning various tasks, domains, and modalities) for evaluating the distributional-robustness/DG capabilities of models under in-the-wild (real-world) distributions shifts.

Part I

BEGINNING

This part comprises the introduction to the thesis – wherein I set the scene for what is to come and the order in which that is to occur – along with some rudimentary background on those topics thought relevant to Part II.

1 | INTRODUCTION

"It is good to have an end to journey toward; but it is the journey that matters, in the end."

The Left Hand of Darkness
Ursula K. Le Guin

Preface

It is well-accepted that every good story should have a beginning, a middle, and an end; for if a story had not those things it could scarcely be called a story at all, at most it would be a nonsense one and nonsense is only sens*ible* when founded on *able* sense. This is where this thesis begins, a thesis that hopefully satisfies some of the reader's sensibilities regarding what makes a good story; at the very least I hope it bears some scintilla of sense, even perhaps an estimate of erudition. When I say *story* I mean not, of course, to say that the contents are in any way fictitious or embellished; as a story, this is a work of chapters and bridging those chapters is evolution, both academically and personally. For a story to *become* – to tell itself or let itself be told – it must grow, by nature, by contrivance, by necessity; stories reflect life and life is a process of growth, of betterment – where each step carries us onward on a journey, one without a destination, but a journey one should never yield on regardless of the times one stumbles. I am not ashamed admitting that the journey paved by this thesis itself was marked by many such stumbles, by many foibles and follies, times when I felt I could not stand again for the weight of the past and the murk of the future; I cannot say that I conquered all, if any, of my frets and fears and failings, but I can at least say that I forged on and became *better*, not by own mettle alone but as much by the support of those who have staunchly companioned me each step of the way, through summer-sweetened meadow and gloom-drenched thicket alike. I would also call this thesis a *story* simply for my fondness of stories, for all that they might teach us about others, and, most of all, ourselves – by exploring fabulous *other* worlds we also come to better know the inner one; to want to be fond of something one has given so much to is, I think, natural and should there be any lesser elements to it, I can fancy them, as in some dualistic tale, the darkness that makes the light shine all the brighter.

The thesis of stories and the journey without an end

On the sea of themes

Every story has a theme: an unbroken thread that weaves all into one. This story is no exception; its theme is not one of valour, of defiance in the face of impossible odds, or of taking the next step

when the path ahead is fogged and the path behind beset with demons – indeed, the theme is not quite so elevating – having the power to rouse our best selves – but it bears its own importance, nevertheless – not to the human condition but rather to the autonomous one.

A theme to connect all other themes

This is a machine learning (ML) thesis and the themes are appropriately related to ML; if there is one central theme that unites all themes across all chapters, it is the *disconnect* between the *statistical* nature of ML and the *causal* nature of reality – to mistake the notion of *correlation* – born from the former – with the notion of *causation* is a great fallacy, as every new Statistics student learns before most else. That is to say, if two events, X and Y, routinely coöccur, with the former preceding the latter, they are correlated, in that one can use the occurrence of X to predict the occurrence of Y at an above-chance rate – such is the indispensable utility of statistics, as the quantification of *patterns* (and ML is but sophisticated *pattern recognition*) – but one cannot reasonably extend this deduction that 'X *predicts* Y' to the much deeper one, 'X *causes* Y' without employing interventions with the view to eliminate confounding factors, i.e. factors causing both (and thereby correlating) X and Y. I should emphasise – to stem right away any misconceptions about what this thesis is or aspires to be – that despite their being rooted in it, the themes of this thesis are not *of* causality itself, by which I mean that, while I may use the calculus of causality to characterise and interconnect phenomena, I never venture to solve the formidable problem of causal discovery head on, only those problems emanating from it, representing the aforementioned disconnect.

The simplest explanations are usually the best ones, and by *best* I mean *true* (or approximately so) when it comes to *reality* – indeed, such was the consummate genius of Newtonian, Einsteinian and Darwinian theories to collect myriad related-but-seemingly-distinct phenomena within unified models of startling (relatively) simplicity. However, what is simplest in a statistical sense, with respect to a finite representation of reality – a *dataset* – does not always – and often does not – align with what is simplest in the general sense that we can presume *true*. Modern physics frames the universe in terms of *symmetries* (as permitted by Noether's celebrated theorem (Noether, 1918)), in terms of what changes and what does not change subject to a particular action or group of actions or *interventions* – *variances* and *invariances* – and this same notion, one of *modularity of subsystems*, is naturally expressible under a causal (and group-theoretic) framework. Through this lens, the *best* model is the one that completely accounts for all observations as a function of the fewest variables – one that is maximally invariant or modular.

This thesis is not a story of learning a true, causally-complete, world model – for that would be too lofty a goal – however, the notion of invariance – to specific concepts inducing specific (open or closed) sets of transformations – is at the heart of all problems broached – each on its own quite humble. 'Concepts' is, of course, a rather nebulous term but generality presupposes a degree of nebulousness; in relation to the works contained herein, I specifically and concretely mean, for instance, some indicator of group-membership, or the site, or context, of collection (the *domain*), though the exact nature of said concepts is indeed both conceptually and practically (as far as the methods I introduce are concerned) arbitrary.

The reader has assuredly heard of the many astounding feats accomplished by ML in the last decade, borne on the winds of the deep learning (DL) revolution heralded by Krizhevsky et al. (2012) and mediated by advances in highly-parallelisable hardware. Indeed, the statistical-learning paradigm has given rise to, no less than, autonomous agents capable of new, more-efficient

algorithms for age-old mathematical problems (Fawzi et al., 2022); of deeply comprehending language – in all its daedal complexity – and in turn generating it with remarkable coherence, consistency and – on occasion – expert-level insight (Brown et al., 2020); of going toe-to-toe with, or even trouncing, the most adept players of the most strategically- and physically-demanding games conceived by humanity (Silver et al., 2017; Berner et al., 2019; Vinyals et al., 2019; (FAIR)† et al., 2022). Yet, where there is light there is shadow, and for every marvel there is a misstep; the annals of ML present no exception to that. For all the mystique that enshrouds it, ML is (as I have mentioned in passing), au fond, *statistical* modelling: an ML algorithm ingests a set of data, collected via some (generally imperfect) mechanism, and models correlations between the covariate and response variables (and between the covariate variables themselves) in some higher-dimensional space, assuming for the sake of simplicity (and consistency with the works in this thesis) a *supervised-learning* task (and if otherwise then the response variables are but some (dynamic) subset of the covariate variables).

A misstep for every marvel: the problem with correlations

There are *correlations* that are causally-supported, and these are the correlations one hopes are learned by one's ML algorithm of choice, because they are *real* and thus *generalisable*; there are also those that are *not* causally-supported but are nonetheless present due to deficiencies on the data-collection (or data-curation) side – I will refer to such correlations as being *spurious* henceforth. Of course, I am being 'nebulous' again when I say 'deficiencies', and again it is for generality's sake, for this covers anything from systemic bias – as is the remit of AF – to insufficient coverage (geographically, demographically, etc.), as a result, for instance, of constrained resources – as is more the remit of DA and DG, though the lines between these and AF are often blurred. These deficiencies have led to a spate of (in)famous cases, within both academic and journalistic spheres, igniting public concern and redoubling research efforts to allay said concern. To illustrate: a 2018 investigation (Dastin, 2018) into the automated-hiring system trialled by Amazon revealed said system to vastly prefer male candidates to female ones for software-development roles based on their résumés; the same year, Buolamwini and Gebru (2018) investigated the behaviour of three commercial gender-classification algorithms w.r.t. different skin types ('lighter' vs'darker') and observed marked disparities in the resulting accuracies between lighter-skinned and darker-skinned subjects, and between male and female subjects, with a compounding of the two trends; again the same year, Zech et al. (2018) demonstrated that ML classifiers can predict incidences of pneumonia with near-perfect accuracy based on only site-specific tags and the prevalence rates associated thereof. Such systems can be said to *not* be *distributionally robust* as they either fail to generalise beyond the training distribution or to exhibit approximate performance-parity between sub-distributions even within the training data, corresponding to different concepts (gender and race in the foregoing examples).

The knots and nots of data and bias

The solution – the sword to this Gordian knot – seems a staggeringly simple one, so simple as to perhaps invite the reader to question the whys and wherefores of this thesis and cry: 'just collect more/better data!' Unfortunately, if this best-of-all-possible-worlds were so perfect as to always provide us this knot-cutter, there would likely be no need for it in the first place; in this imperfect world, we often have little choice but to make the best of what we have. There is little hope, for instance, of deriving diverse, bias-free data from any significant population of people given how rife, and arguably intrinsic (from an evolutionary-psychology perspective; Kurzban and Leary, 2001), out-group biases (conscious or subconscious) are in us humans. There

is also little hope of capturing, by camera trap, every species within a given region under every condition and in every locale given that different animals have different ecologies (nocturnal vs. diurnal, being a clear distinguisher) and there are only so many devices one can afford to place and thereafter monitor; it may not even be possible to capture every locale alone (but wish to later generalise to the thitherto-unseen locales) and in such case we have a pure DG problem. While we cannot correct the problem at the source, we can intervene to *mitigate* the downstream biases; given that we know what these biases are, we can seek to be *invariant* to their cause. Thus, in AF, we can (depending on our definition of 'fairness') couch our desideratum as learning a predictor that is invariant to (does not take into account) the designated sensitive attribute(s) – race, gender, and age being the usual candidates; in DG, it is the eponymous *domain* one targets for invariance, this being the locale in the forgoing camera-trap example.

I have used the all-encompassing word 'data' above to mean 'annotated data', for it is the in the context of supervised learning that I speak of these things, a context wherein the task, and thus the function to be approximated, is defined by annotations, or at least some target attribute, thinking of the tabular data where there may not be such a clear divide between the annotations and the annotated (in contrast with the prototypical example of image-classification where the inputs are composed of pixels and the annotations are human-conferred labels). While 'annotated data' is often indeed scarce due to reasons, inter alia, of a budgetary, geographic, or inability-to-travel-retrograde-along-the-fourth-dimension nature, data of the 'unannotated' (or 'partially annotated') variety is more readily obtained, and it is this reality by which the embers of unsupervised learning (UL) have been stoked, into what has become a bustling blaze of research in but the short time since this thesis was begun. This leads us, hopefully not-too-word-weathered, to the final of port in our thematic voyage: though this theme is the first written of in the title – 'Semi-Supervised Methods for…' – it is the last I shall write of here. In this thesis, I entertain two particular motives for using supplementary, unannotated data, unified in their premise of enabling invariance to 'concepts' of interest. Since the unannotated data is being used concurrently with the usual task-defining, annotated data, we tread in semi-supervised learning's (SemiSL's) domain. The aforementioned motives are as such: First, while for a bias-bearing dataset the desired invariances may not be learnable from the annotated data, due to statistical entanglement between the target and concept variables, they may be so – or to a greater extent, at least – given better-covering and more diverse, but unannotated, data; second, in the context of DG, one may view the domains as defining a finite perturbation set, informing our model of the types of invariances it need learn in order to generalise to domains yet unseen; by augmenting this perturbation set with said unannotated data we may also aspire to augment the robustness of said model.

Charting the course ahead

This is a thesis of three parts and six chapters. As this is the first chapter, it is necessarily the beginning yet, at the same time, it is not the end of the beginning. In the subsequent chapter – what remains of the beginning – I provide background on the main topics I think relevant to the works that constitute the 'middle' and marrow of this thesis. That second chapter has its own

introduction and so I will not dwell here on the precise nature of its contents. The 'middle' is a part of three chapters, each chapter corresponding to a distinct paper and a distinct problem; these papers appear in an order that is chronological yet also most thematically-contiguous. In the 'end', I discuss the works holistically, both in the context of one another and in the context of more recent developments in the germane fields; taking stock of this, I also ponder future avenues of work in similar vein. To help orient the reader, I adumbrate below the three 'middle' chapters of the thesis, stating in each case their motivations, methods, and merits. I will shift to using 'we' here, in self-reference, as all these works were done collaboratively, yet I claim credit enough (as first or second author) to feel deserving of their ownership and of their inclusion in this thesis, a text that represents work of my shaping and doing. To substantiate this claim, I include at the end of each of the corresponding chapters an estimate of the contributions made by myself and each coäuthor.

Chapter 3

In this first of the middle chapters, we tackle what we term an *aggravated fairness problem*, characterised by strong SCs between the sensitive and target attributes, to the extent of one-to-one correspondence which is not maintained at test time. This is, consequently, a shortcut learning (SCL) problem in which the former attribute serves as a proxy for the latter attribute, by virtue of its lower-complexity; to ensure generalisability across the spectrum of intersectional groups, we require a model that learns the correct, causal mapping for the features to the target, which is to say that is invariant to the sensitive attribute, though we are harried by the entailed problem of identifiability. Similar characterisations had been considered in the DG literature (Arjovsky et al., 2019; Jacobsen et al., 2019), though we approached and, argued for the validity of, the problem from an AF perspective, partly motivated by the cases of systemic censoring adduced in Kallus and Zhou (2018). Given the intractability of disentangling the two attributes, we assume the existence of a supplementary dataset that contains all intersectional groups but is annotated only partially, in the sense that the target annotations are absent – we assume such data is more easily obtained than fully-annotated data, with census data being one potential source, for example. This gives rise to a kind of transfer-learning setup in which the quality to be transferred is invariance to the sensitive attribute for which we propose an interpretable framework exploiting the unique properties of invertible models, in particular their losslessness connoted by their bijectivity and their namesake exact invertibility (whereas more traditional approaches based on auto-encoders (AEs) furnish only approximate invertibility).

We demonstrate that these unique properties are practically meritable, giving rise to models that perform more robustly over a range of datasets and degrees of correlations, compared with baseline models, and especially so in full transfer-learning scenarios where the partially-annotated and fully-annotated datasets are drawn from disparate distributions sharing a sensitive attribute. Moreover, the exact invertibility allows us insight into what the model has learned and diagnose potential failures, such as unforeseen entanglements between the sensitive attribute and certain dimensions of the inputs.

Chapter 4

The above work required only partially-annotated unbiased data but annotated data nonetheless and there are realistically many cases where this requirement may be preclusive. Accordingly, in Chapter 4, we entertain a relaxed version of the problem – this time, couched in non-AF-centric manner – which is soluble using supplementary data that is unannotated in the truest sense, in that neither the sensitive – here, 'subgroup' – nor target annotations are provided. The problem is still one of SCL yet the biasing is imposed in a hierarchical (with the targets – specifically classes – constituting the top level the implied tree, the subgroup the bottom level) and asymmetric fashion such that the identifiability is possible; its general formulation – which admits the problem from the previous chapter as a special case – is as much a contribution of the chapter as the solution we ultimately propose to it, a solution predicated on the idea of aligning support rather than distributions as historically practised in DA (Ben-David et al., 2006). By 'asymmetric' I mean that for we observe all subgroups and classes expected at deployment time and for at least one class we observe more than one subgroup. In this context, we refer to the intersectional groups – target-subgroup combinations – as 'sources' and the aforementioned problem, characterised by their missingness, as one of 'missing sources'. The problem is strongly redolent of the classic unsupervised domain adaptation (UDA) one, in that one has access to unlabelled data from a *target* domain, distributionally-shifted relative to the training data, and seeks to maximise positive transfer, or adaptation, to that test data from that same domain. However, the distinction lies in the missing-sources problem being of a hierarchical/class-conditional nature, whereas in UDA there would be entire subgroups missing from the training data yet at the same time no shortcuts between subgroup and targets induced.

To accomplish the alluded-to support-alignment, or *support-matching*, we look to semi-supervised clustering to estimate the sources in the unannotated dataset, or *deployment set*, which we assume to be source-complete w.r.t. the test set, and may even be the test set itself in a transductive setting. With the estimates in hand, we proceed to use a hierarchical-sampling procedure to construct batches from the training and deployment sets representative of their respective support over the sources, training an encoder to generate representations of them that are *dataset*-invariant by means of an adversarial set-discriminator. We find this approach can generate, and with surprisingly swift and stable convergence, invariant representations in a way that is robust to the approximation-error incurred by clustering, significantly more so than instance-wise and supervised (using clusters for balancing and as direct targets) baselines.

Chapter 5

In the forgoing chapters, we considered setups in which the concepts (sensitive subgroup attributes for Chapters 3 and 4 respectively) constituted a closed set, which is to plainly say that their possible test-time values were known and represented – in some form, annotated, partially-annotated, or entirely unannotated – at training time. We also assumed that the target played a role in the distribution shift that consequently corresponded to one of target and subpopulation/subgroup/domain (all these terms being synonymous here and throughout the literature, differing only by context)

shift combined yet could not be treated with conventional methods, such as reweighting, due to the emergence of spurious correlations. The focus of this third chapter, in contrast, is not on any spurious component but on the domain-shift one, such that is the well-established problem of DG that we tackle, albeit with a semi-supervised learning (SemiSL) slant introduced by then-recent benchmarks (Sagawa et al., 2022). We may distil this focus into the motivating question "Given annotated training data drawn from a finite set of domains, disjoint from those of the test set, might we use additional unannotated data from again disjoint [w.r.t. both training and test sets] domains to improve generalisation [to the aforesaid unseen domains]?"

The springboard for our proposed method was statistical-matching algorithm developed by my coäuthors, as appearing in Romiti et al. (2022), that pairs samples from different (in this case) domains based on certain (robust) statistical criteria. In the context of Romiti et al. (2022), this algorithm is applied in a pseudo-post-hoc fashion to construct a 'patched' dataset, with which the initial model is trained anew in a distributionally-robust manner. In the context of the chapter-being-discussed, this algorithm is generalised and embedded in an online-learning (online) framework – motivated by improved efficiency and the finding that the original two-stage (offline) framework proved insufficient for problems of the kind in question – with the generated matches used to define matches for a consistency-regularised SemiSL objective that is task- and modality-agnostic. Here, the use of the term 'online-learning' is perhaps somewhat-Pickwickian, for I refer not to the algorithm *in toto* but to the matching component of it specifically, viewed from a bilevel perspective: the former, as characteristic of *offline learning*, may draw upon any given sample an arbitrary number of times and draw upon the samples collectively and in an unprescribed order (that is, with full control over the sampling mechanism; in the online-learning *seteup* the sequence of samples is prescribed and as such may not, and often will not, be i.i.d.); the latter, conversely, observes only a subset of the dataset (with some amortisation) at each match-generation step – this is in contrast to Romiti et al. (2022) where the matching is performed over the full dataset in a post-hoc fashion. Thus, one should understand the meaning of 'online' to be akin to 'bootstrapped' and, indeed, throughout the paper we use the terms near-interchangeably. This aside complete, we find the resulting algorithm consistently outperforms baseline methods, including fully-supervised ones which Sagawa et al. (2022) showed many existing SemiSL methods failed to accomplish, and, as in Chapter 3, it comes with a valuable interpretable aspect, conferred by the trained-with matching algorithm.

BIBLIOGRAPHY

Noether, E (1918). 'Invariante Variationsprobleme'. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1918, pp. 235–257.

Kurzban, Robert and Mark R Leary (2001). 'Evolutionary origins of stigmatization: the functions of social exclusion.' In: *Psychological bulletin* 127.2, p. 187.

Ben-David, Shai, John Blitzer, Koby Crammer and Fernando Pereira (2006). 'Analysis of Representations for Domain Adaptation'. In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt and T. Hoffman. Vol. 19. MIT Press. URL: https://proceedings.neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C.J. Burges, L. Bottou and K.Q. Weinberger. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Silver, David et al. (2017). 'Mastering the game of go without human knowledge'. In: *nature* 550.7676, pp. 354–359.

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

Dastin, Jeffrey (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women'. In: *Ethics of data and analytics*. Auerbach Publications, pp. 296–299.

Kallus, Nathan and Angela Zhou (2018). 'Residual Unfairness in Fair Machine Learning from Prejudiced Data'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 2444–2453.

Zech, John R, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano and Eric Karl Oermann (2018). 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study'. In: *PLoS medicine* 15.11, e1002683.

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

Berner, Christopher et al. (2019). 'Dota 2 with large scale deep reinforcement learning'. In: *arXiv preprint arXiv:1912.06680*.

Jacobsen, Jörn-Henrik, Jens Behrmann, Richard S. Zemel and Matthias Bethge (2019). 'Excessive Invariance Causes Adversarial Vulnerability'. In: *International Conference on Learning Representations (ICLR)*.

Vinyals, Oriol et al. (2019). 'Grandmaster level in StarCraft II using multi-agent reinforcement learning'. In: *Nature* 575.7782, pp. 350–354.

Brown, Tom B. et al. (2020). 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

(FAIR)†, Meta Fundamental AI Research Diplomacy Team et al. (2022). 'Human-level play in the game of Diplomacy by combining language models with strategic reasoning'. In: *Science* 378.6624, pp. 1067–1074.

Fawzi, Alhussein et al. (2022). 'Discovering faster matrix multiplication algorithms with reinforcement learning'. In: *Nature* 610.7930, pp. 47–53.

Romiti, Sara, Christopher Inskip, Viktoriia Sharmanska and Novi Quadrianto (2022). 'RealPatch: A Statistical Matching Framework for Model Patching with Real Samples'. In: *CoRR* abs/2208.02192.

Sagawa, Shiori et al. (2022). 'Extending the WILDS Benchmark for Unsupervised Adaptation'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=z7p2V6KR0OV.

2 | SOME FAIRLY-INCOMPREHENSIVE BACKGROUND ON SOME FAIRLY-RELATED THINGS

Preamble

In this chapter, I aim to provide background on the topics spanned by the works in this thesis, both individually, and holistically. While I will on occasion point to exemplar methods, generally I 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 paradigms. This is to say, this chapter does not aspire to be a comprehensive survey of DA, DG, AF, 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 ML literature, the field having grown precipitously over the last decade since the onset of the deep-learning revolution. 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 I wish to avoid repetition in this respect.

The main themes of this thesis – as just introduced – are SemiSL and distributional robustness (DR), the latter in the context of AF (Chapters 3 and 4) and DG (Chapter 5), specifically. I 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, I begin with discussion of the classical supervised learning (SL) 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. I 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, or (statistical) shortcuts (I 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 SCL has close ties to DG (Arjovsky et al., 2019), the focus of Chapter 5; in light of this, I afford dedicated discussion of the SCL problem and what conditions are needed to engender that problem.

After introducing, in turn, AF, DA and DA – as different realisations of the DRO problem – term, I attempt to unify their problem setting distribution shift by drawing upon causal principles.

To reiterate, while this thesis does not directly tangle with questions of causal inference, the field of causality (Pearl, 2009) provides, through causal Bayesian networks (CBNs) 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, I conclude this section – as alluded to above – with a rundown of specific learning paradigms – departing from the problem setups they can might be applied to – featured throughout the works in this thesis, namely SemiSL, adversarial learning (AdvL), and invertible neural networks (INNs).

2.1 Some notes on notation

I 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, I denote random variables using upper-case (non-calligraphic) letters and their associated observed/deterministic/realised variables with the corresponding lower-case letters. Following convention, I consistently denote by X and Y the input (covariate) and target (response) variables, respectively; by S some auxiliary variable (or 'concept' as I have earlier called it) on which we want to condition (for evaluation and/or optimisation), such as the domain (in DA/DG) or sensitive attribute (in AF); 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, $X : \Omega \to \mathbb{R}^d$, realisations $x \in \mathcal{X} \subset \mathbb{R}^d$ defined on a subset of the d-dimensional space of real numbers. I 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. I 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, I abuse notation by allowing functions of the form $f : X \to Y$ to accept random and observed variables interchangeably; I 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.

2.2 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* 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 (true) risk minimisation seeks the hypothesis $f^* \in \mathcal{F}$ that is the minimiser of the eponymous *risk*, $\mathcal{R}$, defined according to some statistical distance, or *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 required to identify a sample from the true distribution given a coding scheme optimised for the predictive distribution and takes the form

$$H(Y, f(X)) \triangleq \mathbb{E}_{P(Y)}[\log P(f(X))]. \tag{2.1}$$

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

$$\mathcal{R}(f) \triangleq \mathbb{E}_{(X,Y)\sim P^{tr}(X,Y)}[\mathcal{L}(f(X, Y))]. \tag{2.2}$$

Empirical risk minimisation

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 restricted to 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 the substitution of $P^{tr}(X, Y)$ 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

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

Long-tailedness and importance-weighting

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-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. 2.2, with an importance weight (IW) 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 estimation (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 IW 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. 2.2 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 $\mathbb{R}$), 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 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 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.

Yet, despite its long and storied history, with roots in early statistical modelling, the practical usefulness of importance-weighted ERM in the context of modern deep learning (DL) 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 standard (unaugmented) ERM. Further empirical evidence for this was provided by Sagawa et al. (2019), wherein the importance of combining aggressive regularisation with (a dynamic form of) importance-weighting for strong worst-group generalisation is stressed. Theoretical support for these results was later adduced by Zhai et al. (2022) with proofs showing that equivalence of the implicit biases of these IW-based algorithms and ERM. Summarily, for all its intuitiveness, importance-weighting, 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).

The pitfalls of statistical modelling

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 arise disconcertingly often in real-world – many, moreover, safety-critical – applications. This is to say, the predictions of a statistical model are only trustable 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).

The true causal relationships between independent and dependent variables is in general *unidentifiable* given the training data alone, due to confounding variables; additional information, as provided by interventions, or *environments* (Peters et al., 2016) – a tack popular within the domain generalisation literature, wherein domain can be viewed as a different intervention on the true distribution – is needed to resolve the immanent statistical ambiguity. By 'confounding variable', or *confounder*, I 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 SCL problems addressed in Chapters 3 and 4, we will see statistical learning break down in a similar way yet for essentially the opposite reason. Namely, instead of having latent variable that explains the statistical dependency between X and 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 $X \rightarrow Y$ (or some subset of X) assumed to be the true causal mechanism. I will broach more deeply SCL is and how the mechanisms that give rise to it in §2.4; for now, however, we will proceed to characterising some of the types of distribution shift that statistical learning might contend with.

Short-cuts make for long delays

2.3 A (BRIEF) TAXONOMY OF DISTRIBUTION SHIFTS

In this section, I 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, I draw heavily upon the works of Moreno-Torres et al. (2012) and Castro et al. (2020) in the definitions, noting that the ML literature is not of a single mind regarding the terminology and its semantics. In §2.8 I 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 I leave that aside for now and seek to present them in more general terms.

2.3.1 *Covariate shift*

As the most well-studied of the shifts, 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), I 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 I disregard the importance of distinguishing between the two causal directions; in the context of DA and SemiSL, I 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').

2.3.2 *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 §2.2, 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).

2.3.3 *Concept shift*

To complete the triad of bivariate distribution shifts (I 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 (alignment via reinforcement learning from human feedback (RLHF) being a prime and topical example of such a task (Bai et al., 2022)). 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 without invoking strict assumptions due to the need to decouple the constituent shifts (a problem of identifiability).

2.3.4 *Sampling bias*

Sampling (which I 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 I consider extreme cases of it in which certain demographics, or outcomes for certain demographics, are omitted from the training data, promoting SCs 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 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.

2.4 Shortcut learning

While the notoriety of shortcut learning (SCL) in ML is relatively recent, the phenomenon underpinning it is a fundamental one in statistics, one that may be summed up with the age-old

Simplicity bias

fallacy *cum hoc ergo propter hoc*, or *correlation does not imply causation*. Deep neural networks (DNNs) define highly 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 does not 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 AF are active – more so than ever – subfields of ML contending with different flavours of the problem and have successfully developed mitigation strategies for them.

Corsican cows

The now-canonical example of shortcut-learning in the DG literature – which I 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 such behaviour could induce 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 elucidated by Zech et al. (2018). There are also obvious ethical concerns that arise when the spurious features in question correspond to sensitive attributes such as '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 facets of SCL 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,

Feature suppression and the multi-view hypothesis

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 ascribed to. High-frequency cues, such as colour and texture, are readily modulated by (unstable under) changes in lighting, for instance, making them less reliable 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 ImageNet-trained DNNs.

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 AF 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.

2.5 Algorithmic fairness

Research into algorithmic fairness (AF) 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 consequentially affect human lives– 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.

Notions of fairness

There are many strands of AF following different criteria for what it means for a predictor to be *fair*; in this thesis I 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; I will generally favour the former). As a reminder, I denote group membership using the random variable $S : \Omega \to \mathcal{S}$ – with realisation s – and assume here that said variable and the target variable are both discrete (and in most cases binary); this is the most common

setup considered in the AF 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 (Grari et al., 2021). Though the focus is on group fairness, at the tail-end of the section, I will touch briefly on the premise of individual fairness (IF; Dwork et al., 2012) to take the opportunity to draw a parallel between it and the idea of consistency/smoothness in SemiSL.

Particularly, I will consider two 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 kind partially subsumes the former: by solving its entailed problem, one can subsequently 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 AF is to learn some predictor, $f : \mathcal{X} \to \mathcal{Y}$ that is, according to some definition, non-discriminatory – that does not deprive a given individual of opportunities by virtue of belonging to a particular sensitive group. A frequently-invoked example is that of a financial institution running 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 f, $\mathcal{X}$, 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', I 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).

2.5.1 *Equalised rates*

A long-standing, and vigorously-debated, problem in AF 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 AF 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, I briefly discuss and formulate *demographic parity (DP)*, *equality of opportunity (EqOp)*, and *equalised odds (EqOd)* 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 DG literature (Sagawa et al., 2019), a perspective which is known under the guise of *minimax fairness* in the AF 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 EqOp (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

$$\forall s \in \mathcal{S} : P(\hat{Y} = 1|S = s) \overset{!}{=} \text{constant}, \tag{2.3}$$

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 random 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.

$$P(\hat{Y} = 1|S) \overset{!}{=} P(\hat{Y} = 1). \tag{2.4}$$

Since requiring this condition be satisfied exactly is, generally, over-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:

$$\forall s \in \mathcal{S} : P(\hat{Y} = 1|S = s)P(S = s) \in [1 - \epsilon, 1 + \epsilon]. \tag{2.5}$$

We can measure the 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.

This idea of statistical independence, or invariance, 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:

$$\mathcal{I}(Y; S) \triangleq D_{KL}\Big(P(Y, S) \| P(Y) \otimes P(S)\Big). \tag{2.6}$$

Mutual-information minimisation

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 (FRL), DA, and DG), for example, a common method for imparting

independence, $Z \perp S$ – which is sufficient for $Y \perp S$, given a predictor head, $g : \mathcal{Z} \to \mathcal{Y}$ – between the representations learned by encoder $h : \mathcal{X} \to \mathcal{Z}$ and the sensitive attribute to train h to maximise the conditional entropy $H(\hat{S}|Z)$ of an adversarial predictor, $a : \mathcal{Z} \to \triangle^{|\mathcal{S}|}$ (itself trained via maximum likelihood estimation (MLE)). I shall expatiate on AdvL broadly, and w.r.t. to its use as an infomin engine – as in the forgoing – in §2.10.

Equality of opportunity

Equality of opportunity (EqOp) 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, EqOp can be written, as above, as

$$\forall s \in \mathcal{S} : P(\hat{Y} = 1|Y = 1, S = s) = P(\hat{Y} = 1|Y = 1). \tag{2.7}$$

equalised odds

By making this symmetric w.r.t. Y, such that not only do we demand parity of the true positive rates (TPRs) but also false positive rates (FPRs), we obtain the final of the equalised rate (ER)-based metrics, EqOd (Hardt et al., 2016) – also known as disparate mistreatment:

$$\forall y \in \mathcal{Y}, \forall s \in \mathcal{S} : P(\hat{Y} = 1|Y = y, S = s) = P(\hat{Y} = 1|Y = y). \tag{2.8}$$

Accuracy-fairness trade-off

The inherent trade-off between the utility, as measured by aggregate performance and fairness under conceptions of fairness based on ERs, has long been recognised 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. This accuracy-fairness trade-off can also be said to be a consequence of the fact that the test set itself typically shares the training set's biases; if said data were to be bias-free (a condition which is often unrealistic) then, in theory, no such trade-off would be implied. Nevertheless, given this trade-off does generally apply, 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 (PO) solutions, the discovery of which is the remit of multi-objective optimisation (multi-objective optimisation (MOO); Sawaragi et al., 1985; Deb and Deb, 2013). MOO has appeared both implicitly and explicitly throughout the AF 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 (preference direction) of this solution on the Pareto frontier (PF) controlled by a linear weighting of loss terms optimising for utility and fairness separately.

2.5.2 *Going beyond the accuracy-fairness 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 [the majority] to give to the poor [the minority]' is inapplicable. Healthcare defines a whole host of applications to which this consideration applies, where accurately detecting positive cases amounts to 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 tack 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 (PO) terms, though the same idea has long existed in distributionally-robust optimisation (DRO), and an idea which has been inherited by DG 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), DG (as already noted; Sagawa et al., 2019), or, indeed, AF. 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

$$\inf_{f \in \mathcal{F}} \sup_{P(X,Y) \sim \mathfrak{P}(X,Y)} \mathbb{E}_{(X,Y) \sim P(X,Y)}[\mathcal{L}(f(X), Y)], \tag{2.9}$$

recalling that we use $\mathcal{L} : \mathbb{R}^{\Omega} \times \mathcal{Y}^{\Omega} \to \mathbb{R}$ to denote a loss function of random variables, and overloading f such that its codomain is implied by $\mathcal{L}$, as before in §2.2 (the cross-entropy loss implying a codomain of the standard simplex, for example). 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, f. 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' distribution, else 0. In the context of minimax fairness, $\mathfrak{P}$ would be induced by the sensitive attribute (each element corresponding to the distribution of a sensitive group); the same might naturally be accomplished by conditioning, an alternative I shall pursue in 2.8 after establishing a causal basis.

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 DRO literature this quantity is commonly known as robust accuracy (RobAcc). For convenience and clarity, I first reformulate the standard (non-conditional) accuracy as

$$\text{Acc}(f, \mathcal{D}^{eval}) \triangleq \mathbb{E}_{(x,y) \in \mathcal{D}^{eval}}[\delta_{f(x)y}], \tag{2.10}$$

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 $s \in \mathcal{S}$

$$\text{RobAcc}(f, \mathcal{D}^{eval}) \triangleq \min_{s \in \mathcal{S}} \text{Acc}(f, \mathcal{D}^{eval}_{S=s}), \tag{2.11}$$

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

2.5.3 *Individual fairness*

While notions of group fairness consider fairness at the level of demographic groups, individual fairness (IF) focusses – as expected of the name – on fairness at the individual level, and may be pithily summarised with the apophthegm 'similar individuals should receive similar treatments'. This premise of IF greatly resembles the smoothness assumption from SemiSL that I shall introduce formally in §2.9; glimpsing ahead, however, the smoothness assumption can be seen as a K-Lipschitz constraint on our function class; this is the very manner in which the *Fairness Through Awareness* (FTA) – the most general concept of IF – is couched in Dwork et al. (2012), its name contrasting with FTU. The central question thus implied by FTA then is what constitutes an appropriate measure of similarity – in both the domain and codomain – for the population and task being considered; while being free from the fetters of group fairness and its notional issues, this question is itself far from trivial.

2.5.4 *Bias propagation and systematic censoring*

I conclude this section with a brief discussion of residual unfairness, as termed by Kallus and Zhou (2018), and the problem setting engendering 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.

2.6 Domain adaptation

DA contends 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, $f^\star : \mathcal{X} \to \mathcal{Y}$), 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* ($X \to Y$); practically, there are many cases for which the converse in fact holds true, i.e. the relationship between X and Y is *anticausal* ($Y \to X$). 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 I will revisit the several times more 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.

2.7 Domain generalisation

I will discuss in §2.8 how the underlying problem of DG may be couched in terms of exogenous variables, over which a bilevel optimisation problem is defined, and this section shall be, accordingly, reasonably brief in eschewing redundancy; I here focus on providing a preliminary introduction to the problem, on distinguishing DG from its sister field of DA and in providing intuition for why DG can work despite some impossibilities. Indeed, I say 'sister' for both DA

and DG aim to maximally transfer knowledge between domains while achieving invariance to domain-specific (nuisance) factors. However, the paradigms diverge in the respect that the task entailed by DA is fundamentally, as per its name, one of out-of-distribution (OOD) generalisation – in the strict sense of the transfer in question being to novel domains – rather than adaptation to a closed set of given, and thus known and seen, domains. To endow some formalism, while in UDA 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 $\mathcal{D}_{e^\dagger}$ from a novel domain $e^\dagger \notin \mathcal{E}$. This can be done in the DRO fashion described by Eq. 2.9, with the meta distribution induced by $\mathcal{E}$, as in Eq. 2.12.

The design of domain generalisation

We would, of course, hope our model to generalise (well) to any arbitrary environment (assuming the task remains consistent), yet 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. Justification for DG might then be viewed according to the following perspectives, respectively rooted in DRO and causality. First, given a set of known variances – induced by the domains – one should expect to be able to leverage these variances to learn a predictor able to generalise both within the convex hull (affine combinations of elements of the set) they define as well as to those variances within the vicinity – proximal to – this hull (Krueger et al., 2021). 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. Second, given a set of interventions on the underlying causal graph defined by the set of environments, one would expect to be able to recover – if only partially – 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.

Justifying domain generalisation

2.8 Through the lens of causality

I now introduce a causal formalism of the distribution-shift problem, a formalism which has been frequently exploited in the DG and AF literature as it provides a simple calculus with which to reason about desired (and undesired) variances. It should be again noted that I 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 in any of the constituent works. The background on causality is thus commensurably light and I would refer the reader to Pearl (2009) for full exposition of the topic.

Domain as an exogenous variable

While the term 'domain' typically refers to the observed distributions as a whole in both DA and DG alike (i.e. 'source' vs. '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 arguably more flexible then, consistent with the JCI formulation of 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. I 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 **A** causes (is a parent of) **B**) while the absence of an edge between two nodes **A** and **B** indicates independence between them when conditioned on their parents, i.e. $\mathbf{A}|\text{Pa}(\mathbf{A}) \perp \mathbf{B}|\text{Pa}(\mathbf{B})$, where $\text{Pa}(\cdot)$ denotes the causal parents of its argument node. Formally, a CBN 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 **G** then defines a probability distribution, conditional on its parents, such that the joint distribution of $\mathcal{V}$, $P(V)$, factorises as $P(V) = \prod_{v \in \mathcal{V}} P(v|\text{Pa}(v))$ where we can now define $\text{Pa}(\cdot)$ 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, SCs. To attempt to crystallise this, I illustrate in Fig. 2.1 illustrate minimal CBNs corresponding to different distribution shifts for a (causal) prediction task.

Independent causal mechanisms

This notion of probabilistic independence embodied by CBNs is the very basis for the independent causal mechanisms (ICM) principle (Schölkopf et al., 2021), understanding 'modules' and 'nodes' to be synonymous (for nodes can represent factors of arbitrary abstraction). Namely, the principle posits that independence between factors implies that changing (intervening on) any one mechanism in the system (CBN) leaves all other mechanisms within the same system unchanged

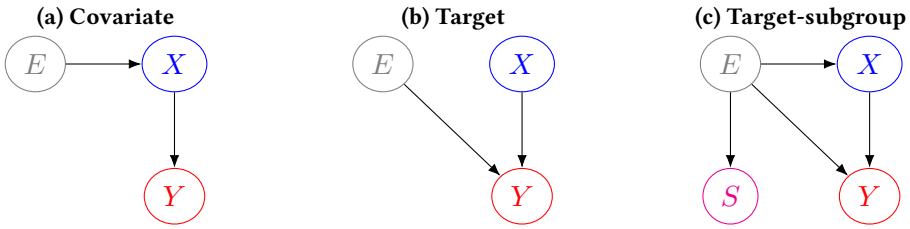


Figure 2.1: CBNs corresponding to different distribution shifts, induced by exogenous variable, E, for a causal prediction ($X \to Y$) task, where X and Y correspond to the covariates and response variable, respectively. (c) introduces an additional, auxiliary label, S, which forms the basis of the problems tackled in Chapters 3 and 4.

(invariant), such that in the above-described covariate shift case the labelling mechanism, $X \to Y$, remains constant, despite influence on the input distribution by the exogenous variable. One may further derive from the ICM principle, by coarsening, the sparse mechanism shift hypothesis, that small distribution shifts tend to manifest in sparse/local way in the causal graph, which is to say, that not all factors should vary simultaneously. By virtue of modularity, a world model based on causally-factorised latents is maximally compressive in general, and by this I mean it provides the simplest explanation to the complex physical world. It is much more efficient, for example, to explain changes in the shape of an object as one moves around it by a change in vantage point, in terms of a single global transformation, rather than by many concerted changes in local structure; by the same token, a change in illumination should not connote a change in colour (the principle of 'colour constancy'), again because the contrary explanation, that multiple objects have undergone concerted changes in their spectral properties (knowing such properties to not be readily mutable) is significantly more complex/prior-defying. The connection between learning causally-faithful models and modern physics thus becomes clear when one thinks as modularity being synonymous with invariances, or *symmetries*, it being little exaggeration to say that modern physics is the study of symmetries and the corresponding conserved quantities, as predicated by Noether's celebrated theorem (Noether, 1918). In light of this, the statistically- and causally-driven approaches are manifestly at odds with one another in the context of SCL – shortcut features provide the simplest, loss-minimising solution based on the training distribution, $P^{tr}(X, Y)$, however they do not provide the simplest solution in the causal sense, as evidenced by the lack of generalisability. Models that are faithful to the underlying causal structure of the problem are inherently more robust to distributional changes, more protean as the learned modules can be reconfigured readily to cater to the problem at hand, and locally updatable, such that new information can be incorporated without degrading modules adapted for orthogonal tasks – this is particularly pertinent in given recent elucidations of the distorting effects of fine-tuning on distributional-robustness (Andreassen et al., 2021; Kumar et al., 2022).

Reformulating the DRO objective

I presented in §2.5.2, within the context of minmax fairness, a formulation of the bilevel DRO objective mathematises the desire to minimise the worst risk, where 'worst' is computed over a collection of sub-distributions, corresponding in said case of minimax fairness to different sensitive attributes. It is common to couch DG in the same terms (Arjovsky et al., 2019; Sagawa

et al., 2019; Krueger et al., 2021); with some restatement of that objective, using the new calculus, one can define the optimal predictor over the (empirical) environment distribution $\mathcal{E}$ from 2.7 as

$$f^*_{\text{robust}} = \inf_{f \in \mathcal{F}} \sup_{e \in \mathcal{E}} \mathbb{E}_{P^{tr}(X,Y|E=e)}[\mathcal{L}(f(X), Y)]. \tag{2.12}$$

Given that domains/environments can be modelled as deriving from different interventions of the causal factorisation of $P(X, Y)$, it follows that, for Eq. 2.12 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. To recall, though generalisation to arbitrary perturbations may be provably hard (or impossible; 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 $\mathcal{E}$ and to an extrapolated region outside of it (Krueger et al., 2021). Indeed, one expects 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 DG (Arjovsky et al., 2019; Ahuja et al., 2020; Creager et al., 2021) literature (as discussed in §2.7), with the caveat that a degree of inductive bias or additional information is necessary to provably identify the correct causal relations (Lin et al., 2022).

2.9 Semi-supervised learning

Given the titular reference to SemiSL [1], it is only appropriate that a part of this background section be devoted to the topic. However, I 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.

I 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-DL era (many of which are still fundamentally applicable today, however), a book I will reference extensively throughout this brief overview of the topic. For a comprehensive and current survey of SemiSL methods in the current DL era, on the other hand, I 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 §2.2, which we can use for standard

1 Since self-supervised learning is also referenced 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.

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 UL 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 SL with additional information (Chapelle et al., 2009).

Closely related to SemiSL is the idea of *Transductive Learning* (TL; 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 DA and DG, 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, I 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.

Transduction

2.9.1 *Justifying semi-supervised learning*

While SemiSL is a tantalising prospect whenever one has a large corpus of unlabelled data coupled with a 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. Indeed, theory prescribes that certain assumptions about the data-generating process be met in order for the learning paradigm to bear fruit (in the 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. I 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), I 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 ('perspective' being perhaps the more apposite term) engendering 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 the Transductive Support Vector Machine (TSVM; Joachims et al., 1999), which shares the inductive Support Vector Machine's (SVM's) aim of maximising the margin between the decision boundary and nearest data-points (the *support vectors*) – yielding the *maximum margin hyperplane* – yet considers both the (labelled) training and (unlabelled) test data during fitting.

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

$$|d_{\mathcal{Y}}(f(x), f(x')) - d_{\mathcal{X}}(x, x')| \leq \epsilon, \tag{2.13}$$

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, hearkening back to the earlier discussion on IF) 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 5 wherein 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.

2.9.2 *Causal connections: when should(n't) semi-supervised learning 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 in §2.8, 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.

2.10 Adversarial learning

AdvL is a general paradigm characterising – in game-theoretical parlance – non-coöperative (competitive) systems, or 'games' in which two or more 'players' compete over shared, or at least 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 I 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).

Minimax

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

Bilevel optimisation

utility, as measured by the player-specific function $\pi : \Sigma \to \mathbb{R}^+$, 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)

Nash equilibria

$$\forall \sigma_{min} \in \Sigma_{min} : \pi_{min}(\sigma^*_{min}, \sigma^*_{max}) \geq \pi_{min}(\sigma_{min}, \sigma^*_{max}), \tag{2.14}$$

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 PO 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 presented 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 DNNs 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.

Notable concrete applications of AdvL include artificial curiosity (Schmidhuber, 1992), Generative Adversarial Networks (generative adversarial network (GAN); Goodfellow et al., 2014), self-play (Silver et al., 2018), adversarial robustness (Szegedy et al., 2013), and, most germanely to this thesis, domain-invariant learning (Ganin et al., 2016; Zhao et al., 2019) and FRL (Edwards and Storkey, 2015; Madras et al., 2018). In the latter applications, AdvL 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 AF, respectively. For this, both players take the form of a neural network 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 $a_{min} : \mathcal{X} \times \Theta_{min} \to \mathcal{Z}$ and $a_{max} : \mathcal{Z} \times \Theta_{max} \to \mathbb{R}$. I note, incidentally, that due to the continuity and non-convexity of the players in this case, 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, respective to the learner and its adversary:

Adversarial infomin

- **learner**: create a censored version (representation) of the input, z, that maximally minimises the amount of information about s that can be extracted.
- **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 of the input, z, given by the learner.

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

$$\min_{\theta_{min}\in\Theta_{min}} \max_{\theta_{max}\in\Theta_{max}} - \mathbb{E}_{(x,s)\sim\mathcal{D}}[\ell(a_{max}(a_{min}(x,\theta),\psi),s)], \tag{2.15}$$

where I 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; 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 AF 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 by Eq. 2.15 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 PF 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.

I have already alluded to some of the challenges entailed in adversarial infomin approaches; here, I summarise the two most salient ones. First, the strength of information-minimisation

is proportional to the strength adversary used to drive it, and, moreover, 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). Second, adversarial setups are generally susceptible to cycles (oscillations) 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 recognise the pitfall of cyclic dynamics and propose a middle ground of training against a stochastic ensemble of adversaries.

Approximation errors

Cyclic dynamics

2.11 Invertible neural networks

Chapter 3 of this thesis explores the application of INNs to FRL and so will afford some brief discussion to their basic workings here. An invertible neural network (Kobyzev et al., 2020), as the name would so suggest, refers to neural networks 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 if 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.

Bijectivity

INNs are foremost used for their density-estimation – and by complement, generative modelling – capabilities due to their hallmark diffeomorphic property that makes 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 evidence lower bound (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 (VAEs; 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,

Normalising flows

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 variational auto-encoders (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 4, normalising flows (NFs) have been applied applied to AF 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 of INNs

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. bidirectional K-Lipschitzness), the optimisation of INNs can be prone

to instabilities that can render them *numerically* non-invertible, despite their design, and thereby invalidate computations made according to Eq. 2.11 (Behrmann et al., 2021).

BIBLIOGRAPHY

Noether, E (1918). 'Invariante Variationsprobleme'. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1918, pp. 235–257.

Brown, George W (1951). 'Iterative solution of games by fictitious play'. In: *Act. Anal. Prod Allocation* 13.1, p. 374.

Scudder, Henry (1965). 'Probability of error of some adaptive pattern-recognition machines'. In: *IEEE Transactions on Information Theory* 11.3, pp. 363–371.

Fralick, S (1967). 'Learning to recognize patterns without a teacher'. In: *IEEE Transactions on Information Theory* 13.1, pp. 57–64.

Sawaragi, Yoshikazu, Hirotaka Nakayama and Tetsuzo Tanino (1985). *Theory of multiobjective optimization*. Elsevier.

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

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

Schmidhuber, Jürgen (1992). 'Learning factorial codes by predictability minimization'. In: *Neural computation* 4.6, pp. 863–879.

Pearl, Judea (1995). 'From Bayesian networks to causal networks'. In: *Mathematical models for handling partial knowledge in artificial intelligence*, pp. 157–182.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu (1996). 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: *Knowledge Discovery and Data Mining*.

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

Joachims, Thorsten et al. (1999). 'Transductive inference for text classification using support vector machines'. In: *Icml*. Vol. 99, pp. 200–209.

Kaplow, Louis and Steven Shavell (1999). 'The conflict between notions of fairness and the Pareto principle'. In: *American Law and Economics Review* 1.1, pp. 63–77.

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

Szummer, Martin and Tommi Jaakkola (2001). 'Partially labeled classification with Markov random walks'. In: *Advances in neural information processing systems* 14.

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

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

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

Chapelle, Olivier, Bernhard Scholkopf and Alexander Zien (2009). 'Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]'. In: *IEEE Transactions on Neural Networks* 20.3, pp. 542–542.

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

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

Henrich, Joseph, Steven J Heine and Ara Norenzayan (2010). 'The weirdest people in the world?' In: *Behavioral and brain sciences* 33.2-3, pp. 61–83.

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

Moreno-Torres, Jose G, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla and Francisco Herrera (2012). 'A unifying view on dataset shift in classification'. In: *Pattern recognition* 45.1, pp. 521–530.

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

Bair, Eric (2013). 'Semi-supervised clustering methods'. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.5, pp. 349–361.

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

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

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

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

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

Kingma, Diederik P and Max Welling (2014). 'Auto-Encoding Variational Bayes'. In: *stat* 1050, p. 1.

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

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

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

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

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

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

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

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

Fefferman, Charles, Sanjoy Mitter and Hariharan Narayanan (2016). 'Testing the manifold hypothesis'. In: *Journal of the American Mathematical Society* 29.4, pp. 983–1049.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky (2016). 'Domain-adversarial training of neural networks'. In: *The journal of machine learning research* 17.1, pp. 2096–2030.

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

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2016). 'Inherent trade-offs in the fair determination of risk scores'. In: *arXiv preprint arXiv:1609.05807*.

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

Silver, David et al. (2017). 'Mastering the game of go without human knowledge'. In: *nature* 550.7676, pp. 354–359.

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

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

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

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

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

Dastin, Jeffrey (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women'. In: *Ethics of data and analytics*. Auerbach Publications, pp. 296–299.

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

Kallus, Nathan and Angela Zhou (2018). 'Residual Unfairness in Fair Machine Learning from Prejudiced Data'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 2444–2453.

Liu, Lydia T, Sarah Dean, Esther Rolf, Max Simchowitz and Moritz Hardt (2018). 'Delayed impact of fair machine learning'. In: *International Conference on Machine Learning*. PMLR, pp. 3150–3158.

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

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

Silver, David et al. (2018). 'A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play'. In: *Science* 362.6419, pp. 1140–1144.

Unterthiner, Thomas, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer and Sepp Hochreiter (2018). 'Coulomb GANs: Provably Optimal Nash Equilibria via Potential Fields'. In: *International Conference on Learning Representations*.

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

Zech, John R, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano and Eric Karl Oermann (2018). 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study'. In: *PLoS medicine* 15.11, e1002683.

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

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

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

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

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

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

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

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

Vinyals, Oriol et al. (2019). 'Grandmaster level in StarCraft II using multi-agent reinforcement learning'. In: *Nature* 575.7782, pp. 350–354.

Wang, Tianlu, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang and Vicente Ordonez (2019). '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*, pp. 5310–5319.

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

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

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

Castro, Daniel C, Ian Walker and Ben Glocker (2020). 'Causality matters in medical imaging'. In: *Nature Communications* 11.1, pp. 1–10.

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

Gulrajani, Ishaan and David Lopez-Paz (2020). 'In Search of Lost Domain Generalization'. In: *International Conference on Learning Representations*.

Kobyzev, Ivan, Simon JD Prince and Marcus A Brubaker (2020). 'Normalizing flows: An introduction and review of current methods'. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11, pp. 3964–3979.

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

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

Mooij, Joris M, Sara Magliacane and Tom Claassen (2020). 'Joint Causal Inference from Multiple Contexts'. In: *Journal of Machine Learning Research* 21.99, pp. 1–108.

Navon, Aviv, Aviv Shamsian, Ethan Fetaya and Gal Chechik (2020). 'Learning the Pareto Front with Hypernetworks'. In: *International Conference on Learning Representations*.

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

Andreassen, Anders, Yasaman Bahri, Behnam Neyshabur and Rebecca Roelofs (2021). 'The evolution of out-of-distribution robustness throughout fine-tuning'. In: *arXiv preprint arXiv:2106.15831*.

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

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

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

Grari, Vincent, Sylvain Lamprier and Marcin Detyniecki (2021). 'Fairness-aware neural Rényi minimization for continuous features'. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 2262–2268.

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

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

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

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal and Yoshua Bengio (2021). 'Toward causal representation learning'. In: *Proceedings of the IEEE* 109.5, pp. 612–634.

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

Wang, Ke Alexander, Niladri Shekhar Chatterji, Saminul Haque and Tatsunori Hashimoto (2021). 'Is Importance Weighting Incompatible with Interpolating Classifiers?' In: *International Conference on Learning Representations*.

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

Bai, Yuntao et al. (2022). 'Training a helpful and harmless assistant with reinforcement learning from human feedback'. In: *arXiv preprint arXiv:2204.05862*.

Cerrato, Mattia, Marius Köppel, Alexander Segner and Stefan Kramer (2022). 'Fair Group-Shared Representations with Normalizing Flows'. In: *arXiv e-prints*, arXiv–2201.

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

Kumar, Ananya, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma and Percy Liang (2022). 'Fine-Tuning Distorts Pretrained Features and Underperforms Out-of-Distribution'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=UYneFzXSJWh.

Lin, Yong, Shengyu Zhu and Peng Cui (2022). 'ZIN: When and How to Learn Invariance by Environment Inference?' In: *arXiv preprint arXiv:2203.05818*.

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

Zhai, Runtian, Chen Dan, Zico Kolter and Pradeep Ravikumar (2022). 'Understanding Why Generalized Reweighting Does Not Improve Over ERM'. In: *arXiv e-prints*, arXiv–2201.

Part II

MIDDLE

This part comprises three papers, two peer-reviewed and published at eminent *ML* conferences, one forever a work-in-progress (but no less a paper for it); the former have been adapted for this thesis textually, with the view to improve clarity, technical correctness, and consistency (both internally and between works) but the adaptations are limited in scope and the preponderance of the text remains as-published.

3 NULL-SAMPLING FOR INTERPRETABLE AND FAIR REPRESENTATIONS

Authors:
Thomas Kehrenberg[1], Myles Bartlett[1], Oliver Thomas[1] & Novi Quadrianto[1,2,3]
Affiliations:
[1]Predictive Analytics Lab (PAL), University of Sussex, Brighton, UK
[2]BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning
[3]Monash University, Indonesia
Conference: *European Conference on Computer Vision* (ECCV), 2020
DOI: 10.1007/978-3-030-58574-7_34
Note: The appendices have been included under §3.6.

Abstract

We propose to learn invariant representations, in the data domain, to achieve interpretability in algorithmic fairness. Invariance implies a selectivity for high level, relevant correlations w.r.t. class label annotations, and a robustness to irrelevant correlations with protected characteristics such as race or gender. We introduce a non-trivial setup in which the training set exhibits a strong bias such that class label annotations are irrelevant and spurious correlations cannot be distinguished. To address this problem, we introduce an adversarially-trained model with a *null-sampling* procedure to produce invariant representations in the data domain. To enable disentanglement, a partially-labelled *representative* set is used. By placing the representations into the data domain, the changes made by the model are easily examinable by human auditors. We demonstrate the effectiveness of our method on both image and tabular datasets: Coloured MNIST, CelebA, and the Adult dataset.

3.1 Introduction

Without due consideration for the data-collection process, machine learning algorithms can exacerbate biases, or even introduce new ones if proper control is not exerted over their learning (Holstein et al., 2019). While most of these issues can be solved by controlling and curating data collection in a fairness-conscious fashion, doing so is not always an option, such as when working with historical data. Efforts to address this problem algorithmically have been centred on developing statistical definitions of fairness and learning models that satisfy these definitions. One popular definition of fairness used to guide the training of fair classifiers, for example, is demographic parity (DP), stating that positive outcome rates should be equalised (or *invariant*) across protected groups.

In the typical fair-classification setup, we have an input $x \in \mathcal{X}$, a sensitive attribute $s \in \mathcal{S}$, that represents some inadmissible (for prediction) information like gender and a class label $y \in \mathcal{Y}$ which is the prediction target. The idea of fair *representation* learning (Zemel et al., 2013; Edwards and Storkey, 2016; Madras et al., 2018) is then to transform the input x to a representation $z \in \mathcal{Z}_{\neg s}$ which is invariant to s, so that in training a downstream predictor on that representation one cannot introduce a forbidden dependence on s. A good fair representation is one that preserves most of the information from x while satisfying the aforementioned constraints.

As unlabelled data is much more freely available than labelled data, it is of interest to learn the representation in an unsupervised manner, for allowing us to draw on a much more diverse pool of data to learn from. While annotations for y are often hard to come by (and often noisy; see Kehrenberg et al., 2020), annotations for the sensitive attribute s are usually less so, as s can often be obtained from demographic information provided by census data. We thus consider the setting where the representation is learned from data that is only labelled with s and not y. This is in contrast to most other representation learning methods. We call the set used to learn the representation the *representative* set, because its distribution is meant to match the distribution of the deployment setting (and is thus representative thereof).

Once we have learnt the mapping from x to z, we can transform the *training* set which, in contrast to the representative set, has the y labels (and s labels). In order to make our method more widely applicable, we consider an *aggravated fairness problem* in which the training set contains a strong spurious correlation (SC) between s and y, making it impossible – without some overriding inductive bias – to learn from it a representation invariant to s but variant to y, variance to y being important as this is the variable we care about predicting accurately. The training set thus does *not* match the deployment setting, thereby rendering the representative set essential for learning the right invariance. Throughout the remainder of the paper, we will use the terms *spurious* and *sensitive* interchangeably, depending on the context, to refer to an attribute of the data we seek invariance to. We can draw a connection between learning in the presence of SCs and what Kallus and Zhou (2018) call *residual unfairness*. Consider the Stop, Question and Frisk (SQF) dataset for example: the data was collected in New York City, but the demographics of the recorded cases do not represent the true demographics of NYC well. The demographic attributes of the recorded individuals might correlate so strongly with the prediction target that the two are nigh-indistinguishable from a statistical perspective. This is the scenario that we are investigating: s and y are so closely correlated in the labelled dataset that they cannot be distinguished, but the learning of s is favoured due to its lower complexity. The deployment setting (i.e. the test set) does not possess this strong correlation and thus a naïve approach will lead to very unfair predictions. In this case, a disentangled representation is insufficient; the representation needs to be explicitly invariant solely with respect to s. In our approach, we make use of the (partially labelled) representative set to learn this invariant representation.

Despite there being a substantial body of existing literature devoted to the problems of fair-representation learning (FRL), exactly how the invariance in question is achieved is often overlooked. When critical decisions, such as who should receive bail or be released from jail, are being deferred to an automated decision making system, it is critical that people be able to trust the logic of the model underlying it, whether it be via semantic or visual explanations. We build on the work of Quadrianto et al. (2019) and learn a decomposition ($f^{-1} : (\mathcal{Z}_s \times \mathcal{Z}_{\neg s}) \to \mathcal{X}$)

of the *data domain* (X) into independent subspaces *invariant* to s ($\mathcal{Z}_{\neg s}$) and *variant* to s ($\mathcal{Z}_s$), which lends an interpretability that is absent from most representation-learning methods. While model interpretability has no strict definition (Zhang and Zhu, 2018), we follow the intuition of Adel et al. (2018) – *a simple relationship to something we can understand*, a definition which representations in the data domain inherently fulfil.

Whether as a result of the aforementioned sampling bias or simply because the features necessarily co-occur, it is not rare for features to correlate with one another in real-world datasets. Lipstick and gender for example, are two attributes that we expect to be highly correlated and to enforce invariance to gender can implicitly enforce invariance to make-up. This is arguably the desired behaviour. However, unforeseen biases in the data may engender cases which are less justifiable. By baking interpretability into our model (by having representations in the data domain), though we still have no better control over what is learned, we might at least diagnose such pathologies.

To render our representations interpretable, we rely on a simple transformation we call *null-sampling* for mapping invariant representations into the data domain. Previous approaches to FRL (Edwards and Storkey, 2016; Louizos et al., 2016; Beutel et al., 2017; Madras et al., 2018, inter alia) rely upon training AEs to jointly minimise the reconstruction (maximising MI w.r.t. X) and invariance (minimising MI w.r.t. S) losses. We discuss first how this can be done with such a model that we refer to as conditional VAE (cVAE), before arguing that the bijectivity of INN (Dinh et al., 2014) makes them better suited to this task. We refer to the variant of our method based on these as conditional flow (cFlow). INNs have several properties that make them appealing for unsupervised representation learning. The focus of our approach is on learning invariant representations that preserve the non-sensitive information maximally, with knowledge of only s – and none of the target y – while at the same time having the ability to easily probe what has been learnt. Our contributions are thus two-fold:

1. We propose a simple approach to generating representations that are invariant to a feature s, while having the benefit of interpretability that comes with being in the data domain. We call our method **NIFR** (**N**ull-sampling for **I**nterpretable and **F**air **R**epresentations).

2. We explore a setting where the labelled training set suffers from varying levels of sampling bias, demonstrating an approach based on transferring information from a more diverse representative set, with guarantees of the non-spurious (semantic) information being preserved.

3.2 Background

3.2.1 *Learning fair representations*

Given a sensitive attribute s (for example, gender or race) and inputs x, a fair representation z of x is then one for which $z \perp s$ holds, while ideally also being predictive of the class label y. Zemel et al. (2013) was the first to propose the learning of fair representations which allow for transfer to new classification tasks. More recent methods are often based on variational auto-encoders (VAEs) (Kingma and Welling, 2014; Edwards and Storkey, 2016; Louizos et al., 2016;

Beutel et al., 2017). The achieved fairness of the representation can be measured with various fairness metrics. These measure, however, usually how fair the predictions of a classifier are and not how fair a representation is.

The appropriate measure of fairness for a given task is domain-specific (Liu et al., 2019) and there is often not a universally accepted measure. However, DP is the most widely used (Edwards and Storkey, 2016; Louizos et al., 2016; Beutel et al., 2017). DP demands $\hat{y} \perp s$ where $\hat{y}$ refers to the predictions of the classifier. In the context of fair representations, we measure the Demographic Parity of a downstream classifier, $f(\cdot)$, which is trained on the representation z, i.e. $\Gamma : \mathcal{Z} \to \mathcal{Y}$.

A core principle of all fairness methods is the *accuracy-fairness trade-off*. As previously stated, the fair representation should be invariant to s ($\to$ fairness) but still be predictive of y ($\to$ accuracy). These desiderata cannot, in general, be simultaneously satisfied if s and y are correlated.

The majority of existing methods for FRL also make use of y labels during training, in order to ensure that z remains predictive of y. This aspect can, in theory, be removed from the methods, but then there is no guarantee that information about y is preserved (Louizos et al., 2016).

3.2.2 *Learning fair, transferable representations*

In addition to producing fair representations, Madras et al. (2018) want to ensure the representations are transferable. Here, an adversary is used to remove sensitive information from a representation z. Auxiliary prediction and reconstruction networks, to predict class label y and reconstruct the input x respectively, are trained on top of z, with s being ancillary input to the reconstruction.

Also related is Creager et al. (2019) who employ a FactorVAE (Kim and Mnih, 2018) regularised for fairness. The idea is to learn a representation that is both disentangled and invariant to multiple sensitive attributes. This factorisation makes the latent space easily manipulable such that the different subspaces can be freely removed and composed at test time. Zeroing out the dimensions or replacing them with independent noise imparts invariance to the corresponding sensitive attribute. This method closely resembles ours when we use an invertible encoder. However, the emphasis of our approach is on interpretability, information-preservation, and coping with sampling bias - especially extreme cases where $|\mathcal{S}^{tr} \times \mathcal{Y}^{tr}| < |\mathcal{S}^{te} \times \mathcal{Y}^{te}|$.

Attempts were made by Quadrianto et al. (2019) prior to this work to learn fair representations in the data domain in order to make it interpretable and transferable. In their work, the input is assumed to be additively decomposable in the feature space into a *fair* and *unfair* component, which together can be used by the decoder to recover the original input. This allows us to examine representations in a human-interpretable space and confirm that the model is not learning a relationship reliant on a sensitive attribute. Though a first step in this direction, we believe such a linear decomposition is not sufficiently expressive to fully capture the relationship between the sensitive and non-sensitive attributes. Our approach allows for the modelling of more complex relationships.

3.2.3 *Learning in the presence of spurious correlations*

Strong spurious correlations make the task of learning a robust classifier challenging: the classifier may learn to exploit correlations unrelated to the true causal relationship between the features and label, and thereby fail to generalise to novel settings. This problem was recently tackled by Kim et al. (2019) who apply a penalty based on the MI between the feature embedding and the spurious variable. While the method is effective under mild biasing, we show experimentally that it is not robust to the range of settings we consider.

Jacobsen et al. (2019) explore the vulnerability of traditional neural networks to spurious variables – e.g. textures, in the case of ImageNet (Geirhos et al., 2019) – and propose an INN-based solution akin to ours. The INN's encoding is split such that one partition, z_b is encouraged to be predictive of the spurious variable while the other serves as the logits for classification of the semantic label. Information related to the nuisance variable is "pulled out" of the logits as a result of maximising $\log p(s|z_n)$. This specific approach, however, is incompatible with the settings we consider, due to its requirement that both s and y be available at training time.

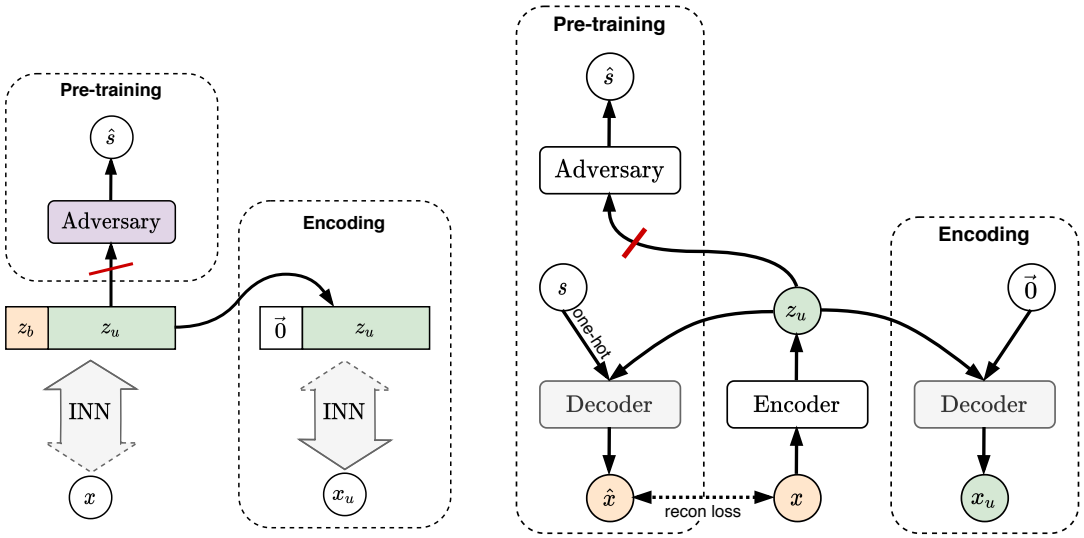
Taking a causal perspective, Arjovsky et al. (2019) propose a variant of ERM they call invariant risk minimisation (IRM). The goal of IRM is to train a predictor that generalises across a large set of unseen environments; because variables with spurious correlations correlations do not represent a stable causal mechanism, the predictor learns to be invariant to them. IRM assumes that the training data is not *i.i.d.* but is partitioned into distinct environments, $e \in E$. The optimal predictor is then defined as the minimiser of the sum of the empirical risk R_e over this set. In contrast, we assume possession of only a single source of *labelled*, albeit spuriously-correlated, data, but that we have a second source of data that is free of spurious correlations, with the benefit being that it need only be labelled w.r.t. s.

3.3 INTERPRETABLE INVARIANCES BY NULL-SAMPLING

3.3.1 *Problem Statement*

We assume we are given inputs $x \in \mathcal{X}$ and corresponding labels $y \in \mathcal{Y}$. Furthermore, there is some spurious variable $s \in \mathcal{S}$ associated with each input x which we do *not* want to predict. Let X, S and Y be the random variables that take on the observed values x, s and y, respectively. The fact that both Y and S are predictive of X implies that $\mathcal{I}(X;Y), \mathcal{I}(X;S) > 0$, where $\mathcal{I}(\cdot;\cdot)$ denotes MI between two random variables. Note, however, that the conditional entropy is non-zero: $H(S|X) > 0$, i.e. S is *not* completely determined by X.

The difficulty of this setup resides in the fact there is a close correspondence between S and Y in the training set such that for a classifier trained via maximum-likelihood estimation, the mappings $\mathcal{X} \to \mathcal{S}$ and $\mathcal{X} \to \mathcal{Y}$ are functionally equivalent, which implies, through transitivity, that $\mathcal{X} \to \mathcal{S} \to \mathcal{Y}$ also is; in many cases, such as those we consider in this paper, the first part of the chain, $\mathcal{X} \to \mathcal{Y}$, is substantially easier to learn than the direct, and, importantly, *causal*, path. This is problematic when we assume that the same correlation does *not* hold in the test set,



(a) cFlow model. **(b)** cVAE model.

Figure 3.1: Training procedure for our models. x: input, s: sensitive attribute, z_u: de-biased representation, x_u: de-biased version of the input in the data domain. The red bar indicates a gradient reversal layer, and $\vec{0}$: the null-sampling operation.

meaning the model cannot rely on shortcuts provided by S if it is to generalise from the training set.

We call this scenario where we only have access to the labels of a biasedly-sampled subpopulation an *aggravated fairness problem*; scenarios of this nature are not uncommon in the real-world. For instance, in long-feedback systems such as mortgage-approval where the demographics of the subpopulation with observed outcomes is *not* representative of the subpopulation on which the model has been deployed. In this case, s has the potential to act as a false (or *spurious*) indicator of the class label and training a model with such a dataset would limit generalisability. Let (X^{tr}, S^{tr}, Y^{tr}) then be the random variables sampled from the training set, and (X^{te}, S^{te}, Y^{te}) likewise be the random variables sampled from the test set. The training and test sets thus induce the following inequality for their MI: $\mathcal{I}(S^{tr}; Y^{tr}) \gg \mathcal{I}(S^{te}; Y^{te}) \approx 0$.

Our goal is to learn a representation Z_u (with realisations z_u), that is independent of S and transferable between downstream tasks. Complementary to z_u, we refer to some abstract component of the model that absorbs the unwanted information related to S as $\mathcal{B}$, the realisation of which we define w.r.t. each of the two models to be described. The requirement for Z_u can be expressed in terms of MI as

$$\mathcal{I}(Z_u; S) \neq 0\,. \tag{3.1}$$

However, for the representation to be useful, we need to capture as much semantically-relevant information from the data as possible. Incorporating this requirement naturally gives rise to the following objective function

$$\min_{\theta} \mathbb{E}_{(X,S)\sim P^{tr}_{(X,S)}}[\lambda \mathcal{I}(f_\theta(X); S) - \log p_\theta(X)], \tag{3.2}$$

where θ refers to the trainable parameters of our model, f_θ, and $p_\theta(\cdot)$ is the likelihood it assigns to the training data, and P^{tr}_{XS} denotes the joint distribution over X^{tr} and S^{tr}. Note that we have slightly abused notation here in allowing f (a Borel Measurable function) to accept random variables X and thereby output random variables, Z_u; the mapping $f(X)$ should be understood to mean $f \circ X(\omega)$ for some event $\omega \in \Omega$, on which basis $f(x)$ can be reinterpreted as $f(X = x)$. In practice, we optimise this loss in an adversarial fashion by playing a minimax game, in which our encoder acts as the generative component from a GAN (Goodfellow et al., 2014) perspective. The adversary is an auxiliary classifier $g :\to \triangle^{|\mathcal{S}|}$ trained to predict the spurious variable s from z_u, with $\triangle^{|\mathcal{S}|}$ being the probability simplex over $\mathcal{S}$. We denote the trainable parameters of the adversary as ϕ; for the parameters of the encoder we use θ, as before. The theoretical objective from Eq. 3.2 can then crystallised as

$$\min_{\theta\in\Theta}\max_{\phi\in\Phi}\mathbb{E}_{(x,s)\sim P^{tr}_{(x,s)}}[\log p_\theta(x) - \lambda H(g_\phi(f_\theta(x)), e_s)], \tag{3.3}$$

where we have substituted $P^{tr}_{(X,S)}$ with the empirical training distribution $P^{tr}_{(x,s)}$, and $H(\cdot,\cdot)$ denotes the cross-entropy between the predicted probabilities and the degenerate target distribution given by the one-hot-encoded labels, $e_s \in \{0,1\}^{|\mathcal{S}|}$. In practice, this adversarial term is realised using a Gradient Reversal Layer (GRL; Ganin et al., 2016) between z_u and g, as is common for adversarial approaches for domain adaptation and fair-representation learning (Edwards and Storkey, 2016).

3.3.2 *The Disentanglement Dilemma*

The objective in equation 3.3 balances the two desiderata: predicting y and being invariant to s. However, in the training set, y and s are so strongly correlated that removing information about s implies removing information about y, causing existing methods to fail under this setting. In order to even define a well-posed learning objective, we require another source of information that allows us to disentangle s and y. For this, we assume the existence of another set of samples that follow a similar distribution to the test set, but while the sensitive attribute is available, the class labels are not. In reality, this is not an unreasonable assumption, as, while properly annotated data is scarce, unlabelled data can be obtained in abundance (with demographic information from census data, electoral rolls, etc.). Indeed, treating the data as unlabelled only w.r.t. y, with the s labels intact, is not without precedence in the fairness literature (Creager et al., 2019; Wick et al., 2019, inter alia). We are restricted only in the sense that the spurious correlations we want to sever are indicated in the features. We call this the *representative set*, with random variables X^{rep} and S^{rep} and satisfying the condition that $\mathcal{I}(S^{rep}; Y^{rep}) \approx 0$ (or rather, it would if the class labels Y^{rep} were available).

We now summarise the training procedure; an outline of the invertible network model (cFlow) can be seen in Fig. 3.1a. First, the encoder network f is trained on (X^{rep}, S^{rep}), during the first phase. The trained network is then used to encode the training set, taking in input x and producing the representation, z_u, decorrelated from the spurious variable. The encoded dataset can then be used to train any off-the-shelf classifier safely, with information about the spurious variable having been absorbed by some auxiliary component $\mathcal{B}$. In the case of the conditional

VAE (cVAE) model, $\mathcal{B}$ takes the form of the decoder subnetwork, which reconstructs the data conditional on a one-hot encoding of s, while for the invertible network $\mathcal{B}$ is realised as a partition of the feature map z (such that $z \triangleq [z_u, z_b]$, where $[\cdot]$ denotes concatenation), given the bijective constraint. Thus, the classifier cannot take the shortcut of learning s and instead must learn how to predict y directly. Obtaining the s-invariant representations, x_u, in the data domain is simply a matter of replacing the $\mathcal{B}$ component of the decoder's input for the cVAE, and z_b for cFlow, with a zero vector of equivalent size. We refer to this procedure used to generate x_u as *null-sampling* (here, with respect to z_b).

Null-sampling resembles the *annihilation* operation described in Xiao et al. (2018), however we note that the two serve very different roles. Whereas the annihilation operation serves as a regulariser to prevent trivial solutions (similar to Jaiswal et al., 2018), null-sampling is used to generate the invariant representations post-training.

3.3.3 *Conditional Decoding*

We first describe a VAE-based model similar to that proposed in Madras et al. (2018), before highlighting some of its shortcomings that motivate the choice of an invertible representation learner.

The model takes the form of a class conditional β-VAE (Higgins et al., 2017), in which the decoder is conditioned on the spurious attribute. We use $\theta_{enc}, \theta_{dec} \in \theta$ to denote the parameters of the encoder and decoder sub-networks, respectively. Concretely, the encoder component performs the mapping $x \to z_u$, while $\mathcal{B}$ is instantiated as the decoder, $\mathcal{B} \triangleq p_{\theta_{dec}}(x|z_u, s)$, which takes in a concatenation of the learned non-spurious latent vector z_u and a one-hot encoding of the spurious label s to produce a reconstruction of the input $\hat{x}$. Conditioning on a one-hot encoding of s, rather than a single value, as done in Madras et al. (2018), is the key to visualising invariant representations in the data domain. If $\mathcal{I}(z_u; s)$ is properly minimised, the decoder can only derive its information about s from the label, thereby freeing up z_u from encoding the unwanted information while still allowing for reconstruction of the input. Thus, by feeding a zero-vector, e_s, to the decoder we achieve $\hat{x} \perp s$. The full learning objective for the cVAE is given as

$$\begin{aligned}\mathcal{L}_{\text{cVAE}} =& \mathbb{E}_{q_{\theta_{enc}}(z_u|x)}[\log p_{\theta_{dec}}(x|z_u, e_s) - \log p_{\theta_{dec}}(s|z_u)] \\ & - \beta D_{\text{KL}}(q_{\theta_{enc}}(z_u|x) \| p(z_u)),\end{aligned} \tag{3.4}$$

where β is a hyperparameter that determines the trade-off between reconstruction accuracy and independence constraints, and $p(z_u)$ is the prior imposed on the variational posterior. For all our experiments, $p(z_u)$ is the standard isotropic Gaussian prior; Fig. 3.1b summarises the procedure diagrammatically.

While we show this setup can indeed work for simple problems, as Madras et al. (2018) before us have, we show that it lacks scalability due to conflict between the components of the loss. Since information about s is only available to the decoder as a binary encoding, if the relationship between s and x is highly non-linear and unsummarisable by a simple on-off mechanism, as is the case if s is an attribute such as gender, off-loading information to the decoder by conditioning

is no longer possible. As a result, z_u is forced to carry information about s in order to minimise the reconstruction error.

The obvious solution to this is to allow the encoder to store information about s in a partition of the latent space as in Creager et al. (2019). However, we question whether an AE is the best choice for this setup, with the view that an invertible model is the better tool for the task. Using an invertible model affords several guarantees, principal of which being complete that of information-preservation and freedom from a reconstruction loss, the importance of which we expatiate on below.

3.3.4 *Conditional Flow*

INVERTIBLE NEURAL NETWORKS. INNs embody a subclass of neural networks characterised by a bijective mapping between their inputs and output (Dinh et al., 2014). The transformations are designed such that their inverses and Jacobians are exactly and efficiently computable. These flow-based models permit *exact* likelihood estimation (Rezende and Mohamed, 2015) through the warping of a base density with a series of invertible transformations and computing the resulting, highly multi-modal, but still normalised, density, using the change-of-variable theorem:

$$\log p(x) = \log p(z) + \sum \log \left| \det \left(\frac{\mathrm{d}h_i}{h_{i-1}} \right) \right|, \quad p(z) = \mathcal{N}(z; 0, \mathbb{I}), \tag{3.5}$$

where h_i denotes the output of the ith layer of the network and $p(z)$ is the base density, which is again an Isotropic Gaussian. Training the INN then reduces to maximising $\log p(x)$ over the training set, i.e. maximising the probability the network assigns to samples in the training set.

THE BENEFITS OF BIJECTIVITY. Using an network to generate our encoding, z_u, carries a number of advantages over other approaches. Ordinarily, the main benefit of flow-based models is that they permit exact density estimation. However, since we are not interested in sampling from the model's distribution, in our case the likelihood term serves as a regulariser, in the same vein as Jacobsen et al. (2018). Critically, this forces the mean of each latent dimension to zero, thereby enabling null-sampling. The invertible property of the network guarantees the preservation of all information relevant to y which is independent of s, regardless of how it is allocated in the output space. Secondly, we conjecture that the encodings are more robust to OOD data. Whereas an auto-encoder (AE) could map a previously seen input and a previously unseen input to the same representation, an invertible network sidesteps this due to the network's bijective property, ensuring all relevant information is stored somewhere. This opens up the possibility of transfer learning between datasets with a similar manifestation of s, as we demonstrate §3.6.7.

Under our framework, the invertible network f maps the inputs x to a representation $z \triangleq f(x)$. We interpret the representation z as being the concatenation of two subembeddings, namely $z \triangleq [z_u, z_b]$. The dimensionality of z_b (and z_u, by complement) is a free parameter (see §3.6.3 for tuning strategies). As f is invertible, x can be recovered as such:

$$x = f^{-1}([z_u, z_b]) \tag{3.6}$$

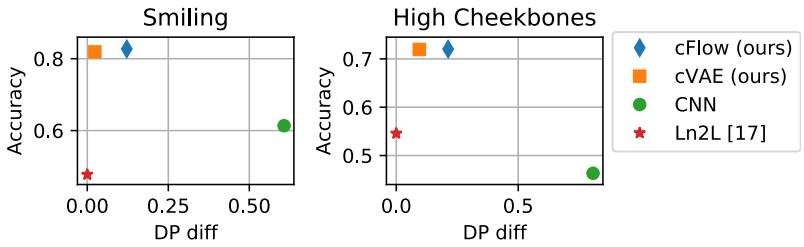


Figure 3.2: Performance of our model for different targets (mixing factor $\eta = 0$). Left: *Smiling* as target, right: *high cheekbones*. *DP diff* measures fairness with respect to demographic parity. A perfectly fair model has a *DP diff* of 0.

where z_b is required for equality of the output dimension and input dimension to satisfy the bijectivity of the network – we cannot output z_u alone, but have to output z_b as well. In order to generate the pre-image of z_u, we perform null-sampling w.r.t. z_b by zeroing-out the elements of z_b (such that $x_u \triangleq f^{-1}([z_u, \mathbf{0}])$), i.e. setting them to the mean of the prior density, $\mathcal{N}(z; 0, I)$.

How can ensure that z_u contains the information about y necessary for downstream classification? The importance of the invertible architecture bears out from this consideration, for so long as z_b does not contain the information about y, z_u necessarily must. We can then raise or lower the information capacity of z_b by adjusting its dimensionality, $\dim(z_b)$; practically, it should be set to the smallest size sufficient to capture all information about s, so as not to sacrifice class-relevant information. §3.6.2 explores the influence of $\dim(z_b)$ empirically.

3.4 Experiments

We present experiments to demonstrate that the null-sampled representations are in fact invariant to s while still allowing a classifier to predict y from them. We run our cVAE and cFlow models on the coloured MNIST (cMNIST) and CelebA dataset, which we artificially bias, first describing the sampling procedure we follow to do so for non-synthetic datasets. As baselines we have the model of Kim et al. (2019) (Ln2L) and the same convolutional neural network (CNN) used to evaluate the cFlow and cVAE models but with the unmodified images as input (CNN). For the cFlow model we adopt a Glow-like architecture (Kingma and Dhariwal, 2018), while both sub-networks of the cVAE model comprise gated convolutions (Oord et al., 2016), where the encoding size is 256. For cMNIST, we construct the Ln2L baseline according to its original description, for CelebA, we treat it as an augmentation of the baseline CNN's objective function. Detailed information regarding model architectures can be found in §3.6.1 and §3.6.3. [1]

3.4.1 *Synthesising Dataset Bias*

For our experiments, we require a training set that exhibits a strong spurious correlation, together with a test set that does not. For cMNIST, this is easily satisfied as we have complete control over the data generation process. For CelebA and UCI Adult, on the other hand, we have to generate the split from the existing data. To this end, we first set aside a randomly selected portion of the dataset from which to sample the biased dataset The portion itself is then split further into two

1 Code can be found at https://github.com/wearepal/nifr.

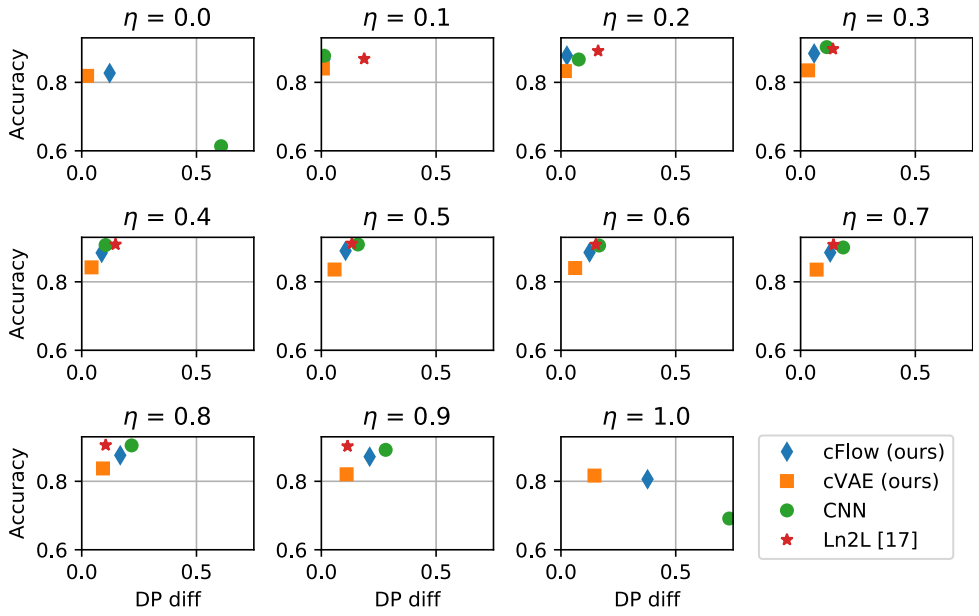


Figure 3.3: Performance of our model for the target "smiling" for different mixing factors η. *DP diff* measures fairness with respect to demographic parity. A perfectly fair model has a *DP diff* of 0, thus the closer to top-left the better it is in terms of we accuracy-fairness trade-off. Only values $\eta = 0$ and $\eta = 1$ correspond to the scenario of a strongly biased training set. The results for $0.1 \leq \eta \leq 0.9$ are to confirm that our model does not harm performance for non-biased training sets.

parts: one in which $(s = -1 \wedge y = -1) \vee (s = +1 \wedge y = +1)$ holds true for all samples, call this part $\mathcal{D}_{eq}$, and the other part, call it $\mathcal{D}_{opp}$, which contains the remaining samples. To investigate the behaviour at different levels of correlation, we mix these two subsets according to a mixing factor η. For $\eta \leq \frac{1}{2}$, we combine (all of) $\mathcal{D}_{eq}$ with a fraction of 2η from $\mathcal{D}_{opp}$. For $\eta > \frac{1}{2}$, we combine (all of) $\mathcal{D}_{opp}$ and a fraction of $2(1-\eta)$ from $\mathcal{D}_{eq}$. Thus, for $\eta = 0$, the biased dataset is just $\mathcal{D}_{eq}$, for $\eta = 1$ it is just $\mathcal{D}_{opp}$ and for $\eta = \frac{1}{2}$ the biased dataset is an ordinary subset of the whole data. The test set is simply the data remaining from the initial split.

3.4.2 *Evaluation protocol*

We evaluate our results in terms of accuracy and fairness. A model that perfectly decouples its predictions from s will achieve near-uniform accuracy across all biasing-levels. For binary s/y we quantify the fairness of a classifier's predictions using *demographic parity* (DP): the absolute difference in the probability of a positive prediction for each sensitive group.

3.4.3 *Experimental results*

We report the results from two image datasets. cMNIST, a synthetic dataset, is a good starting point for evaluating our model due to the direct control we have over the biasing. CelebA, on the other hand, is a more practical and challenging example. We also test our method on a tabular dataset, the Adult dataset.

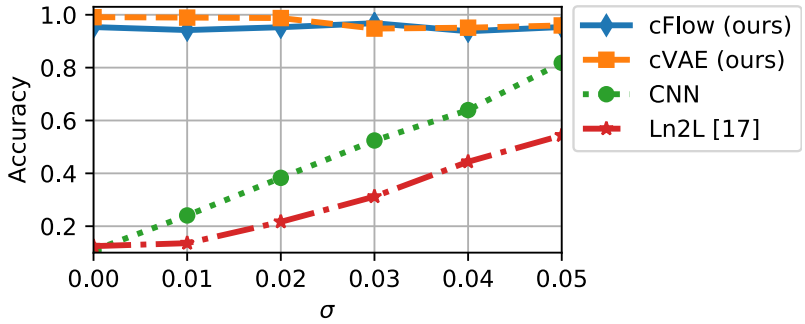
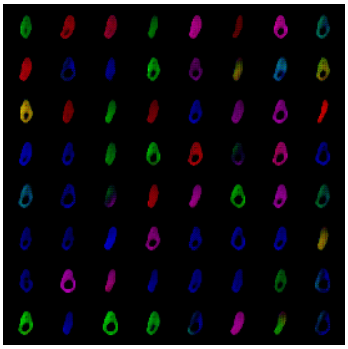


Figure 3.4: Accuracy of our approach in comparison with other baseline models on the cMNIST dataset, for different standard deviations (σ) for the colour sampling.

(a) Samples from the cMNIST training set, $\sigma = 0$.

(b) x_u null-samples from the cFlow model.



(c) x_b null-samples from the cFlow model.

Figure 3.5: Sample images from the coloured MNIST dataset problem with 10 predefined mean colours. (a): Images from the spuriously correlated subpopulation where colour is a reliable signal of the digit class-label. (b-c): Results of running our approach realised with cFlow on the cMNIST dataset. The model learns to retain the shape of the digit shape while removing the relationship with colour. A downstream classifier is now less prone to exploiting correlations between colour and the digit label class.

cMNIST. The coloured MNIST (cMNIST) dataset is a variant of the MNIST dataset in which the digits are coloured. In the training set, the colours have a one-to-one correspondence with the digit class. In the test set (and the representative set), colours are assigned randomly. The colours are drawn from Gaussians with 10 different means. We follow the colourisation procedure outlined by Kim et al. (2019), with the mean colour values selected so as to be maximally dispersed. The full list of such values can be found in §3.6.4. We produce multiple variants of the cMNIST dataset corresponding to different standard deviations σ for the colour sampling: $\sigma \in \{0.00, 0.01, ..., 0.05\}$.

For this specific dataset, we can establish an additional baseline by simply grey-scaling the dataset which only leaves the luminosity as spurious information. We also evaluate the model, with all the associated hyperparameters, from Kim et al. (2019). The only difference between the setups is the dataset creation, including the range of σ values we consider. Our versions of the dataset, on the whole, exhibit much stronger colour bias, to the point of the mapping the digit's colour and class being bijective. Fig. 3.4 shows that the model significantly underperforms even the naïve baseline, aside from at $\sigma = 0$, where they are on par.

Inspection of the null-samples shows that both the cVAE and cFlow model succeed in removing almost all colour information, which is supported quantitatively by Fig. 3.4, and qualitatively by Fig. 3.5. While the cVAE outperforms cFlow marginally at low σ values, performance degrades

as this increases. This highlights the problems with the conditional decoder we anticipated in §3.3.3. The lower σ, and therefore the variation in sampled colour, is, the more reliably the s label, corresponding to the mean of RGB distribution, encodes information about the colour. For higher σ values, the sampled colours can deviate far from the mean and so the encoder must incorporate information about s into its representation if it is to minimise the reconstruction loss. cFlow, on the other hand, is consistent across σ values.

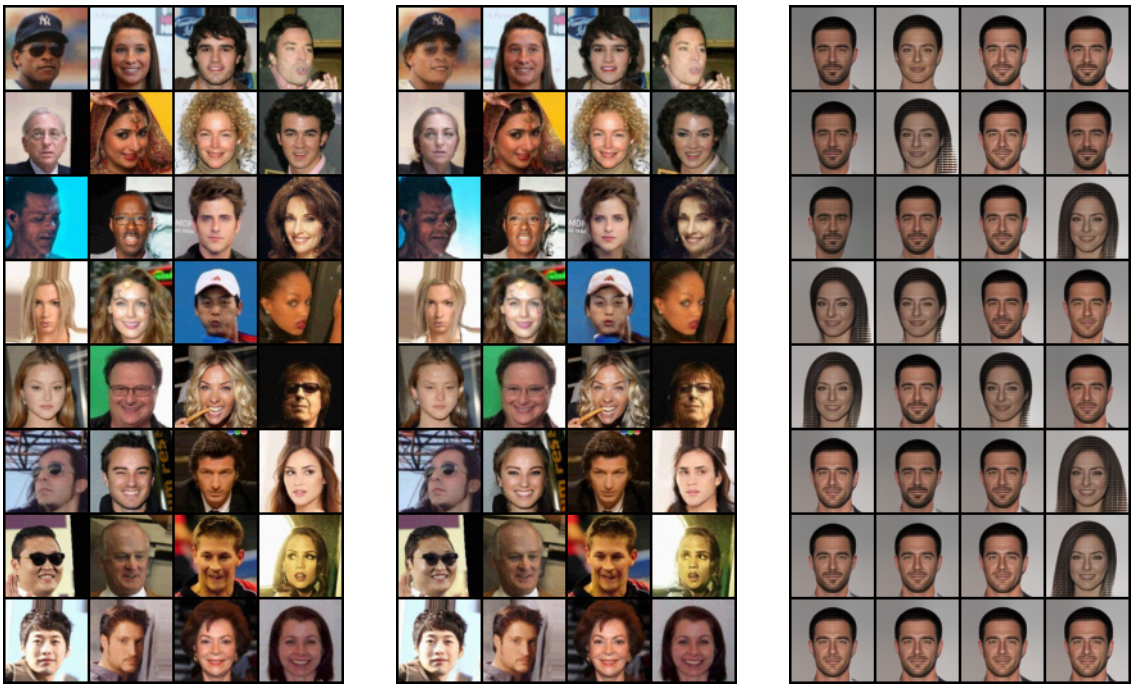
CelebA. To evaluate the effectiveness of our framework on real-world image data we use the CelebA dataset (Liu et al., 2015), consisting of 202,599 celebrity images. These images are annotated with various binary physical attributes, including "gender", "hair colour", "young", etc., from which we select our sensitive and target attributes. The images are centre cropped and resized to 64×64, as is standard practice. For our experiments, we designate "gender" as the sensitive attribute, and "smiling" and "high cheekbones" as target attributes. We chose gender as the sensitive attribute as it a common sensitive attribute in the fairness literature. For the target attributes, we chose attributes that are harder to learn than gender and which do not correlate too strongly with gender in the dataset ("wearing lipstick" for example being an attribute too closely correlated with gender). The model is trained on the representative set (normal subset of CelebA) and is then used to encode the artificially biased training set and the test set. The results for the most strongly biased training set ($\eta = 0$) can be found in Fig. 3.2. Our method outperforms the baselines in accuracy and fairness.

We also assess performance for different mixing factors (η) which correspond to varying degrees of bias in the training set (see Fig. 3.3). This is to verify that the model does not *harm* performance when there is not much bias in the training set. For these experiments, the model is trained once on the representative set and is then used to encode different training sets. The results show that for the intermediate values of η, our model incurs a small penalty in terms of accuracy, but at the same time makes the results *fairer* (corresponding to an accuracy-fairness trade-off). Qualitative results can be found in Fig. 3.6 (images from cVAE can be found in §3.6.6).

To show that our method can handle multinomial, as well as binary, sensitive attributes, we also conduct experiments with $s =$ hair colour as a ternary attribute ("Blonde", "Black", "Brown"), excluding "Red" because of the paucity of samples and the noisiness of their labels. The results for these experiments can be found in §3.6.2.

Results for the UCI Adult dataset. The UCI Adult dataset consists of census data and is commonly used to evaluate models focused on algorithmic fairness. Following convention, we designate "gender" as the sensitive attribute s and whether an individual's salary is \$50,000 or greater as y. We show the performance of our approach in comparison to baseline approaches in Fig. 3.7. We evaluate the performance of all models for mixing factors (η) 0 and 1. Results shown in Fig. 3.7 show that we match or exceed the baseline. In terms of fairness metrics, our approach generally outperforms the baseline models for both of η. Detailed results can be found in §3.6.2.

We also did experiments to show that the encoder transfers to other tasks. These transfer-learning experiments can be found in §3.6.7.



(a) Original images.

(b) $\boldsymbol{x}_u$ null-samples from the cFlow model.

(c) $\boldsymbol{x}_b$ null-samples from the cFlow model.

Figure 3.6: CelebA null-samples learned by our cFlow model, with gender as the sensitive attribute. (a) The original, untransformed samples from the CelebA dataset (b) Reconstructions using only information unrelated to s. (c) Reconstruction using only information related to $\neg s$. The model learns to disentangle gender from the non-gender related information. Note that some attributes like skin tone seem to change along with gender due to the correlation between the attributes. This is especially visible in images (1,1) and (3,2). Only because our representations are produced in the data-domain can we easily spot such instances of entanglement.

3.5 Conclusion

We have proposed a general and straightforward framework for producing invariant representations, under the assumption that a representative but partially-labelled *representative* set is available. Training consists of two stages: an encoder is first trained on the representative set to produce a representation that is invariant to a designated spurious feature. This is then used as input for a downstream task-classifier, the training data for which might exhibit extreme bias with respect to that feature. We train both a VAE- and INN-based model according to this procedure, and show that the latter is particularly well-suited to this setting due to its losslessness. The design of the models allows for representations that are in the data domain and therefore

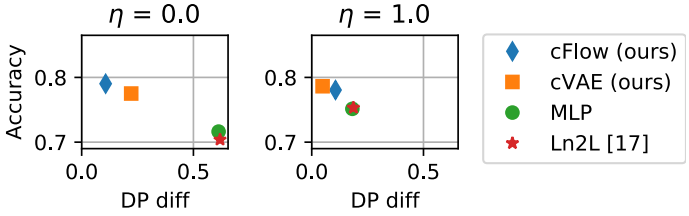


Figure 3.7: Results for the Adult dataset. The x-axis corresponds to the difference in positive rates. An ideal result would occupy the top-left.

exhibit meaningful invariances. We characterise this for synthetic as well as real-world datasets for which we develop a method for simulating sampling bias.

Table 3.1: INN architecture used for each dataset.

Dataset	Levels	Level depth	Coupl. chan.	Input to discr.
UCI Adult	1	1	35	Null-samples
cMNIST	3	16	512	Encodings
CelebA	3	32	512	Encodings

Table 3.2: cVAE encoder architecture used for each dataset. The decoder architecture in each case mirrors that of its encoder counterpart through use of transposed convolutions. For the adult dataset we apply ℓ_2 and cross-entropy losses to the reconstructions of the continuous features and discrete features, respectively.

Dataset	Initial channels	Levels	β	Recon. loss
UCI Adult	35	–	0	ℓ_2 + CE
cMNIST	32	4	0.01	ℓ_2
CelebA	32	5	1	ℓ_1

3.6 Appendix

3.6.1 *Model Architectures*

For both cMNIST and CelebA we parametrise the coupling layers with the same convolutional architecture as in Kingma and Dhariwal (2018), consisting of 3 convolutional layers each with 512 filters of, in order, sizes 3×3, 1×1, and 3×3. Following Ardizzone et al. (2019), we Xavier initialise all but the last convolutional layer of the s and t sub-networks which itself is zero-initialised so that the coupling layers begin by performing an identity transform. We use a Glow-like architecture (Kingma and Dhariwal, 2018) (affine coupling layers together with chequerboard reshaping and invertible 1×1 convolutions) for the convolutional INNs. Table 3.1 summarises the INN architectures used for each dataset.

For the image datasets each level of the cVAE encoder consists of two gated convolutional layers (Oord et al., 2016) with ReLU activation. At each subsequent level, the number of filters is doubled, starting with an initial value 32 and 64 in the case of CelebA and cMNIST respectively. In the case of the Adult dataset, we use an encoder with one fully-connected hidden layer of width 35, followed by SeLU activation (Klambauer et al., 2017). For both cMNIST and CelebA, we downsample to a feature map with spatial dimensions 8×8, but with 3 and 16 channels respectively. For the Adult dataset, the encoding is a vector of size 35. The output layer specifies both the parameters (mean and variance) of the representation's distribution. In all cases the KL-divergence is computed with respect to a standard isotropic Gaussian prior. Details of the encoder architectures can be found in table 3.2. The loss pre-factors were sampled from a logarithmic scale; without proper balancing the networks can exhibit instability, especially during the early stages of training.

3.6.2 *Additional results*

DETAILED RESULTS FOR UCI ADULT DATASET. This census data is commonly used to evaluate models focused on algorithmic fairness. Following convention, we designate "gender" as the s and whether an individual's salary is \$50,000 or greater as y. We show the performance of our approach in comparison to baseline approaches in figure 3.8. We evaluate the performance of all models for mixing factors (η) of value $\{0, 0.1, \ldots, 1\}$. Results shown in figure 3.8 show that while our model fails to surpass the baseline models in terms of accuracy for the balanced case (and those close to it), we match or exceed the baseline as η moves the dataset to a more imbalanced setting. In terms of fairness metrics, our approach generally outperforms the baseline models regardless of η.

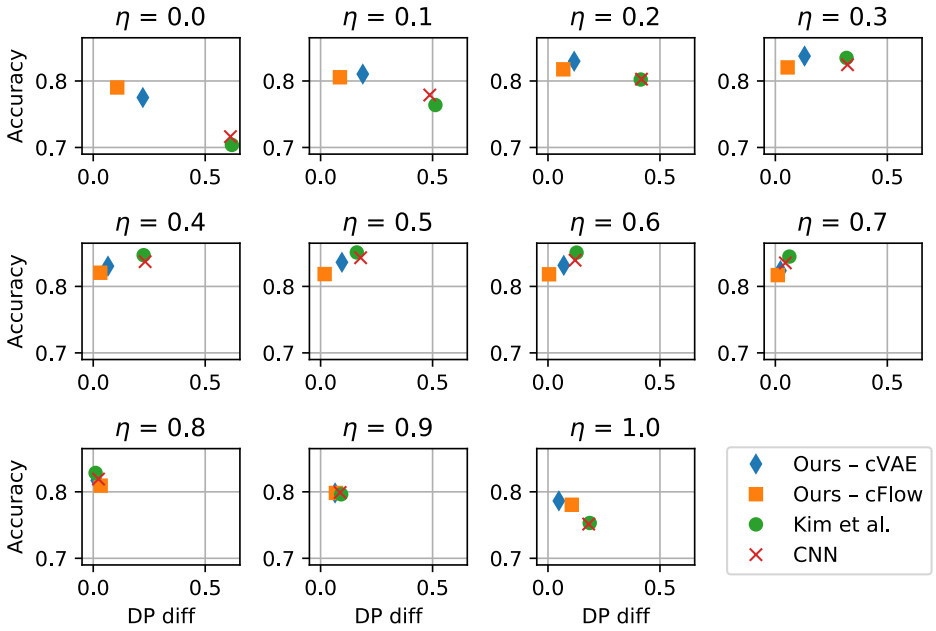


Figure 3.8: Results for the ADULT dataset. The x-axis corresponds to the difference in positive rates. An ideal result would occupy the TOP-LEFT.

MULTINOMIAL SENSITIVE ATTRIBUTES. In addition to binary sensitive attribute s, we also investigate multinomial s in the CelebA dataset. First, we do experiments with hair colour, where s has three possible values: blond hair, brown hair and black hair. The other experiment is with a combination of age and gender, where s has four possible values, each of which is a combination of a gender and an age: Young/Female, Young/Male, Old/Female and Old/Male. To evaluate the fairness for multinomial s, we use Hirschfeld-Gebelein-Rényi maximal correlation (HGRMC, (Mary et al., 2019)), defined on the domain $[0, 1]$ and yielding $\text{HGR}(Y, S) = 0$ iff $Y \perp S$, 1 if there is a deterministic mapping between the variables. Results can be found in Fig. 3.9.

INVESTIGATION INTO THE SIZE OF z_b. In the cFlow model, the size of z_b is an important hyperparameter which can affect the result significantly. Here we investigate the sensitivity of the model to the choice of z_b size. Table 3.3 shows accuracy and fairness (as measured by *DP*

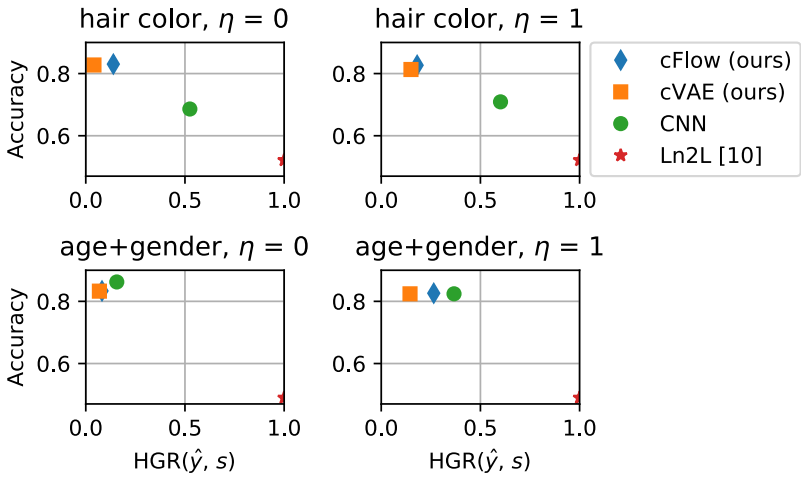


Figure 3.9: For *hair colour*, s takes on the values Blond, Brown and Black. For *age+gender*, s takes on the values Young/Female, Young/Male, Old/Female and Old/Male.

Table 3.3: Results on the CelebA dataset with different sizes of z_b.

$\lvert z_b \rvert$	$\lvert z_b \rvert / \lvert z \rvert$	Accuracy	DP diff
1	0.0082%	0.60	0.63
3	0.0245%	0.60	0.63
5	0.0410%	0.84	0.12
10	0.0820%	0.84	0.12
30	0.2442%	0.74	0.23
50	0.4070%	0.68	0.27

diff) for different sizes of z_b. The results show that both too large and too small z_b is detrimental. However, they also show that the model is not overly sensitive to this parameter: both sizes 5 and 10 achieve nearly identical results.

ADDITIONAL FAIRNESS METRICS. In addition to *DP diff*, we report here the result from other fairness measures. These results are from the same setup as those reported in the main paper. We report the difference in TPRs between the two groups (male and female), which corresponds to a measure of equalised odds (EqOd), and the difference in true negative rates (TNRs) between the two groups.

Table 3.4: Additional fairness metrics for the experiments on the CelebA dataset (Fig. 3.3 from the main text). *TPR diff.* refers to the difference in true positive rate. *TNR diff.* refers to the difference in true negative rate. LEFT: $\eta = 0$. RIGHT: $\eta = 1$.

Method	Accuracy	DP diff	TPR diff	TNR diff
cFlow	0.83	0.10	0.15	0.25
cVAE	0.82	0.05	0.09	0.18
CNN	0.61	0.63	0.70	0.64
Ln2L	0.52	0.00	0.00	0.00

Method	Accuracy	DP diff	TPR diff	TNR diff
cFlow	0.82	0.33	0.28	0.21
cVAE	0.81	0.16	0.10	0.05
CNN	0.67	0.75	0.66	0.76
Ln2L	0.51	0.08	0.06	0.09

3.6.3 *Optimisation Details*

All our models were trained using the RAdam optimiser (Liu et al., 2020) with learning rates 3×10^{-4} and 1×10^{-3} for the encoder/discriminator pair and classifier respectively. A batch size of 128 was used for all experiments.

We now detail the optimisation settings, including the choice of adversary, specific to each dataset. Details of the cVAE and cFlow architectures can be found in table 3.2 and table 3.1, respectively.

UCI ADULT. For this dataset our experiment benefited from using null-samples as inputs to the adversary of the cFlow model. Unlike for the image datasets, we found a single adversary to be sufficient. This was realised as a multi-layer perceptron (MLP) with one hidden layer, 256 units wide. The INN performs a bijection of the form $f : \mathbb{R}^n \to \mathbb{R}^n$. However, the adult dataset is composed of mostly discrete (binary/categorical) features. To achieve good performance, we found it necessary to first pre-process the inputs with a pretrained autoencoder, using its encodings as the input to the cFlow model, as well as to the adversary. The learned representations were evaluated with a logistic regression model from scikit-learn (Pedregosa et al., 2011), using the standard settings. All baseline models were trained for 200 epochs. The Ln2L (Kim et al., 2019) and MLP baselines share the architecture of the cVAE's encoder, only with a classification layer affixed.

COLOURED MNIST. Each level of the architecture used for the downstream classifier and naïve baseline alike consists of two convolutional layers, each with kernel size 3 and followed by Batch Norm (Ioffe and Szegedy, 2015) and ReLU activation. For the Ln2L baseline, we use an a setup identical to that described in Kim et al. (2019). Each level has twice the number of filters in its convolutional layer and half the spatial input dimensions as the last. The original input is downsampled to the point of the output being reduced to a vector, to which a fully-connected classification layer is applied.

To allow for an additional level in the INN (the downsampling operations requiring the number of spatial dimensions to be even), the data was zero-padded to a size of 32×32. The cVAE and cFlow models were trained for 50 and 200 epochs respectively, using ℓ_2 reconstruction loss for the former. The downstream classifier and all baselines were trained for 40 epochs. For both of our models, an ensemble of 5 adversaries was applied to the encodings, with each member taking the form of a fully-connected ResNet, 2 blocks in depth, with SeLU activation (Klambauer et al., 2017). The adversaries were reinitialised independently with probability 0.2 at the end of each epoch. While the adversaries could equally well take null-samples as input, as done for the Adult dataset, doing so requires the performing of both forward and inverse passes each iteration, which, for the convolutional INNs of the depths we require for the image datasets, introduces a large computational overhead, while also showing to be the less stable of the two approaches in our preliminary experiments.

CELEBA. The downstream classifier and naïve baseline take the same form as described above for cMNIST, but with an additional level with 32 filters in each of its convolutions at the top of the

network. For this dataset we adapt the Ln2L model by simply considering it as an augmentation the naïve baseline's objective function, with the entropy loss applied to the output of the final convolutional layer. These models were again trained for 40 epochs, which we found to be sufficient for convergence for the tasks in question. The cVAE and cFlow models were respectively trained for 100 epochs and 30 epochs, using ℓ_1 reconstruction loss for the former. Compared with cMNIST, the size of the adversarial ensemble was increased to 10, the reinitialisation probability to 0.33, but no changes were made to the architectures of its members.

THE PITFALLS OF ADVERSARIAL TRAINING. Adversarial learning has become one of the go-to methods for enforcing invariance in fair representation learning (Ganin et al., 2016) with maximum mean discrepancy (MMD) (Louizos et al., 2016) and HSIC (Quadrianto et al., 2019), being popular non-parametric alternatives. Ganin et al. (2016) proposed adversarial learning for domain adaptation (DA) problems, with Edwards and Storkey (2016) soon after making this and learning a representation promoting DP. The adversarial approach carries the benefits of being both efficient and scalable to multi-class categorical variables, which many sensitive attributes are in practice, whereas the non-parametric methods only permit pair-wise comparison.

However, when realised as a neural network, the adversary is both sensitive to the values of the inputs as well as their ordering (though exchangeable architectures, such as Zaheer et al. (2017) do exist, but which sacrifice expressiveness). Thus, it can happen that the representation learner optimises for the surrogate objective of eluding the adversary rather than the real objective of expelling s-related information. Moreover, the non-stationarity of the dynamics can lead to cyclic equilibria, irrespective of the capacity of the adversary.

When working with a partitioned latent space, this behaviour can be averted by instead encouraging z_b to be predictive of s, acting as a kind of information "sink", as in Jacobsen et al. (2018). However, this does not have the guarantee of making z_u invariant to s - there are often many indicators for s, not all of which are needed to predict the label perfectly. Training the network to convergence before taking each gradient step with the representation learner is one way one to attempt to tame the unstable minimax dynamics (Feng et al., 2019). However, this does not prevent the emergence of the aforementioned cyclicity.

We try to mitigate the aforementioned degeneracies by maintaining a diverse set of adversaries, as has shown to be effective for GAN training (Durugkar et al., 2017), and by decorrelating the individual trajectories by intermittently re-initialising them with some small probability following each iteration.

TUNING THE PARTITION SIZES. There are several ways of ensuring that the size of z_b is sufficient to capture all s dependencies, but minimal enough that information unrelated to s is maximally preserved We adopt the straightforward search strategy of, starting from some initial guess, calibrating the value according to accuracy attained by a classifier trained to predict s from z_b on a held-out subset of the representative set, which is measured whenever the adversarial loss plateaus. If the accuracy is above chance level then that suggests the size of the z_b partition, $|z_b|$, needs to be increased to accommodate more information about s. If the accuracy is found to be at chance level then are two possibilities: 1) $|z_b|$ is already optimal; 2) $|z_b|$ is large enough that it fully contains both information s as well as that of a portion of y. If the former is true,

Table 3.5: Mean RGB values (in practice normalised to $[0, 1]$) parametrising the Multivariate Gaussian distributions from which each digit's colour is sampled in the biased (training) dataset. In the representative and test sets, the colour of each digit is sampled from one of the specified Gaussian distributions at random.

Digit	Colour Name	Mean RGB
0	Cyan	(0, 255, 255)
1	Blue	(0, 0, 255)
2	Magenta	(255, 0, 255)
3	Green	(0, 128, 0)
4	Lime	(0, 255, 0)
5	Maroon	(128, 0, 0)
6	Navy	(0, 0, 128)
7	Purple	(128, 0, 128)
8	Red	(255, 0, 0)
9	Yellow	(255, 255, 0)

then perturbations around the current value allow us to confirm this; if the latter is true then decreasing the value was indeed the correct decision.

3.6.4 *Synthesising Coloured MNIST*

We use a colourised version of MNIST as a controlled setting investigate learning from biased data in the image domain. In the biased training set, each digit is assigned a unique mean RGB value parametrising the multivariate Gaussian from which its colour is drawn. These values were chosen to be maximally dispersed across the 8-bit colour spectrum and are listed in table 3.5. By adjusting the standard deviation, σ, of the Gaussians, we adjust the degree of bias in the dataset. When $\sigma = 0$, there is a perfect and noiseless correspondence between colour and digit class which a classifier can exploit. The classifier can favour the learning of the low-level spurious feature over those higher level features constituent of the digit's class. As the standard deviation increases, the sampled RGB values are permitted to drift further from the mean, leading to overlap between the samples of the colour distributions and reducing their reliability as indicators of the digit class. In the test and representative sets alike, however, the colour of each sample is sampled from one of the 10 distributions randomly, such that colour can no longer be leveraged as a shortcut to predicting the digit's class.

3.6.5 *Stabilising the Coupling layers*

Heuristically, we found that applying an additional non-linear function to the scale coefficient of the form

$$s = \sigma(f(u)) + 0.5 \tag{3.7}$$

greatly improved the stability of the affine coupling layers. Here, σ is the logistic function, which we shift to be centred on 1 so that zero-initialising f results in the coupling layers initially performing an identity-mapping.

(a) Original images.

(b) x_u null-samples generated by the cVAE model.

(c) x_b null-samples generated by the cVAE model.

Figure 3.10: CelebA null-samples learned by our cVAE model, with gender as the sensitive attribute. (a) The original, untransformed samples from the CelebA dataset (b) Reconstructions using only information unrelated to s. (c) Reconstruction using only information related to $\neg s$. The model learns to disentangle gender from the non-gender related information. Compared with the cFlow model, there is a severe degradation in reconstruction quality due to the model trying to simultaneously satisfy conflicting objectives.

3.6.6 *Qualitative Results for CelebA*

Learning a representation alongside its inverse mapping, be it approximate or exact, enables us to probe the behaviour of the model that produced it, and any biases it may have implicitly captured due to entanglement between the sensitive attribute and other attributes present in the data. We highlight a few examples of such biases manifesting in the cFlow model's CelebA null-samples in Fig. 3.11. In these cases, make-up and hair style have been inadvertently modified during the null-sampling due to the tight correlation between these two attributes and the sensitive attribute, gender, to which we had aimed to make our representations invariant. Additionally, in all highlighted images, the skin tone has changed: from male to gender-neutral, the skin becomes lighter and from female to gender-neutral, the skin becomes darker; in the change from male to gender-neutral, glasses are also often removed. As the model cannot know that the label is meant to only refer to gender, and not to these other (correlated) attributes, the links cannot be disentangled by the model. However, the advantage of our method is that we can at least identify such biases due to the interpretability that comes with the representations being in the data domain.

(a) Original images.

(b) x_u null-samples generated by the cFlow model.

(c) $\mathbf{x}_b$ null-samples generated by the cFlow model.

Figure 3.11: CelebA null-samples learned by our cFlow model, with gender as the sensitive attribute. (a) The original, untransformed samples from the CelebA dataset (b) Reconstructions using only information unrelated to s. (c) Reconstruction using only information related to $\neg s$. The model learns to disentangle gender from the non-gender related information. Attributes such as *make-up* and *hair length* are also often modified in the process (prime examples framed with red) due to inherent correlations between them and the sensitive attribute, which the interpretability of our representations allows us to easily identify.

3.6.7 *Transfer Learning*

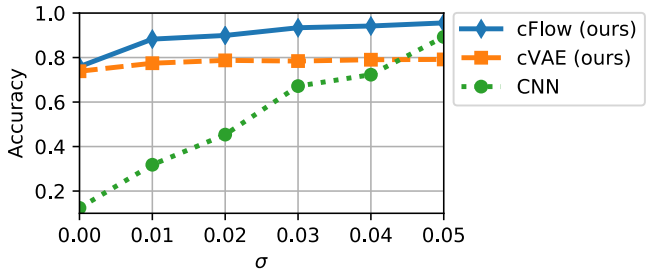
For our method, we require a representative set which follows the same distribution as that observed during deployment. Such a representative set might not always be available. In such a scenario, we can resort to using a set that is merely *similar* to that in the deployment setting and leverage transfer learning.

One of the advantages of using an invertible architecture over conventional, *surjective* ones that we stressed in the main text is its *losslessness*. Since the transformations are necessarily bijective, the information contained in the input can never be destroyed, only redistributed. This makes such models particularly well-suited, in our minds, for transferring learned invariances: even if the input is unfamiliar, no information should be lost when trying to transform it. This works as long as only the information about s ends up in the z_b partition. If s takes a form similar to that which we pre-trained on, and can thus be correctly partitioned in the latent space, by complement we have the information about $\neg s$ stored in the z_u partition, without presupposing similarity to the $\neg s$ observed during pre-training.

TRANSFERRING FROM MIXED-NIST TO MNIST. We test our hypothesis by comparing the performance of the cFlow and cVAE models pre-trained on a mixture of datasets belonging to the NIST family, colourised in the same way as cMNIST, while the downstream train and test sets remain

the same as in the original cMNIST experiments. Specifically, we create this representative set by sampling 24,000 images (to match the cardinality of the original representative set) from EMNIST (letters only; Cohen et al., 2017), FashionMNIST (Xiao et al., 2017) and KMNIST (Clanuwat et al., 2018), in equal proportion. We use the same architectures for the cVAE and cFlow models as we did in the non-transfer learning setting. In terms of hyperparameters, the only change made was to the KL-divergence's pre-factor, finding it necessary to increase it to 1 to guarantee stability.

The results for the range of σ values are shown in Fig. 3.12a. Unsurprisingly, the performance of both models suffers when the representative and test sets do not completely correspond. However, the cFlow model consistently outperforms the cVAE model, with the gap increasing as the bias decreases. Although some colour information is retained in the cFlow null-samples, symptomatic of an imperfect transfer, semantic information is almost entirely retained as well. Conversely, the cVAE is very much flawed in this respect; as can be seen in the bottom row of Fig. 3.12a, for some samples, semantic information is degraded to the point of the digit's identity being altered. As a result of this semantic degradation, the performance of the downstream classifier is curtailed by the noisiness of the digit's identity and is relatively unchanging across σ-values, in contrast to the monotonic improvement of that achieved on the cFlow null-samples.



(a) Performance on cMNIST test data after pre-training on the mixed NIST dataset.

(b) Test data input to the cFlow model.

(c) x_u null-samples generated by the cFlow model.

(d) Test data input to the cVAE model.

(e) x_u null-samples generated by the cVAE model.

Figure 3.12: Results for the transfer learning experiments in which the representative set consists of colourised samples from EMNIST, KMNIST, and FashionMNIST, while the downstream dataset remains as cMNIST. (a) Quantitative results for different σ-values. (b-c) Qualitative results for the cFlow model. (d-e) Qualitative results for the cVAE model. The qualitative results provide comparisons of the images before (left) and after (right) null-sampling. Note that for some of the cVAE samples, the clarity of the digits has clearly changed due to null-sampling, serving as an explanation for the non-increasing downstream performance.

3.7 Authorial contributions

T. Kehrenberg conceived of the idea of using an INN combined with a representative set to learn an invariant representation in the face of spurious correlations, ran many of the experiments – especially so in the initial stages – and wrote much of the original text and code.

I wrote a significant part of the code (being responsible for several refactorings as part of the debugging process), the much of the text, helped crystallise the initial idea, ran many of the experiments, aided in experimental analysis, and developed much of the technical tricks needed to train the INN stably within the adversarial framework (grappling with the issues later elucidated by Behrmann et al. (2021)).

O. Thomas aided in writing the code, in running the experiments, partook in discussion and analysis of results, helped writing and formatting the paper, and served as an unwavering source of optimism.

N. Quadrianto supervised the project, providing feedback on current progress and iterations of the paper and advising which directions to pursue.

BIBLIOGRAPHY

Pedregosa, F. et al. (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Zemel, Richard S., Yu Wu, Kevin Swersky, Toniann Pitassi and Cynthia Dwork (2013). 'Learning Fair Representations'. In: *International Conference on Machine Learning (ICML)*. Vol. 28. JMLR Workshop and Conference Proceedings, pp. 325–333.

Dinh, Laurent, David Krueger and Yoshua Bengio (2014). 'NICE: Non-linear Independent Components Estimation'. In: *International Conference on Learning Representations (ICLR)*.

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

Kingma, Diederik P. and Max Welling (2014). 'Auto-Encoding Variational Bayes'. In: *International Conference on Learning Representations (ICLR)*.

Ioffe, Sergey and Christian Szegedy (2015). 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'. In: *International Conference on Machine Learning (ICML)*. Vol. 37. JMLR Workshop and Conference Proceedings, pp. 448–456.

Liu, Ziwei, Ping Luo, Xiaogang Wang and Xiaoou Tang (2015). 'Deep Learning Face Attributes in the Wild'. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. DOI: 10.1109/ICCV.2015.425.

Rezende, Danilo Jimenez and Shakir Mohamed (2015). 'Variational Inference with Normalizing Flows'. In: *International Conference on Machine Learning (ICML)*. Vol. 37. JMLR Workshop and Conference Proceedings, pp. 1530–1538.

Edwards, Harrison and Amos J. Storkey (2016). 'Censoring Representations with an Adversary'. In: *International Conference on Learning Representations (ICLR)*.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky (2016). 'Domain-adversarial training of Neural Networks'. In: *Journal of Machine Learning Research* 17.1, pp. 2096–2030.

Louizos, Christos, Kevin Swersky, Yujia Li, Max Welling and Richard S. Zemel (2016). 'The Variational Fair Autoencoder'. In: *International Conference on Learning Representations (ICLR)*.

Oord, Aäron van den, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals and Alex Graves (2016). 'Conditional Image Generation with PixelCNN Decoders'. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 4790–4798.

Beutel, Alex, Jilin Chen, Zhe Zhao and Ed H. Chi (2017). 'Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations'. In: *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*.

Cohen, Gregory, Saeed Afshar, Jonathan Tapson and Andre Van Schaik (2017). 'EMNIST: Extending MNIST to handwritten letters'. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 2921–2926.

Durugkar, Ishan P., Ian Gemp and Sridhar Mahadevan (2017). 'Generative Multi-Adversarial Networks'. In: *International Conference on Learning Representations (ICLR)*.

Higgins, Irina, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner (2017). 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'. In: *International Conference on Learning Representations (ICLR)*.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr and Sepp Hochreiter (2017). 'Self-Normalizing Neural Networks'. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 971–980.

Xiao, Han, Kashif Rasul and Roland Vollgraf (2017). 'Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms'. In: *arXiv preprint arXiv:1708.07747*.

Zaheer, Manzil, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov and Alexander J. Smola (2017). 'Deep Sets'. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 3391–3401.

Adel, Tameem, Zoubin Ghahramani and Adrian Weller (2018). 'Discovering Interpretable Representations for Both Deep Generative and Discriminative Models'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 50–59.

Clanuwat, Tarin, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto and David Ha (2018). 'Deep learning for classical japanese literature'. In: *arXiv preprint arXiv:1812.01718*.

Jacobsen, Jörn-Henrik, Arnold W. M. Smeulders and Edouard Oyallon (2018). 'i-RevNet: Deep Invertible Networks'. In: *International Conference on Learning Representations (ICLR)*.

Jaiswal, Ayush, Rex Yue Wu, Wael Abd-Almageed and Prem Natarajan (2018). 'Unsupervised Adversarial Invariance'. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5097–5107.

Kallus, Nathan and Angela Zhou (2018). 'Residual Unfairness in Fair Machine Learning from Prejudiced Data'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 2444–2453.

Kim, Hyunjik and Andriy Mnih (2018). 'Disentangling by Factorising'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 2654–2663.

Kingma, Diederik P. and Prafulla Dhariwal (2018). 'Glow: Generative Flow with Invertible 1x1 Convolutions'. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10236–10245.

Madras, David, Elliot Creager, Toniann Pitassi and Richard S. Zemel (2018). 'Learning Adversarially Fair and Transferable Representations'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 3381–3390.

Xiao, Taihong, Jiapeng Hong and Jinwen Ma (2018). 'DNA-GAN: Learning disentangled representations from multi-attribute images'. In: *ICLR workshop*.

Zhang, Quanshi and Song-Chun Zhu (2018). 'Visual interpretability for Deep Learning: a survey'. In: *Frontiers of Information Technology & Electronic Engineering* 19.1, pp. 27–39.

Ardizzone, Lynton, Carsten Lüth, Jakob Kruse, Carsten Rother and Ullrich Köthe (2019). 'Guided Image Generation with Conditional Invertible Neural Networks'. In: *arXiv preprint arXiv:1907.02392*.

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

Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi and Richard S. Zemel (2019). 'Flexibly Fair Representation Learning by Disentanglement'. In: *International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research, pp. 1436–1445.

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

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel (2019). 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness'. In: *International Conference on Learning Representations (ICLR)*.

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík and Hanna M. Wallach (2019). 'Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?' In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, p. 600. DOI: 10.1145/3290605.3300830.

Jacobsen, Jörn-Henrik, Jens Behrmann, Richard S. Zemel and Matthias Bethge (2019). 'Excessive Invariance Causes Adversarial Vulnerability'. In: *International Conference on Learning Representations (ICLR)*.

Kim, Byungju, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim and Junmo Kim (2019). 'Learning Not to Learn: Training Deep Neural Networks With Biased Data'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9012–9020. DOI: 10.1109/CVPR.2019.00922.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz and Moritz Hardt (2019). 'Delayed Impact of Fair Machine Learning'. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 6196–6200. DOI: 10.24963/ijcai.2019/862.

Mary, Jérémie, Clément Calauzènes and Noureddine El Karoui (2019). 'Fairness-Aware Learning for Continuous Attributes and Treatments'. In: *International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research, pp. 4382–4391.

Quadrianto, Novi, Viktoriia Sharmanska and Oliver Thomas (2019). 'Discovering Fair Representations in the Data Domain'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8227–8236. DOI: 10.1109/CVPR.2019.00842.

Wick, Michael L., Swetasudha Panda and Jean-Baptiste Tristan (2019). 'Unlocking Fairness: a Trade-off Revisited'. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8780–8789.

Kehrenberg, Thomas, Zexun Chen and Novi Quadrianto (2020). 'Tuning Fairness by Balancing Target Labels'. In: *Frontiers in Artificial Intelligence* 3, p. 33. DOI: 10.3389/frai.2020.00033.

Liu, Liyuan, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao and Jiawei Han (2020). 'On the Variance of the Adaptive Learning Rate and Beyond'. In: *International Conference on Learning Representations (ICLR)*.

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

4 | ADDRESSING MISSING SOURCES WITH ADVERSARIAL SUPPORT-MATCHING

Authors:
Thomas Kehrenberg[1], Myles Bartlett[1], Viktoriia Sharmanska[1,2] & Novi Quadrianto[1,3,4]
Affiliations:
[1]Predictive Analytics Lab (PAL), University of Sussex, Brighton, United Kingdom
[2]Imperial College London
[3]BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning
[4]Monash University, Indonesia

Abstract

When trained on diverse labelled data, machine learning models have proven themselves to be a powerful tool in all facets of society. However, due to budgetary limitations, deliberate or non-deliberate censorship, and other problems during data collection and curation, the labelled training set might exhibit a systematic dearth of data for certain groups. This problem is particularly pertinent in medical imaging where the number of positive samples typically outweigh the number of negative samples by an orders of magnitude and certain demographics may be excluded on safety grounds (e.g. pregnant women) or due to socioeconomic biases. We investigate a scenario in which the absence of certain data is linked to the second level of a two-level hierarchy in the data. Inspired by the idea of protected groups from algorithmic fairness, we refer to the partitions carved by this second level as "subgroups"; we refer to combinations of subgroups and classes, or leaf nodes in aforementioned hierarchy, as *sources*. To characterize the problem, we introduce the concept of classes with *incomplete subgroup support*. The representational bias in the training set can give rise to spurious correlations between the classes and the subgroups which cause standard classification models to generalize poorly to unseen sources. To overcome this bias, we make use of an additional, diverse but unlabelled dataset, called the *deployment set*, to learn a representation that is invariant to subgroup. This is done by adversarially matching the support of the training and deployment sets in representation space using a set discriminator operating on sets, or *bags*, of samples. In order to learn the desired invariance, it is paramount that the bags are balanced by class; this is easily achieved for the training set, but requires using semi-supervised clustering for the deployment set. We demonstrate the effectiveness of our method on several datasets and realisations of the problem.

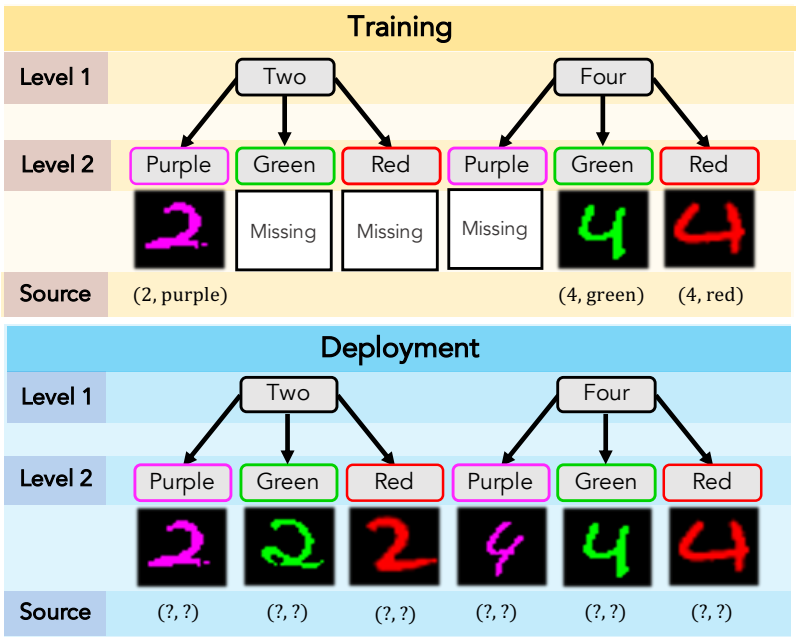


Figure 4.1: Illustration of our general problem setup. We assume the data follows a two-level hierarchy in which the first level corresponds to the class-level information (digit) and the second level corresponds to subgroup-level information (color). While all digits appear in the training set (Top), not all digit-color combinations (sources) do; these gaps in conditional support give rise to a spurious correlation between digit and color, where the former is completely determined by the latter in the training set (giving the mappings purple $\rightarrow$ 2 and green $\vee$ red $\rightarrow$ 4 as degenerate solutions to the classification problem), yet this same correlation does not hold for the deployment set (Bottom) which contains samples from the missing combinations. To disentangle the (spurious) subgroup- and class-related information, we make use of an additional dataset that is representative of the data the model is expected to encounter at deployment time, in terms of the sources present.

4.1 Introduction

Machine learning has burgeoned in the last decade, showing the ability to solve a wide variety of tasks with unprecedented accuracy and efficiency. These tasks range from image classification (Krizhevsky et al., 2012) and object detection (Ren et al., 2015), to recommender systems (Ying et al., 2018) and the modelling of complex physical systems such as precipitation (Ravuri et al., 2021) and protein folding (Jumper et al., 2021). In the shadow of this success, however, one finds less cause for optimism in frequent failure in equitability and generalisation-capability, failure which can have serious repercussions in high-stakes applications such as self-driving cars (Sun et al., 2019), judicial decision-making (Mayson, 2018), and medical diagnoses (AlBadawy et al., 2018). ML's data-driven nature is a double-edged sword: while it opens up the ability to learn patterns that are infeasibly complex for a practitioner to encode by hand, the quality of the solutions learned by these models depends primarily on the quality of the data with which they were trained. If the practitioner does not properly account for this, models ingesting data ridden with biases will assimilate, and sometimes even amplify, those biases. The problem boils down to not having sufficiently diverse annotated data, however collecting more labelled data is not always feasible due to temporal, monetary, legal, regulatory, or physical constraints.

While data can be intrinsically biased (such as in the case of bail records), *representational bias* is more often to blame, where socioeconomic or regulatory factors resulting in certain demographics

being under- (or even un-) represented. Clinical datasets are particularly problematic for ML due to the frequency of the different outcomes being naturally highly imbalanced, with the number of negative cases (`healthy`) typically greatly outweighing the number of positive cases (`diseased`); even if a subgroup is well-represented overall, that may well not be the case when conditioned on the outcome. Equally, it is entirely possible that certain subgroups may be completely absent. For example, pregnant women are often excluded from clinical trials due to safety concerns, and if they do participate it is often at too low a rate to be meaningful (Afrose et al., 2021).

Like many prior works (Kim et al., 2019; Sagawa et al., 2020; Sohoni et al., 2020; Creager et al., 2021), we consider settings where there is a two-level hierarchy, with the second level partitioning the data into *subgroups* that are causally independent of the class (constituting the first level) which is being predicted. This second level of the data is assumed to be predictable by the classifiers in the considered hypothesis class. In both Sohoni et al. (2020) and Creager et al. (2019) the entailed subgroups are unobserved and need to be inferred in a semi-supervised fashion. We consider a similar problem but one where the second level is partially observed. Specifically, we focus on problems where some outcomes are available for some subgroups and not for others. This particular form of the problem has – so far as we are aware – been hitherto overlooked despite pertaining to a number of real-world problems.

If the labelled training set is sufficiently balanced in terms of classes and subgroups, a standard ERM (empirical risk minimisation) classifier can achieve good performance. However, we consider the added difficulty that, in the labelled training set, some outcomes (classes) are not observed for all subgroups, meaning some of the classes do not overlap with all the subgroups. In other words, in the training set, some of the classes have *incomplete support* with respect to the subgroup partition, while in the deployment setting we expect all possible combinations of subgroup and class to appear. We illustrate our problem setup in Fig. 4.1, using Coloured MNIST digits as examples; here, the first level of the hierarchy captures digit class, the second level, colour. While the (unlabelled) deployment set contains all digit-colour combinations (or *sources*), half of these combinations are missing from the (labelled) training set. A classifier trained using only this labelled data would wrongly learn to classify 2s based on their being purple and 4s, based on their being green or red (instead of based on shape) and when deployed would perform no better than random due to the new sources being coloured contrary to their class (relative to the training set).

We address this problem by learning representations that are invariant to subgroups and that thus enable the model to ignore the subgroup partition and to predict only the class labels. In order to train an encoder capable of producing these representations, the information contained in the labelled training set alone is not sufficient to break the *spurious correlations*. To learn the "correct" representations, we make use of an additional unlabelled dataset with support equivalent to that of the deployment set (which includes the possibility of it being the actual deployment set). We do not consider this a significant drawback as such data is almost always far cheaper and less labour-intensive to procure than *labelled* data (which may require expert knowledge).

This additional dataset serves as the inductive bias needed by the encoder to disentangle class- and subgroup-related factors. The encoder is trained adversarially to produce representations whose source (`training` or `deployment`) is indeterminable to a set-classifier. To ensure subgroup- (not class-) invariance is learned, the batches fed to the discriminator need to be approximately

balanced, such that they reflect the support, and not the shape, of the distributions. We propose a practical way of achieving this based on semi-supervised clustering.

We empirically show that our proposed method can effectively disentangle subgroup and semantic factors on a range of classification datasets and is robust to noise in the bag-balancing, to the degree of outperforming the baseline methods even when no balancing of bags from the deployment set is performed. Furthermore, we prove that the entailed objective is theoretically guaranteed to yield representations that are invariant to subgroups and that we can bound the error incurred due to imperfect clustering.

4.2 Problem setup

In this section, we illustrate and formalise the problem of classes with incomplete subgroup-support. We start by defining requisite notation for conveying our setup and in §4.2.2 expand on this notation to construct a more general and compact description of said problem. Let $x \in \mathcal{X} \subset \mathbb{R}^d$, $y \in \mathcal{Y}$ and $s \in \mathcal{S}$ denote the observed input features, class labels and subgroup labels, respectively, with $\mathcal{Y}$ and $\mathcal{S}$ being non-empty, finite sets (i.e. $\mathcal{Y}, \mathcal{S} \in \{A | A \in \mathcal{P}(V), A \neq \varnothing, |A| < \aleph_0\}$), and with upper-case letters denoting observed variables' random-variable counterparts here (X, Y, and S) and throughout. We refer to the values, $g \in \mathcal{G}$ as *sources*, representing unique pairs of s and y, such that $\mathcal{G} \subseteq \mathcal{S} \times \mathcal{Y}$. As in a standard supervised learning task, we have access to a labelled training set $\mathcal{D}^{tr} \triangleq \{(x_n, s_n, y_n)\}_{n=1}^{N^{tr}} \subset (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})$, that is used to train a classifier $\Gamma : \mathcal{X} \to \mathcal{Y}$ that is then deployed on test set $\mathcal{D}^{te} \triangleq \{(x_n, s_n, y_n)\}_{n=1}^{N^{te}} \subset (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})$. We use superscript, to denote association of a domain with a correspondingly superscripted dataset, e.g. $\mathcal{G}^{tr}$ and $\mathcal{G}^{te}$ respectively denote the sources in the training and test sets. Lastly, for some functions, we abuse notation and allow the random and observed variables to be interchanged as inputs; we presuppose such functions (and their domain) are Borel measurable and thus preserve the type of variable, i.e. a function of a random variable is also a random variable. For example, given function $f : \mathcal{X} \to \mathbb{R}$, we may write both $f(x)$ and $f(X)$, meaning by the latter $f \circ X(\omega)$ for some event $\omega \in \Omega$.

4.2.1 *Spurious correlations from missing sources*

The spurious correlation (SC; Arjovsky et al., 2019), or shortcut-learning (Valle-Perez et al., 2018; Geirhos et al., 2020), problem is characterised by the presence of some secondary attribute s (such as background (Beery et al., 2018), texture (Geirhos et al., 2018), or gender (Sagawa et al., 2019; Seyyed-Kalantari et al., 2020) that confounds the prediction task. We refer to this attribute as the "subgroup", in line with algorithmic fairness (AF; Barocas et al. (2019)) that is strongly correlated with the target attribute, y, in the training set, but spuriously so in the sense that the correlation the mapping $\mathcal{S} \to \mathcal{Y}$ is acausal and thus cannot be expected to hold at deployment time. This correlation is pernicious when S is of lower complexity (which can be formalised in the Kolmogorov sense; Scimeca et al. (2021)) than the causal cues contained in X, and thereby becomes the preferred cue by virtue of simplicity bias (Valle-Perez et al., 2018). Such problems have garnered considerable attention in recent years (Sohoni et al., 2020; Krueger et al., 2021; Liu

et al., 2021; Pezeshki et al., 2021) due to their pervasive, and potentially catastrophic (Codevilla et al., 2019; De Haan et al., 2019; Castro et al., 2020), nature. In this paper, we introduce, and propose a semi-supervised solution for, a hierarchical and class-asymmetric variant of the SC problem that we term the *missing source (MS)* problem .

To illustrate the general SC problem and the MS problem as a particular instantiation of it, we define the conditional-probability matrix, $\mathbf{P}^{tr} \in [0,1]^{|\mathcal{S}|\times|\mathcal{Y}|}$, where each element $\mathbf{P}^{tr}_{ij}$ encodes the conditional probability $P^{tr}(Y=j|S=i)$ in the training set, $\mathcal{D}^{tr}$. When $\mathbf{P}^{tr}$ is both binary and doubly stochastic (that is, has all rows and columns summing to 1) we have that y is completely determined by s in $\mathcal{D}^{tr}$ – this is an extreme form of the SC problem which is statistically intractable without access to additional sources of data (Kehrenberg et al., 2020) or multiple environments (Arjovsky et al., 2019). The MS problem can be viewed as a relaxation of this SC wherein the elements of $\mathbf{P}^{tr}$ respect the constraint that all columns contain at least one non-zero value, i.e. we observe all class labels but not all possible pairs of class and subgroup labels – we say that we have *missing sources*, $\mathcal{M} \triangleq \mathcal{G}^{te} \setminus \mathcal{G}^{tr}$. This setup still leads to spurious correlations but ones that are statistically tractable due to asymmetry. Practically speaking, considering only cases where sources are entirely missing is overly restrictive, and as such we instead view the problem setup as extending to cases where sources may not be altogether missing but have sample sizes too small to constitute meaningful supervision. To understand the non-triviality of this problem, and why aiming for invariance to s in the training set alone – as is characteristic of many representation-learning methods in AF (Edwards and Storkey, 2015; Madras et al., 2018; Quadrianto et al., 2019) and domain adaptation (DA; (Ganin et al., 2016; Saito et al., 2018; Zhao et al., 2018; Lee et al., 2019a)) – will assuredly fail, consider a binary classification problem with binary subgroups, where $\mathcal{Y}=\{0,1\}$ and $\mathcal{S}=\{0,1\}$ and for which $\mathbf{P}^{tr}$ takes the form

$$\mathbf{P}^{tr} = \begin{array}{c} \\ S=0 \\ S=1 \end{array} \begin{array}{c} \begin{array}{cc} Y=0 & Y=1 \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \\ 1.0 & 0.0 \end{pmatrix} \end{array}. \tag{4.1}$$

This represents a special case of the MS problem that we refer to as the *Subgroup Bias* (SB) problem, distinguished by the fact that we observe all subgroups. Here, we have samples from $S=0$ evenly distributed across both the negative and positive classes; for $Y=1$, however, we only observe samples from the negative class. This setup might appear somewhat benign at first blush, given that all classes are present in the training set, however, the fact that s serves as a proxy for y in the case of $S=1$ frustrates our goal of subgroup-invariant classification. The reason for this becomes obvious when decompose a classifier into a mixture of experts (MoE), where s indicates which expert to choose for the given sample. Such a model naturally arises in practice due to the tendency of deep neural networks to strongly favour shortcut solutions (Geirhos et al., 2020). We note that for this, and throughout the paper, we assume that s is inferable, to some extent, from x, that is $\mathcal{I}(X;S)>0$, with $\mathcal{I}(\cdot;\cdot)$ denoting the mutual information between two variables – this is almost always the case in practice but we make the dependence explicit here by denoting by X_Y the causally-relevant component of X), that is independent of S, and by including $s \in \{0,1\}$

explicitly in the set of inputs. With this noted, we may then define the MoE classifier, c_{MoE}, that 'solves' the training set with labels distributed according to $\mathbf{P}^{tr}$ as

$$c_{MoE}(X_Y, S) = \begin{cases} c_{S=0}(X_Y) & S = 0 \\ 0 & S = 1 \end{cases}, \tag{4.2}$$

using $c_{S=0}(\cdot)$ to denote the expert that learns to classify only the subset of the data for which $S = 0$. Such a classifier is clearly undesirable, as should it ever encounter a sample belonging to subgroup 1 with a positive label, the classifier will automatically declare it negative without needing to attend to X_Y – it is invariant to X_Y while being variant to S, which is the opposite of what we desire. This is often done by learning an encoder $f : \mathcal{X} \to \mathcal{Z}$ that maps an input x into a representation, $z \in \mathcal{Z} \subset \mathbb{R}^l$, which has the desired property of S-invariance, $Z \perp S$, while also maximising $\mathcal{I}(Z; Y)$ so that the representation is useful for classification. A popular way of imparting this invariance is with adversarial methods (Ganin et al., 2016; Madras et al., 2018; Zhao et al., 2018) where f is trained to the equilibrium point, f^*, of the (non-convex) minimax equation

$$\min_{f \in \mathcal{F}} \max_{a \in \mathcal{A}} \mathbb{E}_{(x,s,y)\sim\mathcal{D}^{tr}} \Big[\overbrace{a(f(x))_s}^{\text{invariance}} - \underbrace{\lambda \mathcal{I}(f(x); y)}_{\text{classification}} \Big], \tag{4.3}$$

where $a : \mathcal{Z} \to \triangle^{|\mathcal{S}|}$ is a parametric adversary with codomain the standard simplex over $\mathcal{S}$, and $\lambda \in \mathbb{R}^+$ is a positive scalar controlling the trade-off between the two constituent objectives. Under ideal conditions, when all possible pairs of s and y are observed, f^* corresponds to the point at which a is maximally entropic and occurs when Z is invariant to S, and only S, while mutual information w.r.t. Y is jointly maximised – from an optimisation standpoint, the gradients of first and second objectives are non-conflicting (i.e. have non-negative inner products; Yu et al., 2020) and there is no trade-off. However, in cases where we have missing sources, the waters are muddied: satisfying the first part of the objective connotes invariance not only to S, but also to Y, since S can be predicted from Y with above-random accuracy due to the skewed statistics of the dataset. This is patently problematic as Y is the very thing we wish to predict and achieving invariance to S does little good if our classifier can no longer utilise features predictive of Y.

Since we cannot achieve optimality for the competing invariance and classification terms simultaneously, we instead have a set of PO solutions that collectively make up the Pareto front – learning the solutions corresponding to different trade-offs, or preference vectors, is the domain of multi-objective optimisation (MOO; Deb, 2014). Specifically, Eq. 4.3, with λ controlling the preference direction, characterises the most straightforward approach to MOO, called *linear scalarisation* (Boyd et al., 2004). MOO has recently been explored in the context of UDA, for controlling the descent direction, in unsupervised domain adaptation (UDA) in light of the conflict arising between the gradients of the alignment and classification terms (Liang et al., 2021), and in algorithmic fairness (AF) for controlling the inherent trade-off between predictive performance and fairness (typified by the *Accuracy-Fairness trade-off*; Martinez et al. (2020)). While our missing-sources problem admits a MOO-based approach, we are instead interested in leveraging unlabelled data to sidestep the implied trade-off altogether.

4.2.2 *Formalising the problem*

In order to provide a general formulation of the MS problem exemplified above, we begin by defining additional notation for reasoning over label-conditioned subsets and their support. For a given dataset, $\mathcal{D}$, we denote by $\mathcal{D}_{S=s'}$ its subset with subgroup label $s' \in \mathcal{S}$, by $\mathcal{D}_{Y=y'}$ its subset with class label $y' \in \mathcal{Y}$, and – combining the two – by $\mathcal{D}_{S=s',Y=y'}$ its subset with subgroup label s' and class label y'. According to this scheme, $\mathcal{D}_{S=\text{purple},Y=2}$ should then be read as "the set of all samples in $\mathcal{D}$ with class label '2' and subgroup label 'purple"'. We apply similar syntax to the subgroups, writing $\mathcal{S}^{tr}_{Y=y'}$ to mean the observed subgroups within class y in the training set. For instance, $\mathcal{S}^{tr}_{Y=1} = \{0\}$ prescribes that for class 1, only subgroup 0 is present in the training set.

We assume a problem of a hierarchical nature. While the full set of class labels is observed in both the training and test sets, we do not observe all pairs of s and y in the former, i.e. $\mathcal{G}^{tr} \subset \mathcal{G}^{te}$ or $\mathcal{M} \neq \varnothing$. Equivalently, we say that for some class, $y^\dagger$, we have $\mathcal{S}^{tr}_{Y=y^\dagger} \subset \mathcal{S}$, subject to the constraint that $\mathcal{S}^{tr} = \mathcal{S}^{te}$. With this, we can succinctly notate the SB problem realised by Eq. 4.1, in which class $Y = 1$ has no overlap with subgroup $S = 1$, as $\mathcal{S}^{tr}_{Y=1} = \{0\}$ (while $\mathcal{S}^{tr}_{Y=0} = \{0, 1\}$), corresponding to $\mathcal{M} = \{(1, 1)\}$, and distinguish SB problems generally by the inclusion of the additional constraint $\mathcal{S}^{tr} = \mathcal{S}^{te}$. To illustrate a more complex case, the SB problem depicted in Fig. 4.1, in which for we observe exclusively purple '2's and green and red '4's, can be notated with the pair $\mathcal{S}^{tr}_{Y=2} = \{\text{purple}\}$, $\mathcal{S}^{tr}_{Y=4} = \{\text{green}, \text{red}\}$.

4.2.3 *A way forward*

In this paper, we propose to alleviate the SB problem by mixing labelled data with *unlabelled* data that is usually much cheaper to obtain (Chapelle et al., 2006), referring to this set of *unlabelled* data as the *deployment set* [1] $\mathcal{D}^{dep}_{\star} = \{(x_n, s^\star_n, y^\star_n)\}_{n=1}^{N^{dep}} \subset (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})$, using "$\star$" to denote that the labels are *unobserved*, and in practice we only have access to $\mathcal{D}^{dep} \triangleq \{(x_n)\}_{n=1}^{N^{dep}} \subseteq \mathcal{X}$ and must estimate the corresponding sources. We assume that this deployment set is source-complete w.r.t. the test set, $\mathcal{G}^{dep} = \mathcal{G}^{te}$. Leveraging this deployment set, we seek to learn a classifier, Γ, that can generalise well to the missing sources appearing in the test set without seeing any labelled representatives in the training set. In practice, we treat Γ as a composition, $c \circ f$, of two subfunctions: an encoder $f : \mathcal{X} \to \mathcal{Z}$, which maps a given input x to a representation $z \in \mathcal{Z} \subseteq \mathbb{R}^l$, and a classifier head $c : \mathcal{Z} \to \mathcal{Y}$ which completes the mapping to the space of class labels, $\mathcal{Y}$. Since the task of achieving independence between the predictions and subgroup labels can be reduced to the task of learning the invariance $Z \perp S$; we next discuss how one can learn an encoder satisfying this condition in a theoretically-principled manner.

4.3 ADVERSARIAL SUPPORT-MATCHING

We cast the problem of learning a subgroup-invariant representation as one of *support-matching* between a dataset that is *labelled* but has *incomplete* support over sources, G, and one, conversely,

1 In our experiments, we report accuracy and bias metrics on another independent test set instead of on the unlabelled data that is available at training time.

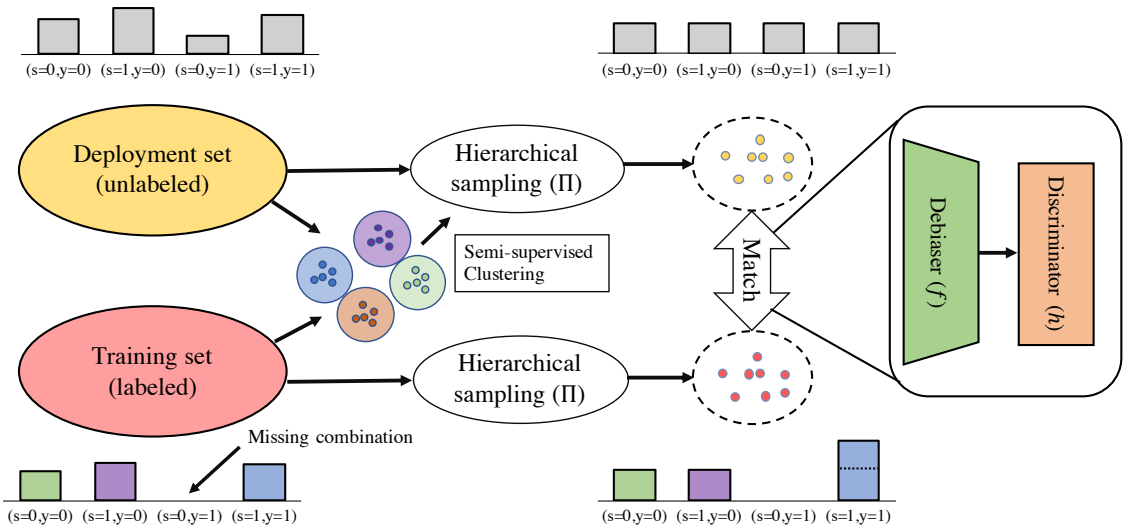


Figure 4.2: Visualisation of our support-matching pipeline. Bags are sampled from the training and deployment sets using the hierarchical sampling procedure described in §4.3 and defined functionally in Eq. 4.17. Since we cannot use ground-truth labels for hierarchical sampling of the deployment set, we use a semi-supervised clustering algorithm to produce balanced batches. In the event that certain combinations are missing, as shown here for $(s = 0, y = 1)$, the sampling on the training set substitutes the missing combinations with combinations that ensure equal representation of the target classes. The debiaser is adversarially trained to produce representations from which the source dataset cannot be reliably inferred by the discriminator. Assuming the bags are sufficiently balanced and $\mathcal{G}^{tr} \subsetneq \mathcal{G} = \mathcal{S} \times \mathcal{Y}$, the optimal debiaser is one that produces a representation z that is invariant to s, which we prove in Appendix 4.7.2.

that has *complete* support over G, but is *unlabelled*. The idea is to produce a representation that is invariant to this difference in support, and thus invariant to the subgroup. However, it is easy to learn the wrong invariance if one is not careful. To measure the discrepancy in support between the two distributions, we adopt an adversarial approach, but one where the adversary is operating on small sets – which we call *bags* – instead of individual samples. These bags need to be balanced with respect to (s, y), such that we can interpret them as approximating $\mathcal{G}$ as opposed to the joint probability distribution, $P(S, Y)$. Details on how these bags are constructed can be found in in §4.3.2 and §4.3.3.

4.3.1 *Objective*

We now present our overall support-matching objective. As alluded to before, the goal, in summary, is to learn an encoder, f, which preserves all information relating to Y, but is invariant to S. Let $P^{tr}(f(X) = z', S = s', Y = y')$ be the joint probability that a data point x drawn from $P^{tr}(X)$ – the training set – results in the encoding z' and is at the same time labelled as subgroup s' and class y'. We also define the following shorthand: $p_f(Z = z') = P^{tr}(f(X) = z')$, the distribution resulting from sampling x from P^{tr} and then transforming x with f. Analogously for the deployment set: $q_f(Z = z') = P^{dep}(f(X) = z')$. For the conditioned distributions we write $p_f|_{S=s',Y=y'}$, following the convention established in §4.2.2 but with the added '|' to clearly delimit the conditioning.

The objective makes a distinction between those classes, $y \in \mathcal{Y}$, for which there is overlap with all subgroups $s \in \mathcal{S}$ in the training set and those classes for which there is not. To formalise this, we define the following helper function Π which maps (s', y') to a set of subgroup identifiers depending on whether the class y has full s-support:

$$\Pi(s', y') = \begin{cases} \{s'\} & \text{if } \mathcal{S}^{tr}_{Y=y'} = \mathcal{S} \\ \mathcal{S}^{tr}_{Y=y'} & \text{otherwise.} \end{cases} \tag{4.4}$$

$\Pi(s, y)$ ensures that the correct invariance is learned and is discussed in more detail further below. Our objective is then

$$\mathcal{L}_{\text{match}}(f) = \sum_{s' \in \mathcal{S}} \sum_{y' \in \mathcal{Y}} d(p_f|_{s \in \Pi(s', y'), Y=y'}, q_f|_{S=s', Y=y'}) \tag{4.5}$$

where $d(\cdot, \cdot)$ is a distance measure for probability distributions. The optimal encoder f^* is found by solving the following optimisation problem:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathcal{L}_{\text{match}}(f) - \mathcal{I}(f(X); X) \tag{4.6}$$

where $\mathcal{I}(\cdot; \cdot)$ again denotes the mutual information. As written, Eq. 4.5 requires knowledge of s and y on the deployment set for conditioning. That is why, in practice, the distribution matching is not done separately for all combinations of $s' \in \mathcal{S}$ and $y' \in \mathcal{Y}$. Instead, we compare *bags* that contain samples from all combinations in the right proportions. For the deployment set, Eq. 4.5 implies that all s-y-combinations have to be present at the same rate in the bags, but for the training set, we need to implement $\Pi(s', y')$ with hierarchical balancing.

As the implications of the given objective might not be immediately clear, we provide the following proposition. The proof can be found in Appendix 4.7.

THEOREM 4.1. *If f is such that*

$$p_f|_{s \in \Pi(s', y'), Y=y'} = q_f|_{S=s', Y=y'} \quad \forall s' \in \mathcal{S}, y' \in \mathcal{Y} \tag{4.7}$$

and P^{tr} and P^{dep} are data distributions that correspond to the real data distribution P, except that some s-y-combinations are less prevalent, or, in the case of P^{tr}, missing entirely, then, for every $y' \in \mathcal{Y}$, there is either full coverage of s for y' in the training set ($\mathcal{S}^{tr}_{Y=y'} = \mathcal{S}$), or the following holds:

$$P(S = s'|f(X) = z', Y = y') = \frac{1}{n_s}\,. \tag{4.8}$$

In other words: for $Y = y'$, $f(x)$ is not predictive of s.

4.3.2 *Implementation*

The implementation of above objective combines elements from unsupervised representation-learning and adversarial learning. In addition to the invariant representation z, our model also

outputs $\tilde{s}$, in a similar fashion to Kehrenberg et al. (2020) and Creager et al. (2019). This can be understood as a reconstruction of the subgroup information from the input x and is necessary to prevent z from being forced to encode s by the reconstruction loss. We note that this need could potentially be obviated through use of self-supervised approaches, but refrain from exploring this avenue in the interest of simplicity.

The model, Γ, is composed of three core modules: 1) two *encoder* functions, f (which we refer to as the "debiaser") and t, which share weights and map x to $z \in \mathcal{Z}$ and $\tilde{s} \in \tilde{\mathcal{S}}$, respectively; 2) a *decoder* function $r : \mathcal{Z} \times \tilde{\mathcal{S}} \to \mathcal{X}$ that learns to invert f and t; and 3) a *discriminator* function $h : \mathcal{Z} \to (0, 1)$ that predicts which dataset a bag of samples embedded in $\mathcal{Z}$ was sampled from. The encoder f is then tasked with learning a representation z such that it is indeterminable to the adversary h whether a given bag originated from the deployment set ('positive') or the training set ('negative'). Formally, given bags $\mathcal{B}^{tr}$, sampled according to Π from the training set, and balanced bags from the deployment set, $\mathcal{B}^{dep}$, we first define, for notational convenience, the loss w.r.t. to the encoder networks, f and t as

$$\mathcal{L}_{\text{enc}}(f, t, r, h) = \sum_{b^{dep} \in \mathcal{B}^{dep}} \sum_{b^{tr} \in \mathcal{B}^{tr}} \Bigg[\overbrace{\sum_{x \in b^{dep} \cup b^{tr}} \|x - r(f(x), t(x))\|_p^p}^{\mathcal{L}_{\text{recon}}} \Bigg] \underbrace{\lambda_{\text{match}} \Big[\log h\big(\{\mathtt{sg}[f(x)] \mid x \in b^{dep}\}\big) - \log h\big(\{f(x) \mid x \in b^{tr}\}\big) \Big]}_{\mathcal{L}_{\text{match}}}, \tag{4.9}$$

where $\mathcal{L}_{\text{recon}}$ denotes the reconstruction loss defined by the p-norm ($p = 1$ and $p = 2$ yielding MAE and MSE, respectively), $\mathcal{L}_{\text{match}}$ denotes the adversarial loss, $\lambda_{\text{match}} \in \mathbb{R}_*^+$ is a pre-factor controlling the trade-off between the loss terms, and $\mathtt{sg}[\cdot]$ denotes the "stop-gradient" operator that behaves as the identity function but with zero partial derivatives. The overall objective, encompassing f, t, and h can then be formulated in terms of $\mathcal{L}_{\text{enc}}$ as

$$\min_{f,t,r} \max_{h} \mathcal{L}_{\text{enc}}(f, t, h) \,. \tag{4.10}$$

This equation is computed over batches of bags and the discriminator is trained to map a bag of samples from the training set and the deployment set to a binary label: 1 if the bag is adjudged to have been sampled from the deployment set, 0 if from the training set. For the discriminator to be able to classify sets of samples, it needs to be permutation-invariant along the bag dimension – that is, its predictions should take into account dependencies between samples in a bag while being invariant to the order in which they appear. For aggregating information over the bags, we use a self-attention-based (Vaswani et al., 2017) pooling layer in which the query vector is averaged over the bag dimension. For more details see Appendix 4.9.2. Furthermore, in Appendix 4.10.1, we validate that having the discriminator operate over sets (bags) of samples rather than independent samples (with the same balancing scheme) is essential for achieving good and robust (w.r.t. balancing quality) performance.

Our goal is to disentangle x into two subspaces: a subspace z, representing the class, and a subspaces $\tilde{s}$, representing the subgroup. For the problem to be well-posed, it is crucial that the bags differ only in terms of which sources are present and not in terms of other aspects. We thus

sample the bags according to the following set of rules which operationalize Π. Please refer to Fig. 4.2 for a visualisation of the effect of these rules.

1. Bags of the deployment set are sampled so as to be approximately balanced with respect to s and y (all combinations of s and y should appear in equal number).

2. For bags from the training set, all possible values of y should appear with equal frequency. Without this constraint, there is the risk of y being encoded in $\tilde{s}$ instead of s.

3. Bags of the training set should furthermore exhibit equal representation of each subgroup within classes so long as rule 2 is not violated. For classes that do not have complete s-support, the missing combinations of (s, y) need to be substituted with a sample from the same class – i.e., if $s \notin \mathcal{S}^{tr}(y)$ we instead sample randomly from a uniform distribution over $\mathcal{S}^{tr}(y)$.

Algorithm 1 Adversarial Support Matching

INPUT: Number of encoder updates N^{enc}, number of discriminator updates N^{disc}, encoders f and t, decoder r, discriminator h, training set $\mathcal{D}^{tr}$, deployment set $\mathcal{D}^{dep}$
OUTPUT: Debiaser f with learned invariance to s

for $i \leftarrow 1$ to N^{enc} **do** ▷ Encoder update loop
  Sample batches of perfect bags $\mathcal{B}^{tr} \sim \mathcal{D}^{tr}$ and $\mathcal{B}^{dep} \sim \mathcal{D}^{dep}$ using Π (Eq. 4.4)
  Compute $\mathcal{L}^{\text{enc}}$ using Eq. 4.10
  Update f, t, and r by descending in the direction $\nabla\mathcal{L}^{\text{enc}}$
  for $j \leftarrow 1$ to N^{disc} **do** ▷ Discriminator update loop
    Sample batches of perfect bags $\mathcal{B}^{tr} \sim \mathcal{D}^{tr}$ and $\mathcal{B}^{dep} \sim \mathcal{D}^{dep}$ using Π (Eq. 4.4)
    Compute $\mathcal{L}^{\text{match}}$ using Eq. 4.10
    Update h by ascending in the direction $\nabla\mathcal{L}^{\text{match}}$
  end for
end for

4.3.3 *Perfect bags*

A visual overview of our pipeline is given in Fig. 4.2. Borrowing from the AF literature (Kleinberg et al., 2016; Chouldechova, 2017), we refer to a bag in which all elements of $\mathcal{G}$ appear in equal proportions as a "perfect bag" (even if the balancing is only approximate). Our pipeline can be broken down into two steps: 1) sample perfect bags from an unlabelled deployment set; and 2) produce disentangled representations using the perfect bags via adversarial support-matching as described in §4.3.2.

Constructing perfect bags via clustering. We cluster the data points from the deployment set into $N^C = |\mathcal{G}|$ clusters by applying spherical k-means to CLIP (Radford et al., 2021) embeddings. Specifically, we use the ResNet-50 version of CLIP, finding this to work better than the ViT-based variants. We inject labelled knowledge into the k-means algorithm by initialising the centroids of the known sources the mean of their features in the labelled (training) data. We find this works reasonably well for the considered datasets; since, the aim of this work is to propose a pipeline for

effectively leverage unlabelled data for invariance-learning, not to set a new state-of-the-art in clustering, we adopt this simple clustering method for a practical proof-of-concept compared with the artificial approach of injecting noise into the ground-truth labels. The latter procedure is useful, however, for performing a fine-grained sensitivity analysis of our algorithm w.r.t. clustering accuracy, in that we can simulate the runs of the algorithm at different levels of noisiness in the bag-sampling. Given, the cluster assignments, we can then stratify the deployment set into perfect bags, to be used by the subsequent support-matching phase.

As a result of clustering, the data points in the deployment set $\mathcal{D}^{dep}$ are labelled with cluster assignments generated by clustering algorithm, C, giving $\mathcal{D}_C^{dep} = \{(x_i, c_i)\}$, $c_i = C(z_i)$, so that we can form perfect bags from $\mathcal{D}_C^{dep}$ by sampling all clusters at equal rates; there is no need for application of the Π operator since the deployment set is complete w.r.t. $\mathcal{G}$. We note that we do *not* have to associate the clusters with specific s or y labels as the labels are not directly used for supervision.

Balancing bags based on clusters instead of the true labels introduces an error, which we can try to bound. For this error-bounding, we assume that the probability distribution distance measure used in Eq. 4.5 is the *total variation distance* TV. The proof can be found in Appendix 4.7.

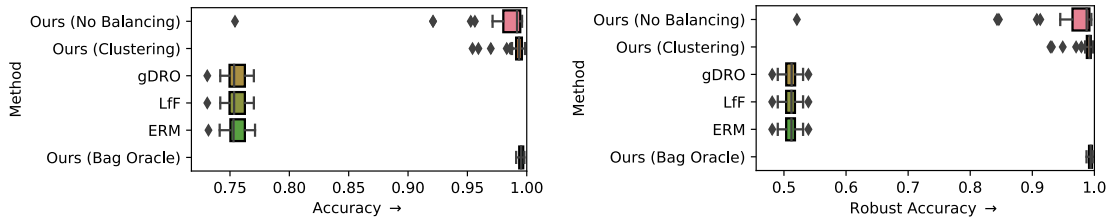
THEOREM 4.2. *If $q_f(Z)$ is a data distribution on $\mathcal{Z}$ that is a mixture of $n_y \cdot n_s$ Gaussians, which correspond to all the unique combinations of $y \in \mathcal{Y}$ and $s \in \mathcal{S}$, and $p_f(Z)$ is any data distribution on $\mathcal{Z}$, then without knowing y and s on q_f, it is possible to estimate*

$$\sum_{s' \in \mathcal{S}} \sum_{y' \in \mathcal{Y}} TV(p_f|_{s \in \Pi(s',y'), Y=y'}, q_f|_{S=s', Y=y'}) \tag{4.11}$$

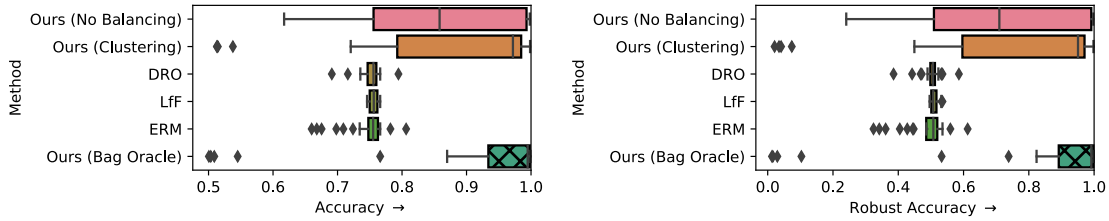
with an error that is bounded by $\tilde{O}(\sqrt{1/N})$ with high probability, where N is the number of samples drawn from q_f for learning.

4.3.4 *Limitation and intended use*

Although having zero labelled examples for some subgroups is not uncommon due to the effects of systematic bias or dataset curation, we should make a value-judgement on the efficacy of the dataset with respect to a task. We can then decide whether or not to take corrective action as described in this paper. A limitation of the presented approach is that, for constructing the perfect bags used to train the disentangling algorithm, we have relied on knowing the number of clusters *a priori*, something that, in practice, is perhaps not the case. However, for person-related data, such information can, for example, be gleaned from recent census data. (see also Appendix 4.10.2 for results with misspecified numbers of clusters.) One difficulty with automatic determination of the number of clusters is the need to ensure that the small clusters are correctly identified. A cluster formed by an underrepresented subgroup can be easily overlooked by a clustering algorithm in favour of larger but less meaningful clusters.



(a) Results for the *subgroup-bias* scenario where purple fours constitute the missing source. The clustering accuracy for `Ours (No Balancing)` was 96% ± 6%. Our method consistently outperforms the baselines, which fare no better than random on the subgroup with the missing source. As we would expect, the median and IQR of our method are positively- and negative- correlated, respectively, with how well the bags of the deployment set are balanced, with `Ours (Bag Oracle)` providing an upper bound for this. Indeed, in one case `Ours (No Clustering)` failed to surpass the baselines, but through use of clustering the `Robust Accuracy` is kept within the region of 95% at worst.



(b) Results for the *missing-subgroup scenario* where purple digits constitute the missing subgroup. The clustering accuracy for `Our (No Balancing)` was 88% ± 5%. This scenario is significantly more difficult to solve than the subgroup-bias as there is insufficient inductive bias in the labels and the deployment set for the support matching to be well-posed. This is reflected in the high variance of our method, variance, however, which can be drastically reduced by improving the quality of balancing. Nevertheless, all variants of our method perform significantly better than the baselines in terms of the median `Robust Accuracy`, and the rate at which they produce degenerate solutions (marked by performance worse than `ERM`'s) relatively low.

Figure 4.3: Results for two-digit Colored MNIST for two different scenarios (subgroup bias (Top) and missing subgroup (Bottom)) in the form of box plots of the `Robust Accuracy` (the minimum accuracy computed over the subgroups) over **30 repeats**.

4.4 Experiments

We perform experiments on a combination of publicly-available image datasets – Coloured MNIST (following a similar data-generation process to Kehrenberg et al. (2020)) and CelebA (Liu et al., 2015). We report the `Robust Accuracy` – the minimum accuracy over the subgroups – for all datasets. For this dataset, we instead report `Robust TPR` – analogously, the minimum TPR over the subgroups – as the primary metric given the emphasis on positive classifications in medical contexts. Additional plots showing the Accuracy, Positive Rate, TPR, and TNR ratios can be found in Appendix 4.9.5.

We compare the performance of our disentangling model when paired with each of three different bag balancing methods: 1) with clustering via rank statistics (`Clustering`); 2) without balancing, when the deployment set $\mathcal{D}^{dep}$ is used as is (`No Balancing`); 3) with balancing using the ground-truth class and subgroup labels (`Oracle Bag`) that would in practice be unobservable; this provides insight into the performance under ideal conditions and how sensitive the method is to bag imbalance.

4.4.1 *Coloured MNIST*

Appendix 4.8 provides description of the dataset and the settings used for D^{dep} and D^{tr}. Each source is then a combination of digit-class (class label) and colour (subgroup label). We begin by considering a binary, 2-digit/2-colour, variant of the dataset with $\mathcal{Y} = \{2, 4\}$ and $\mathcal{S} = \{\text{green}, \text{purple}\}$. (Appendix 4.6.1 provides results for 3-digit/3-colour variant.) For this variant we explore both the SB (subgroup bias) setting and a more extreme *missing subgroup* setting. To simulate the SB setting, we set $\mathcal{S}^{tr}_{Y=4} = \{\text{green}\}$. In the *missing subgroup* setting, $S = \text{purple}$ is missing from $\mathcal{S}^{tr}_{Y=2}$ as well, so that all classes only have support in $\{\text{green}\}$. However, for this scenario, the disentangling procedure has more than one possible solution – apart from the natural solution, it is also possible to consider ($Y = 2$, $S = \text{green}$) and ($Y = 4$, $S = \text{purple}$) as forming one factor in the disentangling, with the other factor comprising the two remaining s-y-combinations. Such an "unnatural" disentangling (spanning digit class *and* colour) is avoided only by the tendency of neural networks to prefer simpler solutions (Occam's razor) and in general we cannot guarantee that this pathological case be avoided based only on the information provided by the training labels and deployment set.

To establish the effectiveness of our method, we compare against four baselines. The first is `ERM`, a classifier trained with cross-entropy loss on this data; the second is `DRO` (Hashimoto et al., 2018), which functions without subgroup labels by minimising the worst-case training loss over all possible groups that are above a certain minimum size; the third is `gDRO` (Sagawa et al., 2019), which minimises the worst-case training loss over predefined subgroups but is only applicable when $|\mathcal{S}^{tr}| > 1$; the fourth is `LfF` (Nam et al., 2020) which reweights the cross-entropy loss using the predictions of a purposely-biased sister network. For fair comparison, the training set is balanced according to the rules defined in §4.3 for all baselines.

Fig. 4.3a shows the results for the SB setting. We see that the performance of our method directly correlates with how balanced the bags are, with the ranking of the different balancing methods being `Oracle` > `Clustering`> `No Balancing`. Even without balancing, our method greatly outperforms the baselines, which all perform similarly.

Fig. 4.3b shows that the problem of *missing subgroups* is harder to solve. For all balancing strategies, the IQR is significantly higher than observed in the SB setting, with the latter also giving rise to a large number of extreme outliers. The median, however, remains high, indicative of a "hit-or-miss" aspect to the method, albeit with the number of hits far outweighing the misses. Visualisations of the reconstructions (Appendix 4.9.4) suggest that the extreme outliers correspond to the degenerate solution mentioned above.

4.4.2 *CelebA*

To demonstrate our method generalises to real-world computer vision problems, we consider the CelebA dataset (Liu et al., 2015) comprising over 200,000 images of different celebrities. The dataset comes with per-image annotations of physical and affective attributes such as 'Smiling', 'Gender', hair colour, and 'Age'. Since the dataset exhibits natural imbalance with respect to $\mathcal{G}$, we perform no additional sub-sampling of either the training set or the deployment set. We

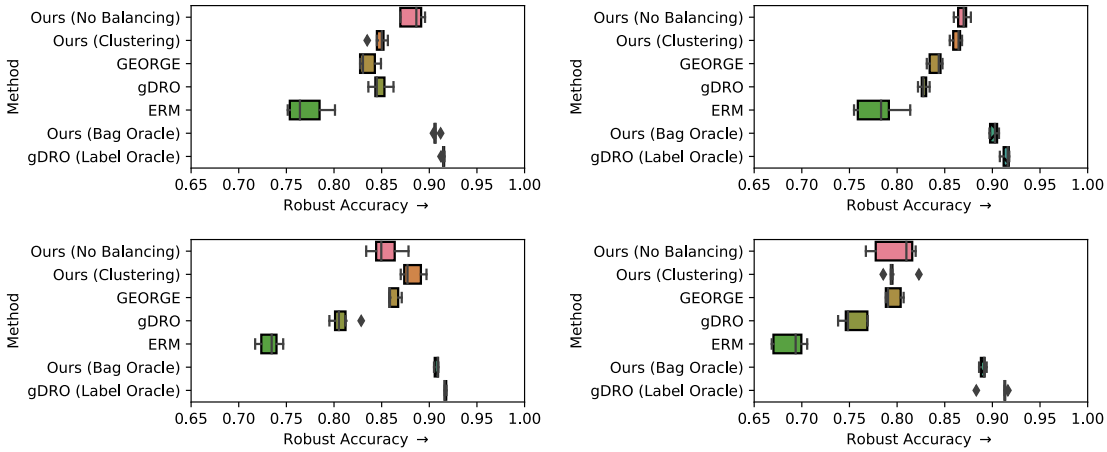


Figure 4.4: Results from **5 repeats** for the CelebA dataset for the *subgroup bias* scenario. The task is to predict "smiling" vs "non-smiling" and the subgroups are based on gender. The four sources are dropped one at a time from the training set (**Top Left**: smiling females; **Top Right**: smiling males; **Bottom Left**: non-smiling females; **Bottom Right**: non-smiling males), while the deployment set is kept fixed. `Robust Accuracy` refers to the minimum accuracy computed over the subgroups. Our method consistently performs on par with or outperforms `GEORGE` (which in turn outperforms `ERM`). We note that in some of the runs, `GEORGE` performed no better than random – these results were truncated for visibility but can be found in Fig. 4.6. Given *indirect* supervision from the deployment set in the form of oracle-balancing, our method performs similarly to `gDro (Label Oracle)` that receives *direct* supervision.

predict "smiling" as the class label and use the binary attribute, "gender", as the subgroup label. Here, we consider the SB setting but rather than just designating one missing source, we repeat our experiments with each source being dropped in turn. As before, we evaluate our method under three balancing schemes and compare with ERM and gDRO trained on only the labelled training data. We also compare with two other variants of gDRO: 1) `gDRO (Label Oracle)`, a variant that is trained with access to the ground-truth labels of the deployment set, thus providing an upper-bound on the downstream classification performance; 2) `GEORGE` (Sohoni et al., 2020), which follows a two-step procedure of first clustering to obtain the labels for hidden subgroups, and then using these labels to train a robust classifier using gDRO. Sohoni et al. (2020) consider a different version of the problem (termed *hidden-stratification*) in which the class labels are known for all samples but the subgroup-labels are missing entirely. We adapt `GEORGE` to our setting by modifying the semi-supervised clustering algorithm to predict the marginal distributions ($P(Y|X)$ and $P(S|X)$ instead of the joint distribution $P(Y, S|X)$, allowing us to propagate the class labels from D^{tr} to D^{dep} (see Appendix 4.11 for details).

Fig. 4.4 shows the results for experiments for each missing source, showing similar trends across all instantiations of the SB scenario. `gDRO (Label Oracle)` consistently achieves the best performance according to both metrics, with `Our Method (Bag Oracle)` consistently coming in second. We note that while both methods use some kind of oracle, the *label oracle* provides *all* class/subgroup labels to its algorithm, whereas the *bag oracle* only balances the bags. Despite the large difference in the level of supervision, the margin between the two oracle methods is slim. We observe that clustering in many cases impairs performance which can be explained by poor clustering of the missing source (∼60% accuracy). CelebA exhibits a natural imbalance with respect to gender/smiling but not a significant one, allowing for random sampling to yield a reasonable approximation to the desired perfect bags. We believe adjustments to the clustering

algorithm – e.g. using a self-supervised loss instead of a reconstruction loss for the encoder – could close the gap between clustering-based and oracle-based balancing. Nonetheless, among the non-oracle methods, variants of our method consistently match or exceed the performance of the baselines. While the plots show GEORGE can perform strongly in this SB scenario, we note that for several of the missing sources, the method failed catastrophically in one out of the five runs. We have cut off those data points here so as not to compromise the visibility of the other results; the full versions of the plots can be found in Appendix 4.6.2. The fact that GEORGE leverages both the training and deployment sets in a semi-supervised way with clustering makes it the baseline most comparable to our method. However, its performance is much more dependent on the clustering step than our method.

4.5 Related work

Invariant learning. Sohoni et al. (2020) and Creager et al. (2021) both consider a similar problem, where the data also exhibits a two-level hierarchy formed by classes and subgroups. In contrast to our work, however, there is no additional bias in the data; while they may be unobserved, the labelled data is assumed to have complete class-conditional support over the subgroups. As such, these methods are not directly applicable to the particular form of the problem we consider. Like us, Sohoni et al. (2020) uses semi-supervised clustering to uncover the hidden subgroups, however their particular clustering method requires access to the class labels not afforded by the deployment set, as does the training of the robust classifier.

Unsupervised domain adaptation. In unsupervised domain adaptation (UDA), there are typically one or more source domains, for which training labels are available, and one or more unlabelled target domains to which we hope to generalise the classifier. A popular approach for solving this problem is to learn a representation that is invariant to the domain using adversarial networks (Ganin et al., 2016) or non-parametric discrepancy measures such as MMD (Gretton et al., 2012).

There are two ways in which one can compare UDA to our setting: 1) by treating the subgroups as domains; and 2) by treating the training and the deployment set as "source" and "target" domains, respectively. The first comparison is exploited in algorithm fairness, yet does not carry over to our setting in which the labelled data contains *incomplete* domains. When all sources from a given domain are missing then there are no domains to be matched, and even when this is not the case, matching will result in misalignment due to differences in class-conditional support. The second comparison is more germane but ignores an important aspect of our problem: the presence of SCs.

Similar to us, Tong et al. (2022) utilise adversarial methods to align the support of two distributions in a semi-supervised regime – specifically, they propose to use symmetric support difference as a divergence measure which they realise using a discriminator. However, their method focuses on label-shift in the UDA setting and does not consider the hierarchical structure that exists within the source (training) and target (deployment) domains, and as such they do not consider the notion of "missing sources" that can arise due to said structure – the characterisation

of this problem is one of the two main contributions of this work (the other being our proposed solution). Furthermore, the discriminator used therein is applied instance-wise; we show in 4.10 that allowing the discriminator to model inter-sample relationships has tangible addition benefits when performing support-matching.

MULTIPLE INSTANCE LEARNING. Multiple instance learning (Maron and Lozano-Pérez, 1998) is a form of weakly-supervised learning in which samples are not labelled individually part as part of a set or *bag* of samples. In the simplest (binary) case, a bag is labelled as positive if there is a single instance of a positive class contained within it, and negative otherwise. In our case, we can view the missing sources as constituting the positive classes, which leads to all bags (a term we will use throughout the paper distinctly from "batches") from the deployment set being labelled as positive, and all bags from the training set being labelled as negative. Given this labelling scheme, we make use of an adversarial set-classifier to align the supports of the training and deployment sets in the representation space of an encoder network.

POSITIVE UNLABELLED LEARNING. Learning from positive and unlabelled data, or *PU learning*, refers to the binary-classification setup in which the labelled training data consists of only positive samples while additional unlabelled data is assumed to contain both positive and negative samples (Liu et al., 2002, 2003; Bekker and Davis, 2020). This is analogous to our problem setting if we consider the positive class to be all samples sources represented in the training set, collectively, while the missing sources collectively make up the negative class. However, the goal here is not merely to learn the classification boundary between the present and missing sources but to learn to classify the target class of a given sample independently of its subgroup. This is equivalent to requiring that a classifier trained to distinguish between positive and negative classes, according to the aforementioned PU learning setup, from the pre-logits layer of our desired classifier be maximally entropic – we propose to use adversarial learning to achieve this.

4.6 CONCLUSION

We have highlighted the problem that systematic bias or dataset curation can result in one or more subgroups having zero labelled data, and by doing so, hope to have stimulated serious consideration for it when planning, building, and evaluating systems. We proposed a two-step approach for addressing the resulting spurious correlations. First, we construct perfect bags from an unlabelled deployment set via semi-supervised clustering. Second, by matching the support of the training and deployment sets in representation space, we learn representations with subgroup-invariance. We empirically validate our framework on the Coloured MNIST and CelebA datasets, showing it possible to maintain high performance on subgroups with incomplete support. Furthermore, we bound the error in the objective due to imperfect clustering. Future work includes exploring other UL methods and addressing the limitations raised in §4.3.4.

4.6.1 *Results for 3-digit 3-colour variant of Coloured MNIST*

To investigate how our method scales with the number of sources, we look to a 3-digit, 3-colour variant of the dataset in the *subgroup bias* setting where four sources are missing from $\mathcal{D}^{tr}$. Results for this configuration are shown in Fig. 4.5. We see that the performance of `Ours (No Balancing)` is quite close to that of `Ours (Bag Oracle)`. We suspect this is because balancing is less critical with the increased number of subgroups strengthening the training signal. As inter-subgroup ratios do not make for suitable metric for non-binary S, we instead quantify the invariance of the predictions to the subgroup with the Hirschfeld-Gebelein-Rényi maximal correlation (HGRMC) (Rényi, 1959).

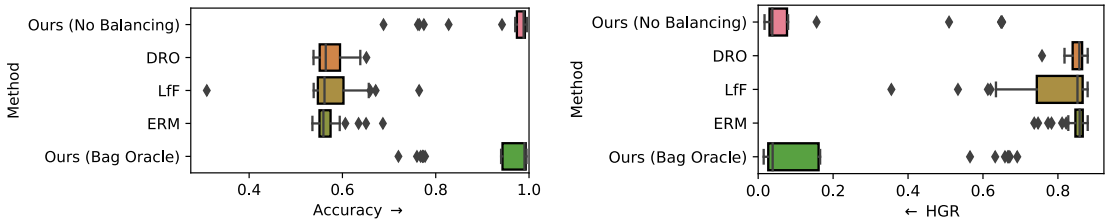


Figure 4.5: Results from **30 repeats** for the Coloured MNIST dataset with three digits: '2', '4' and '6'. Four combinations of digit and colour are missing: green 2's, blue 2's, blue 4's and green 6's. **Left**: Accuracy. **Right**: Hirschfeld-Gebelein-Rényi maximal correlation Rényi, 1959 between S and Y.

4.6.2 *Extended Results for CelebA*

As alluded to in main text, for three out of four of the missing gender/smiling quadrants, the `GEORGE` baseline produced an extreme outlier for one out of the five total repeats - these outliers were omitted from the plots to ensure the discriminability of the other results. We reproduce the full, untruncated versions of these plots here in Fig. 4.6. We have also included `Accuracy` metric in Fig. 4.6.

4.7 THEORETICAL ANALYSIS

In this section, we present our theoretical results concerning the validity of our support-matching objective and the bound on the error introduced into it by clustering. We use notation consistent with that used throughout the main text.

4.7.1 *Sampling function for the objective*

The stated objective uses the following helper function:

$$\Pi(s', y') = \begin{cases} \{s'\} & \text{if } \mathcal{S}^{tr}_{Y=y'} = \mathcal{S} \\ \mathcal{S}^{tr}_{Y=y'} & \text{otherwise .} \end{cases} \tag{4.12}$$

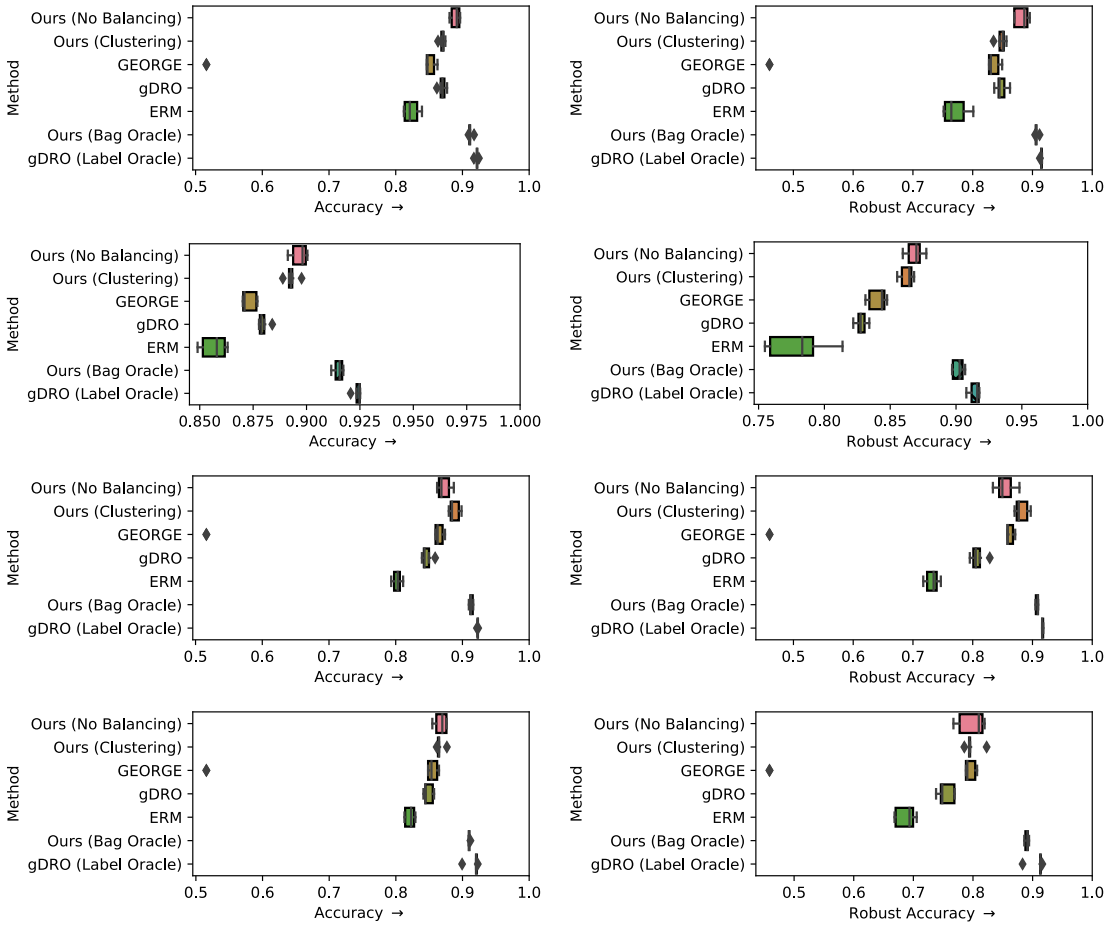


Figure 4.6: Results from **5 repeats** for the CelebA dataset for the *subgroup bias* scenario. The task is to predict "smiling" vs "non-smiling" and the subgroups are based on gender. The four sources are dropped one at a time from the training set (**first row**: smiling females; **second row**: smiling males; **third row**: non-smiling females, **fourth row**: non-smiling males), while the deployment set is kept fixed. "Robust Accuracy" refers to the minimum accuracy computed over the subgroups.

This helper function determines which s value in the training set an s-y pair from the deployment set is mapped to. (The y value always stays *the same* when mapping from deployment set to training set.) To demonstrate the usage of this function, we consider the example of binary Coloured MNIST with $\mathcal{S} = \{\text{purple}, \text{green}\}$ and $\mathcal{Y} = \{2, 4\}$ where the training set is missing $(s = \text{purple}, y = 4)$. In this case, Π takes on the following values:

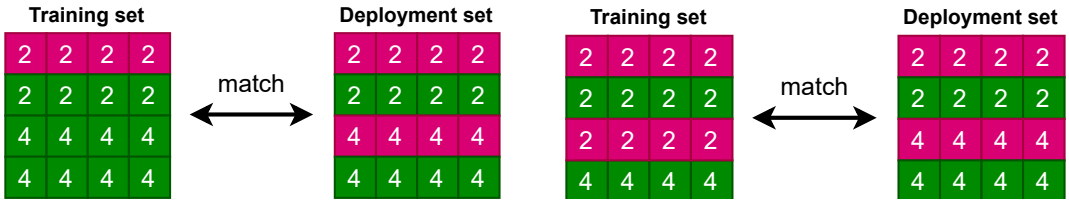
$$\Pi(\text{purple}, 2) = \{\text{purple}\} \tag{4.13}$$
$$\Pi(\text{green}, 2) = \{\text{green}\} \tag{4.14}$$
$$\Pi(\text{purple}, 4) = \mathcal{S}^{tr}_{y=4} = \{\text{green}\} \tag{4.15}$$
$$\Pi(\text{green}, 4) = \mathcal{S}^{tr}_{y=4} = \{\text{green}\} \tag{4.16}$$

It is essential that $(s = \text{purple}, y = 4)$ from the deployment set is mapped to $(s = \text{green}, y = 4)$ from the training set, and not $(s = \text{purple}, y = 2)$. This procedure is illustrated in Fig. 4.7a, and contrasted with an incorrect procedure based on balancing the bag according to s in 4.7b – such a procedure would result in invariance to y instead of s, which is obviously undesirable.

In practice, we use the following sampling function π to implement Π, sampling from it for all $(s, y) \in S \times Y$:

$$\pi(s', y') = \begin{cases} x \sim P^{tr}(x|S = s', y'), \\ \qquad\qquad \text{if } \mathcal{S}^{tr}_{Y=y'} = \mathcal{S} \\ x \sim P^{tr}(x|s = \check{s}, y'), \check{s} \sim \text{uniform}(S^{tr}), \\ \qquad\qquad \text{otherwise .} \end{cases} \tag{4.17}$$

With the assumption that our data follows a two-level hierarchy and all digits appear in the training set, the above sampling function π traverses the first level which corresponds to the class-level information, and *samples* the second level which corresponds to subgroup-level information when we have missing sources.



(a) Correct matching procedure. **(b)** *In*correct matching procedure.

Figure 4.7: The two natural matching procedures for one missing source in the training set. Only figure 4.7a (left) produces the desired invariance.

4.7.2 *Implication of the objective*

We restate proposition 1 and present the proof.

We prove here that an encoding f satisfying the objective is invariant to s, at least in those cases where the class does not have full s-support (which is exactly the case where it matters).

THEOREM 4.3. *If f is such that*

$$p_f|_{s\in\Pi(s',y'),Y=y'} = q_f|_{S=s',Y=y'} \quad \forall s' \in \mathcal{S}, y' \in \mathcal{Y} \tag{4.18}$$

and P^{tr} and P^{dep} are data distributions that correspond to the real data distribution P, except that some s-y-combinations are less prevalent, or, in the case of P^{tr}, missing entirely, then, for every $y' \in \mathcal{Y}$, there is either full coverage of s for y' in the training set ($\mathcal{S}^{tr}_{Y=y'} = \mathcal{S}$), or the following holds:

$$P(S = s'|f(X) = z', Y = y') = \frac{1}{n_s} \,. \tag{4.19}$$

In other words: for $Y = y'$, $f(x)$ is not predictive of s.

Proof. If y' has full coverage of s in the training set, there is nothing to prove. So, assume y' does not have full s-support. That means $\Pi(s', y') = \mathcal{S}^{tr}_{Y=y'}$ for all $s' \in \mathcal{S}$. And so

$$\begin{aligned} &P^{tr}(f(X) = z'|s \in \mathcal{S}^{tr}_{Y=y'}, Y = y') \\ &= P^{dep}(f(X) = z'|S = s', Y = y') \qquad \forall s' \in \mathcal{S} \end{aligned} \tag{4.20}$$

The left-hand side of this equation does not depend on s' and so the right-hand side must have the same value for all $s' \in \mathcal{S}$, which implies:

$$\begin{aligned} &P^{dep}(f(X) = z'|S = s', Y = y') \\ &= P^{dep}(f(X) = z'|Y = y') \end{aligned} \tag{4.21}$$

Now, by assumption, the different data distributions *train* and *deployment* only differ from the "true" distribution by the prevalence of the different s-y-combinations, with the *deployment* data distribution having all combinations but potentially not in equal quantity. However, as we restrict ourselves to a certain combination ($S = s', Y = y'$) in the above equation, the equation also holds in the true data distribution:

$$\begin{aligned} &P(f(X) = z'|S = s', Y = y') \\ &= P(f(X) = z'|Y = y') \end{aligned} \tag{4.22}$$

Then, using Bayes' rule, we get

$$\begin{aligned} &P(S = s'|f(X) = z', Y = y') \\ &= \frac{P(f(X) = z'|S = s', Y = y')P(S = s'|Y = y')}{P(f(X) = z'|Y = y')} \\ &= P(S = s'|Y = y') \,. \end{aligned} \tag{4.23}$$

Finally, in the true data distribution, we have a uniform prior: $P(S = s'|Y = y') = (n_s)^{-1}$. This concludes the proof. □

4.7.3 *Bound on error introduced by clustering*

As previously stated, in practice, no labels are available for the deployment set. Instead, we identify the relevant groupings by clustering. Such clustering cannot be expected to be perfect. So, how will clustering affect the calculation of our objective?

THEOREM 4.4. *If $q_f(Z)$ is a data distribution on $\mathcal{Z}$ that is a mixture of $n_y \cdot n_s$ Gaussians, which correspond to all unique combinations of $y \in \mathcal{Y}$ and $s \in \mathcal{S}$, and $p_f(Z)$ is any data distribution on $\mathcal{Z}$, then without knowing y and s on q_f, we can estimate*

$$\sum_{s' \in \mathcal{S}} \sum_{y' \in \mathcal{Y}} TV(p_f|_{s \in \Pi(s',y'), Y=y'}, q_f|_{S=s', Y=y'}) \tag{4.24}$$

with an error that is bounded by $\tilde{O}(\sqrt{1/N})$ with high probability, where N is the number of samples drawn from q_f for learning.

Proof. First, we produce an estimate $\hat{q}_f$ of q_f using the algorithm from Ashtiani et al. (2020), which gives us a mixture-of-Gaussian distribution of $n_y \cdot n_s$ components with $TV(q_f, \hat{q}_f) \leq \tilde{O}(\sqrt{1/N})$ with high probability, where N is the number of data points used for learning the estimate. Then, by Lemma 3 from Sohoni et al. (2020), *there exists* a mapping i from the components k of the Gaussian mixture $\hat{q}_f$ to the s-y-combinations in q_f such that

$$\begin{aligned} &TV(q_f(Z|S=s', Y=y'), \hat{q}_f(Z|k=i(s',y'))) \\ &\qquad \leq \tilde{O}\left(\frac{1}{\sqrt{N}}\right) . \end{aligned} \tag{4.25}$$

Now, consider the element of the sum in the objective that corresponds to (s', y'):

$$\begin{aligned} &TV(p_f(Z|s \in \Pi(s',y'), Y=y'), q_f(z|S=s', Y=y')) \\ \leq\ &TV(p_f(Z|s \in \Pi(s',y'), Y=y'), \hat{q}_f(Z|k=i(s',y'))) \\ &\quad + TV(\hat{q}_f(z|k=i(s',y')), q_f(z|S=s', Y=y')) \\ \leq\ &TV(p_f(Z|s \in \Pi(s',y'), Y=y'), \hat{q}_f(Z|k=i(s',y'))) \\ &\quad + \tilde{O}(1/\sqrt{N}) \end{aligned} \tag{4.26}$$

Thus, for the whole sum over s and y, the error is bounded by

$$\sum_{s' \in \mathcal{S}} \sum_{y' \in \mathcal{Y}} \tilde{O}(\sqrt{1/N}) \leq n_s n_y \max_{(s',y') \in \mathcal{S} \times \mathcal{Y}} \tilde{O}(\sqrt{1/N}) \tag{4.27}$$

which is equivalent to just $\tilde{O}(\sqrt{1/N})$. □

4.8 Dataset construction

4.8.1 *Coloured MNIST and biasing parameters*

The MNIST dataset (LeCun et al., 1998) consists of 70,000 (60,000 designated for training, 10,000 for testing) images of grey-scale hand-written digits. We colour the digits following the procedure outlined in Kehrenberg et al. (2020), randomly assigning each sample one of ten distinct RGB colours. Each source is then a combination of digit-class (class label) and colour (subgroup label). We use no data-augmentation aside from symmetrically zero-padding the images to be of size 32x32.

To simulate a more real-world setup where the data, labelled or otherwise, is not naturally balanced, we bias the Coloured MNIST training and deployment sets by downsampling certain colour/digit combinations. The proportions of each such combination *retained* in the *subgroup bias* (in which we have one source missing from the training set) and *missing subgroup* (in which we have two sources missing from the training set) are enumerated in table 4.1 and 4.2, respectively. For the 3-digit-3-colour variant of the problem, no biasing is applied to either the deployment set or the training set (the missing combinations are specified in the caption accompanying figure 4.15); this variant was experimented with only under the subgroup-bias setting.

Table 4.1: Biasing parameters for the training (left) and deployment (right) sets of Coloured MNIST in the *subgroup bias* setting.

Combination	Proportion retained	
	training set	deployment set
(Y = 2, S = purple)	1.0	0.7
(Y = 2, S = green)	0.3	0.4
(Y = 4, S = purple)	0.0	0.2
(Y = 4, S = green)	1.0	1.0

Table 4.2: Biasing parameters for the training (left) and deployment (right) sets of Coloured MNIST in the *missing subgroup* setting.

Combination	Proportion retained	
	training set	deployment set
(Y = 2, S = purple)	0.0	0.7
(Y = 2, S = green)	0.85	0.6
(Y = 4, S = purple)	0.0	0.4
(Y = 4, S = green)	1.0	1.0

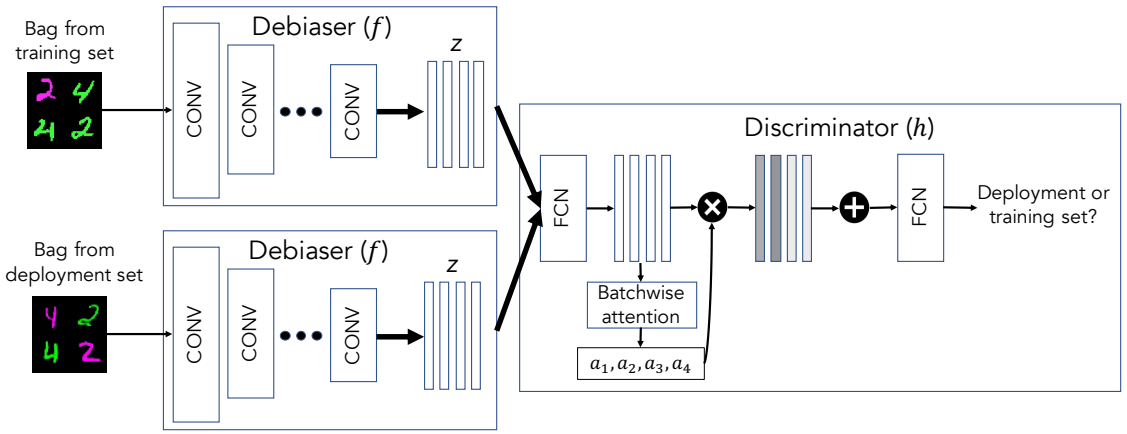


Figure 4.8: (discriminator). The debiaser is trained to produce encodings, z of the data that are invariant the subgroups differences. it. In order to determine whether a bag of encodings originates from the training set or the deployment set, the discriminator performs an attention-weighted aggregation over the bag dimension to model interdependencies between the samples. In the case of Coloured MNIST where purple fours constitute the missing subgroup, the discriminator can identify an encoding of a bag from the training set by the absence of such samples as long as color information is detectable in z. learning a subgroup invariant representation, the debiaser can hide the origin of the bags from the discriminator.

4.9 Model details and optimization

4.9.1 *Overview of model architecture*

We give a more detailed explanation of the model used in our method. Fig. 4.8 shows the core of our method: the debiaser, f, which produces bags of encodings, z – on both the training and the deployment set – which are then fed to a discriminator that tries to identify the origin of the bags. The discriminator uses batch-wise attention in order to consider a bag as a whole, which allows cross-comparisons.

4.9.2 *Details of the attention mechanism*

The *discriminator* function h that predicts which dataset a bag of samples embedded in z was sampled from should have the following property: $h((f(x)|x \in B)) = h((f(x)|x \in \pi(B)))$ for all permutations π, and $f : x \to z$. For the entirety of function h – composed of sub-functions $h_1(h_2(h_3...)))$ – to have this property, it suffices that only the innermost sub-function, ρ, does. While there are a number of choices when it comes to defining ρ, we choose a weighted average $\rho = \frac{1}{|B|} \sum_{x \in B} \text{attention}(f(x), B) \cdot f(x)$, with weights computed according to a learned attention mechanism. The idea of using an attention mechanism for set-wise classification has been previously explored to great success by, e.g. Lee et al. (2019b). We employ an bag-wise attention mechanism based on the scaled dot-product attention of Vaswani et al. (2017), where in our case we define K and V to be linear projections of z – zW_k and zW_v, respectively – and Q to be mean of another linear projection of z, zW_q, taken over the bag dimension.

$$\text{Attention}(Q, K, V) := \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

The output of ρ is then further processed by a series of fully-connected layers and the final output is the binary prediction for a given bag of samples.

4.9.3 *Training procedure and hyperparameters*

The hyperparameters and architectures for the auto-encoder (AE) (`AE`), Predictor and Discriminator sub-networks are detailed in Table 4.3 for all three datasets.We train all models using `Adam` (Kingma and Ba, 2015).

For the Coloured MNIST and CelebA datasets, the baseline `ERM`, `DRO`, `LfF` (in the case of the former) and `gDRO` (in the case of the latter) models use a convolutional backbone consisting of one Conv-BN-LReLU block per "stage", with each stage followed by max-pooling operation to spatially downsample by a factor of two to produce the subsequent stage. This backbone consists of 4 and 5 stages for Coloured MNIST and CelebA, respectively. The output of the backbone is flattened and fed to a single fully-connected layer of size $|Y|$ in order to obtain the class-prediction, $\hat{y}_i$, for a given instance. To evaluate our method, we simply train a linear classifier on top of z; this is sufficient due to linear-separability being encouraged during training by the y-predictor. For the Adult Income dataset, we use an MLP composed of a single hidden layer 35 units in size, followed by a SELU activation (Klambauer et al., 2017), as both the downstream classifier for our method, and as the network architecture of the baselines. All baselines and downstream classifiers alike were trained for 60 epochs with a learning rate of 1×10^{-3} and a batch size of 256.

Since, by design, we do not have labels for all subgroups the model will be tested on, and bias against these missing subgroups is what we aim to combat, properly validating, and thus conducting hyperparameter selection for models generally, is not straightforward. Indeed, performing model-selection for domain generalisation problems is well-known to be a difficult problem (Gulrajani and Lopez-Paz, 2021). We can use estimates of the mutual information between the learned-representation and s and y (which we wish to minimize w.r.t. to the former, maximise w.r.t. the latter) to guide the process, though optimizing the model w.r.t. to these metrics obtained from only the training set does not guarantee generalisation to the missing subgroups. We can, however, additionally measure the entropy of the predictions on the encoded test set and seek to maximise it across all samples, or alternatively train a discriminator of the same kind used for distribution matching as a measure of the shift in the latent space between datasets. We use the latter approach (considering the combination of the learned distance between subspace distributions and reconstruction loss) to inform an extensive grid-search over the hyperparameter space for our method.

For the `DRO` baseline, we allowed access to the labels of the test set for the purpose of hyperparameter selection, performing a grid-search over multiple splits to avoid overfitting to any particular instantiation. Specifically, the threshold (η) parameter for `DRO` was determined by a grid-search over the space $\{0.01, 0.1, 0.3, 1.0\}$.

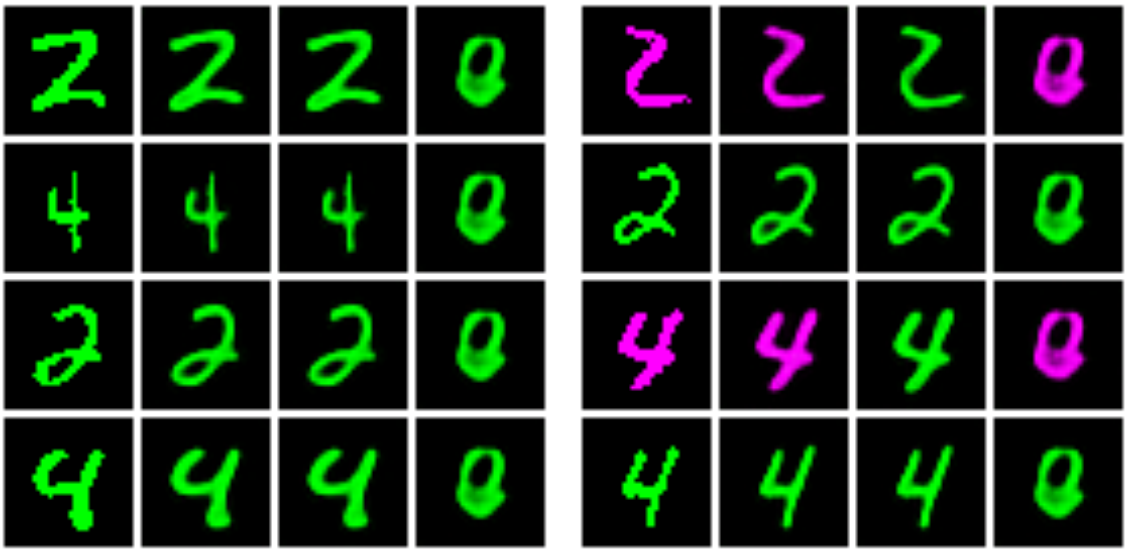
Table 4.3: Selected hyperparameters for experiments with Coloured MNIST, Adult and CelebA datasets.

	Coloured MNIST 2-dig SB / 2-dig MS / 3-dig SB	**Adult**	**CelebA**
Input size	$3 \times 32 \times 32$	61	$3 \times 64 \times 64$
	AutoEncoder		
Levels	4	1	5
Level depth	2	1	2
Hidden units / level	$[32, 64, 128, 256]$	$[61]$	$[32, 64, 128, 256, 512]$
Activation	GELU	GELU	SiLU
Layer-wise Normalisation	-	-	LayerNorm
Downsampling op.	Strided Convs.	–	Strided Convs.
Reconstruction loss	MSE	Mixed[1]	MSE
Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}
	Clustering		
Batch size	256	1000	256
AE pre-training epochs	150	100	10
Clustering epochs	100	300	20
Self-supervised loss	Cosine + BCE	Cosine + BCE	Cosine
U (for ranking statistics)	5	3	8
	Support-Matching		
Batch size	1/32/14	64	32
Bag size	256/8/18	32	8
Training iterations	8k/8k/20k	5k	2k
Encoding (z) size[2]	128	35	128
Binarised $\tilde{s}$	✗ / ✓ / ✓	✗	✗
y-predictor weight (λ_1)	1	0	1
s-predictor weight (λ_2)	1	0	1
Adversarial weight (λ_3)	1×10^{-3}	1	1
Stop-gradient $\left(\nabla_\theta h_\psi(f_\theta(X^{dep})) = 0\right)$	✗	✓	✗
	Predictors		
Learning rate	3×10^{-4}	1×10^{-3} 1×10^{-3}	
	Discriminator		
Attention mechanism[3]	Gated	Gated	Gated
Hidden units pre-aggregation	$[256, 256]$	$[32]$	$[256, 256]$
Hidden units post-aggregation	$[256, 256]$	–	$[256, 256]$
Embedding dim (for attention)	32	128	128
Activation	GELU	GELU	GELU
Learning rate	3×10^{-4}	1×10^{-3}	1×10^{-3}
Updates / AE update	1	3	1

[1] Cross-entropy is used for categorical features, MSE for continuous features.

[2] $|z|$ denotes the combined size of $\tilde{s}$ and z, with the former occupying $\lceil \log_2(\mathcal{S}) \rceil$ dimensions, the latter the remaining dimensions.

[3] The attention mechanism used for computing the sample-weights within a bag. *Gated* refers to gated attention proposed by Ilse et al. (2018), while *SDP* refers to the scaled dot-product attention proposed by Vaswani et al. (2017).



(a) Different reconstructions on the training set. Corresponding to: original, full reconstruction, reconstruction of z ($\tilde{s}$ zeroed out), reconstruction of $\tilde{s}$ (z zeroed out).

(b) Different reconstructions on the deployment set. Corresponding to: original, full reconstruction, reconstruction of z ($\tilde{s}$ zeroed out), reconstruction of $\tilde{s}$ (z zeroed out).

Figure 4.9: Visualisation of our method's solutions for the Coloured MNIST dataset, with purple as the missing subgroup. In each of the subfigures 4.9a and 4.9b: Column 1 shows the original images from x from the respective set. Column 2 shows plain reconstructions generated from $x_{recon} = g(f(x), t(x))$. Column 3 shows reconstruction with zeroed-out $\tilde{s}$: $g(f(x), 0)$, which effectively visualizes z. Column 4 shows the result of an analogous process where z was zeroed out instead.

4.9.4 *Visualisations of results*

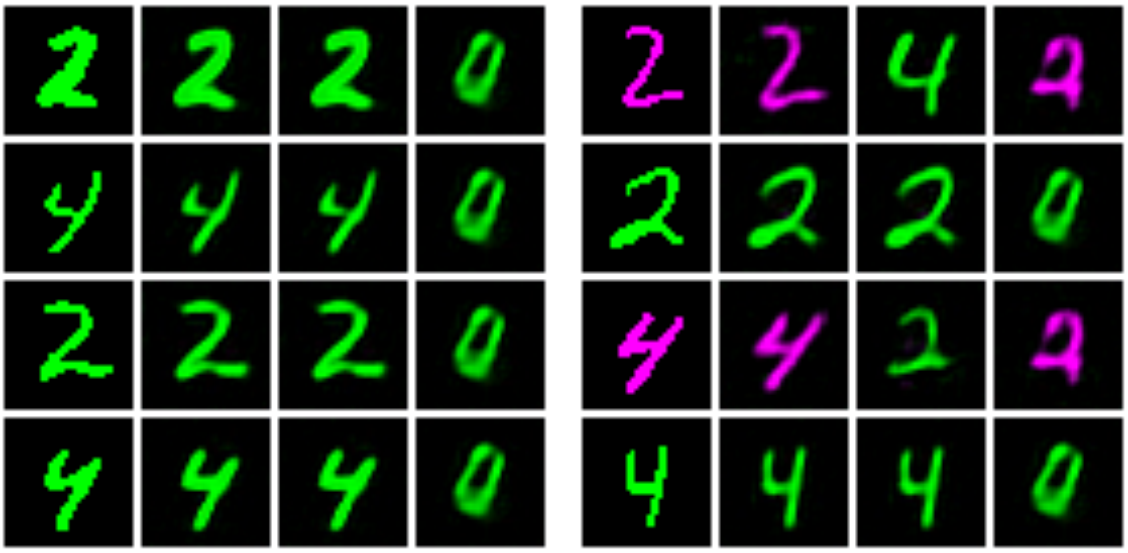
We show qualitative results of the disentangling in figures 4.9, 4.10 (both Coloured MNIST), and 4.11 (CelebA). Fig. 4.9 shows successful disentangling (from a run that achieved close to 100% accuracy); in the deployment set the representation z has lost all colouring (see column 3 in the figures). Fig. 4.10 on the other hand, shows a visualisation from a *failed* run; instead of encoding purple 2's and green 2's with the same representation, the model here encoded purple 2's and green 4's as similar. This is a valid solution of the given optimisation problem – the representation is invariant to training set vs deployment set – but it is definitely not the intended solution.

Fig. 4.11 shows visualisations for CelebA. With a successful disentangling, column 3 (visualisation of z) should show a version of the image that is "gender-neutral" (i.e. invariant to gender). Furthermore, column 4 (visualisation of $\tilde{s}$) should be invariant to the class label (i.e. "smiling"), so the images should either be all with smiles or all without smiles.

Fig. 4.12 shows attention maps for bags from the deployment set. We can see that the model pays special attention to those samples that are not included in the training set. For details, see the captions.

4.9.5 *Additional metrics*

Figures 4.13, 4.14, and 4.16 show the TPR ratio and the TNR ratio as additional metrics for Coloured MNIST (2 digits) and CelebA. These are computed as the ratio of TPR (or TNR) on subgroup $s = 0$ over the TPR (or TNR) on subgroup $s = 1$; if this gives a number greater than 1, the inverse is



(a) Different reconstructions on the training set. Corresponding to: original, full reconstruction, reconstruction of z ($\tilde{s}$ zeroed out), reconstruction of $\tilde{s}$ (z zeroed out).

(b) Different reconstructions on the deployment set. Corresponding to: original, full reconstruction, reconstruction of z ($\tilde{s}$ zeroed out), reconstruction of $\tilde{s}$ (z zeroed out).

Figure 4.10: Visualisation of a failure of our method for the Coloured MNIST dataset, with purple as the missing subgroup. In each of the subfigures 4.10a and 4.10b: Column 1 shows the original images from x from the respective set. Column 2 shows plain reconstructions generated from $x_{recon} = g(f(x), t(x))$. Column 3 shows reconstruction with zeroed-out $\tilde{s}$: $g(f(x), 0)$, which effectively visualises z. Column 4 shows the result of an analogous process where z was zeroed out instead.

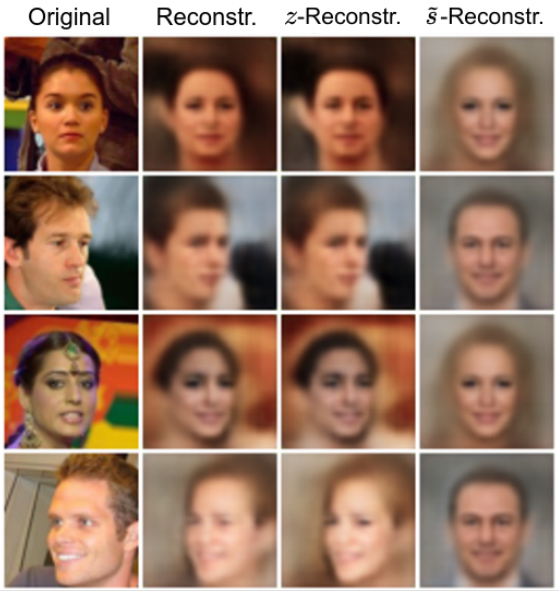


Figure 4.11: Visualisation of our method's solutions for the CelebA dataset, with "smiling females" as the missing subgroup. Column 1 shows the original images from x from the deployment set of CelebA. Column 2 shows plain reconstructions generated from $x_{recon} = g(f(x), t(x))$. Column 3 shows reconstruction with zeroed-out $\tilde{s}$: $g(f(x), 0)$, which effectively visualises z. Column 4 shows the result of an analogous process where z was zeroed out instead.

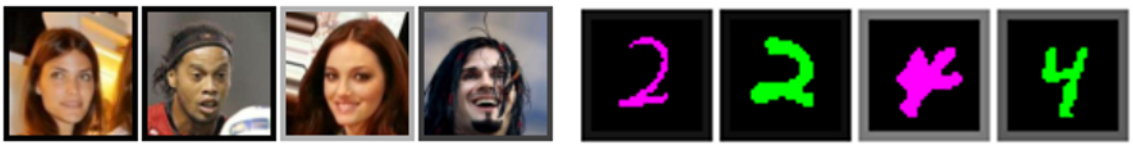


Figure 4.12: Example sample-wise attention maps for bags of CelebA (left) and CMNIST (right) images sampled from a balanced deployment set. The training set is biased according to the *subgroup bias* setting where for CelebA "smiling females" constitute the missing source and for Coloured MNIST purple fours constitute the missing source. The attention weights are used during the discriminator's aggregation step to compute a weighted sum over the bag. The attention-weight assigned to each sample is proportional to the lightness of its frame, with black signifying a weight of 0, white a weight of 1. Those samples belonging to the missing subgroup are assigned the highest weight as they signal from which dataset (training vs. deployment) the bag containing them was drawn from.

taken. Similarly to the PR ratio reported in the main paper, these ratios give an indication of how much the prediction of the classifier depends on the subgroup label s.

Fig. 4.15 shows metrics specific to multivariate s (i.e. non-binary s). We report the minimum (i.e. farthest away from 1 of the pairwise ratios (TPR/TNR ratio min) as well as the largest difference between the raw values (TPR/TNR diff max). Additionally, we compute the HGRMC between S and Y, serving as a measure of dependence defined between two variables with arbitrary support.

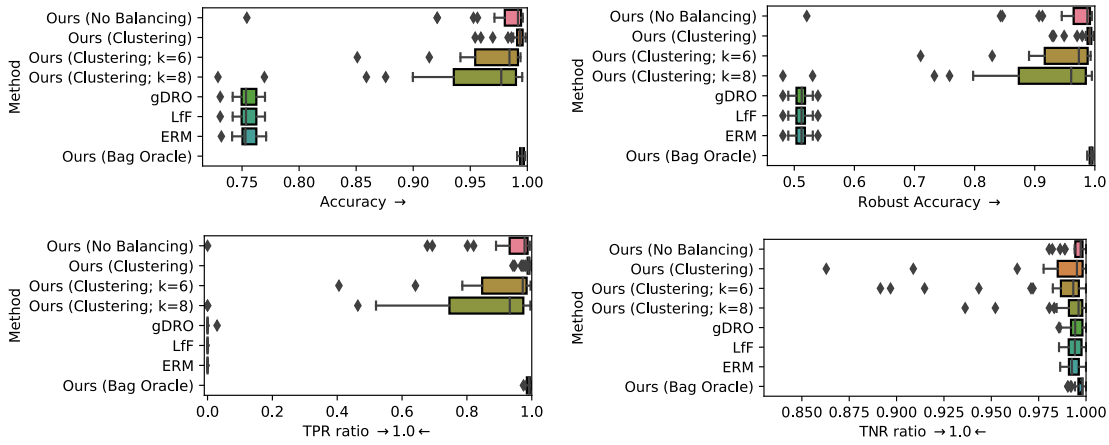


Figure 4.13: Results from **30 repeats** for the Coloured MNIST dataset with two digits, 2 and 4, with *subgroup bias* for the colour 'purple': for purple, only the digit class '2' is present. **Top left**: Accuracy. **Top right**: Positive rate ratio. **Bottom left**: True positive rate ratio. **Bottom right**: True negative rate ratio. For `Ours (Clustering)`, the clustering accuracy was 96% ± 6%. For an explanation of `Ours (Clustering; k=6/8)` see §4.10.2.

4.10 Ablation studies

4.10.1 *Using an instance-wise loss instead of a set-wise loss*

See Fig. 4.17 and Fig. 4.18 for results on 2-digit Coloured MNIST (under the *subgroup bias* and *missing subgroup* settings, respectively) for our method but with the loss computed instance-wise (`Inst.`) as opposed to set-wise, as is typical of adversarial unsupervised domain adaptation methods (e.g. Ganin et al., 2016). All aspects of the method, other than those directly involved

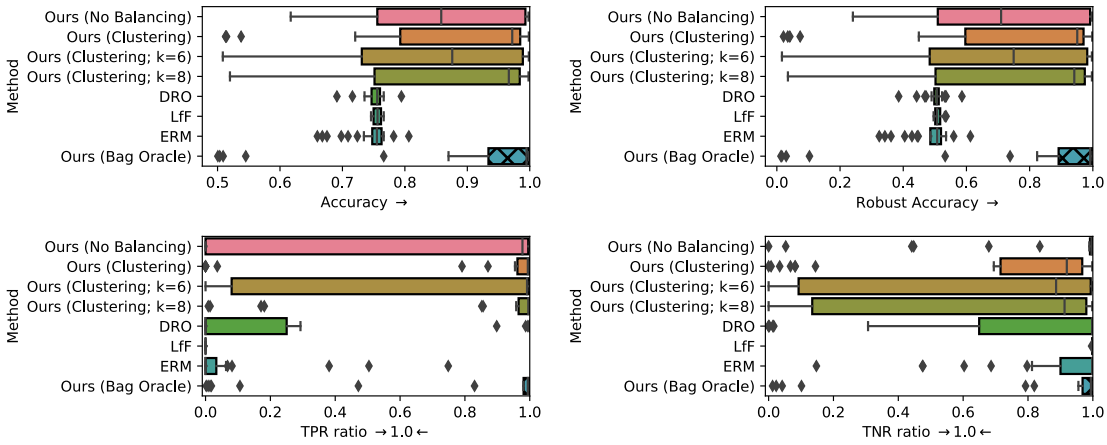


Figure 4.14: Results from **30 repeats** for the Coloured MNIST dataset with two digits, 2 and 4, with a *missing subgroup*: the training dataset only has green digits. **Top left**: Accuracy. **Top right**: Robust Accuracy. **Bottom left**: True positive rate ratio. **Bottom right**: True negative rate ratio. For `Ours (Clustering)`, the clustering accuracy was 88% ± 5%. For an explanation of `Ours (Clustering; k=6/8)` see §4.10.2.

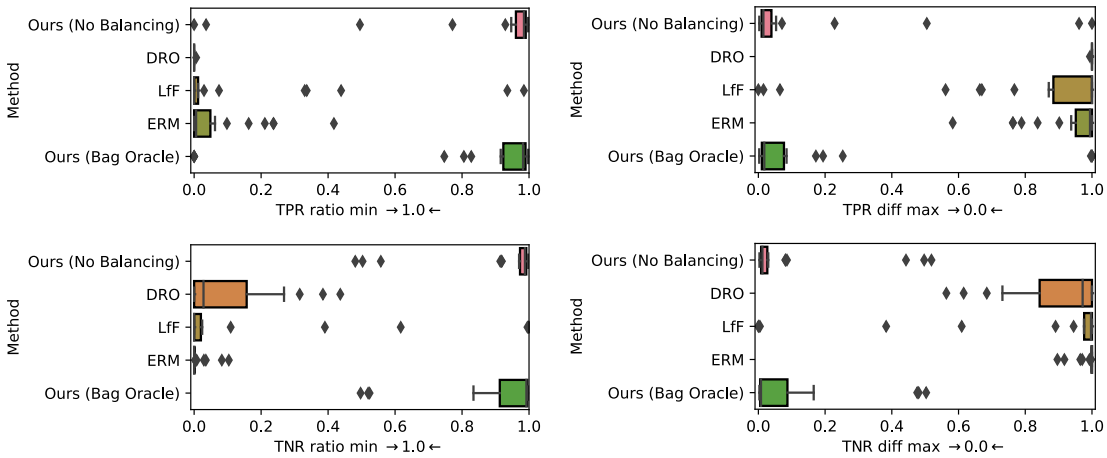


Figure 4.15: Results from **30 repeats** for the Coloured MNIST dataset with three digits: '2', '4' and '6'. Four combinations of digit and color are missing: green 2's, blue 2's, blue 4's and green 6's. **First row, left**: minimum of all true positive rate ratios. **First row, right**: maximum of all true positive rate differences. **Second row, left**: minimum of all true negative rate ratios. **Second row, right**: maximum of all true negative rate differences.

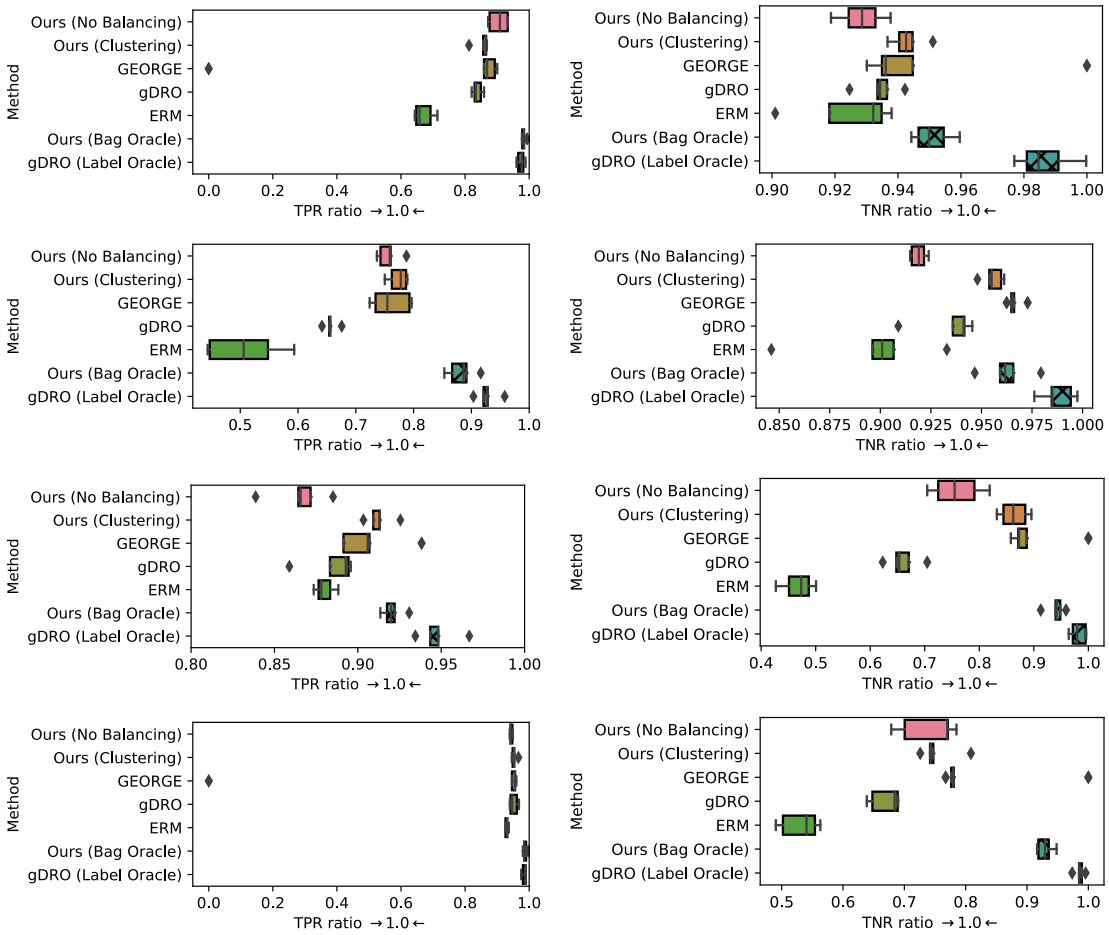


Figure 4.16: Plots of additional metrics for CelebA under the SB setting, where "smiling" is the class label and "gender" is the subgroup label. These metrics are ratios computed between the *Male* and *Female* subgroups with the largest of the two values involved always selected as the denominator. **Left:** TNR ratio. **Right**: TNR ratio.

in the loss-computation, were kept constant – this includes the use of hierarchical balancing, despite the necessary removal of the aggregation layer meaning the discriminator is no longer sensitive to the bagging. It is clear that the aforementioned change to the loss drastically increases the variance (IQR) of the results for both settings and, at the same time, drastically reduces the median `Robust Accuracy` to the point of being only marginally above that of the `ERM` baseline, regardless of the chosen balancing scheme.

4.10.2 *Clustering with an incorrect number of clusters*

We also investigate what happens when the number of clusters is set incorrectly. For 2-digit Coloured MNIST, we expect 4 clusters, corresponding to the 4 possible combinations of the binary class label y and the binary subgroup label s. However, there might be circumstances where the correct number of clusters is not known; how does the batch balancing work in this case? We run experiments with the number of clusters set to 6 and to 8, with all other aspects of the pipeline kept the same. It should be noted that this is a very naïve way of dealing with an unknown number of clusters. There are methods specifically designed for identifying the right number of

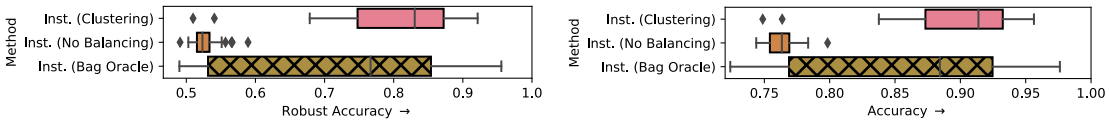


Figure 4.17: Results from **30 repeats** with an *instance-wise* loss for the Coloured MNIST dataset with two digits, 2 and 4, with *subgroup bias* for the colour 'purple': for purple, only the digit class '2' is present. **Left**: Accuracy. **Right**: Positive rate ratio. For `Inst. (Clustering)`, the clustering accuracy was 96% ± 6%.

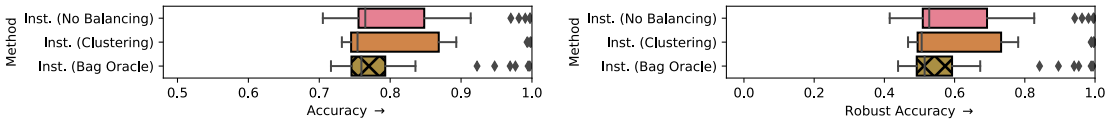


Figure 4.18: Results from **30 repeats** with an *instance-wise* loss for the Coloured MNIST dataset with two digits, 2 and 4, with a *missing subgroup*: the training dataset only has green digits. **Left**: Accuracy. **Right**: Robust Accuracy. For `Inst. (Clustering)`, the clustering accuracy was 88% ± 5%.

clusters (Hamerly and Elkan, 2004; Chazal et al., 2013), and that is what would be used if this situation arose in practice.

The results can be found in Fig. 4.13 and Fig. 4.14. Bags and batches are constructed by drawing an equal number of samples from each cluster. Unsurprisingly, the method performs worse than with the correct number of clusters. When investigating how the clustering methods deal with the larger number of clusters, we found that it is predominantly those samples that do not appear in the training set which get spread out among the additional clusters. This is most likely due to the fact that the clustering is semi-supervised, with those clusters that occur in the training set having supervision. The overall effect is that the samples which are not appearing in the training set are over-represented in the drawn bags, which means it is easier for the adversary to identify where the bags came from, and the encoder cannot properly learn to produce an invariant encoding.

4.11 Adapting GEORGE

As discussed in the main text, GEORGE (Sohoni et al., 2020) was originally developed to address the uneven performance resulting from hidden stratification, though hidden stratification of a different kind to the one we consider. In Sohoni et al. (2020) the training set comes with (super-)class labels but without subclass (or *subgroup* in our terminology) labels. The training set is unlabelled with respect to the subclass, but all superclass-subclass combinations (or "sources") are assumed to be present in the training data and therefore discoverable via clustering. (Note that the clustering in Sohoni et al. (2020) is – in contrast to our method – completely without supervision and there is nothing to guide the clustering towards discovering the subgroups of interest, apart from the assumption that they are the most salient.) On the other hand, in our setting, we do have access to all sources expected at deployment time, but not all of them are present in the training data – some are exclusively found in the *unlabelled* deployment set.

This necessitates propagating the labels from the training set to the deployment set, which can be done within the clustering step to ensure consistency between the cluster labels and the propagated superclass labels. Doing so requires us to modify the clustering algorithm such that

instead of predicting each source independently of one another, we factorise the joint distribution of the super- and subclasses, $P(Y, S)$ into their respective marginal distributions, $P(Y)$ and $P(S)$. In practice, this is achieved by applying two separate cluster-prediction heads to the image representation, z: one, μ_y, predicting the superclass, y, the other, μ_s predicting the subclass, s. This allows us to decouple the supervised loss for the two types of label and to always be able to recover y due to having full supervision in terms of its ground-truth labels from the training set – this means we can identify all the y clusters with the right y labels. This is not necessarily possible for s, because some s values might be completely missing from the labelled training set (missing subgroup setting).

With the outputs structured as just described, we can obtain the prediction for a given sample's source (which is needed to compute the unsupervised clustering loss and for balancing the deployment set), by taking the argmax of the vectorized outer product of the softmaxed outputs of the two heads:

$$\omega_i = \arg\max_k \ \text{vec}(\mu_y(z_i) \otimes \mu_s(z_i))_k \ , \qquad k = 1, ..., |S \times Y| \tag{4.28}$$

where $\mu_y(z_i)$ and $\mu_s(z_i)$ are vectors, $\otimes$ is the outer product, and $\text{vec}(A)$ is the vectorisation of matrix A. After training the clustering model, we can then use it to generate predictions $\hat{Y}^{dep}$, as well as the cluster labels $\hat{\Omega}^{dep}$, for the deployment set, and use them together to train a robust classifier with gDRO (Sagawa et al., 2019), as in Sohoni et al. (2020).

4.12 Code

The code can be found here: https://github.com/wearepal/support-matching.

4.13 Authorial contributions

T. Kehrenberg conceived of the initial idea of overcoming the limitation of the partially-annotated representative set in Chapter 3 through the use of distribution matching, wrote much of the original implementation and text, and was the primary experiment-runner during the nascent distribution-matching stage of the project. Later in the project, he notably worked to establish theoretical guarantees for the support-matching method and a more rigorous formulism of the problem setup, while also continuing to aid with experiment-running and paper-writing (both to a reduced, but still significant, degree).

I proposed converting the problem of distribution-matching, proposed by T. Kehrenberg, into one of support-matching by means of source-balanced bags and a set-discriminator, a necessary element for achieving the desired goal of subgroup-invariance while preserving variance to the target. On the practical front, while much of the original codebase was written by T. Kehrenberg, the lion's share of the (several-times) rewritten and extended (additional datasets, baselines, discriminator methods, etc.) one was authored by myself; a similar story applies both to the text, with much of the latest version (save the theoretical sections) being of my hand, and to the experiment-running (and the implied model-selection).

V. Sharmanska conceived, and gave first form to, the problem setting, and wrote significant portions of the initial versions of the paper – those related to the introduction and problem setup primarily. In later stages of the project, she took on an important advisory role and gave feedback on revisions of the paper.

N. Quadrianto suggested the combining of distribution-matching with clustering-derived sample-weighting during the initial stages – this being a major stepping stone in the development of the eventual support-matching method (for which said weighting was replaced by exact bag-based balancing) – wrote significant portions of the original text, and ran experiments primarily on the clustering side. He was also responsible for generally supervising the project, by which I mean, inter alia, discussing and advising on current progress and future avenues, and providing, feedback on revisions of the paper.

BIBLIOGRAPHY

Rényi, Alfréd (1959). 'On measures of dependence'. In: *Acta Mathematica Academiae Scientiarum Hungarica* 10.3-4, pp. 441–451.

LeCun, Yann, Léon Bottou, Yoshua Bengio and Patrick Haffner (1998). 'Gradient-based learning applied to document recognition'. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Maron, Oded and Tomás Lozano-Pérez (1998). 'A framework for multiple-instance learning'. In: *Advances in neural information processing systems*, pp. 570–576.

Liu, Bing, Wee Sun Lee, Philip S Yu and Xiaoli Li (2002). 'Partially supervised classification of text documents'. In: *ICML*. Vol. 2. Sydney, NSW, pp. 387–394.

Liu, Bing, Yang Dai, Xiaoli Li, Wee Sun Lee and Philip S Yu (2003). 'Building text classifiers using positive and unlabeled examples'. In: *Third IEEE international conference on data mining*. IEEE, pp. 179–186.

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

Hamerly, Greg and Charles Elkan (2004). 'Learning the k in k-means'. In: *Advances in Neural Information Processing Systems* 16, pp. 281–288.

Chapelle, Olivier, Bernhard Schölkopf and Alexander Zien (2006). 'Introduction to Semi-Supervised Learning'. In: *Semi-Supervised Learning*. Ed. by Olivier Chapelle, Bernhard Schölkopf and Alexander Zien. The MIT Press, pp. 1–12.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf and Alexander Smola (2012). 'A kernel two-sample test'. In: *The Journal of Machine Learning Research* 13.1, pp. 723–773.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Chazal, Frédéric, Leonidas J. Guibas, Steve Y. Oudot and Primoz Skraba (2013). 'Persistence-Based Clustering in Riemannian Manifolds'. In: *J. ACM* 60.6. DOI: 10.1145/2535927.

Deb, Kalyanmoy (2014). 'Multi-objective optimization'. In: *Search methodologies*. Springer, pp. 403–449.

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

Kingma, Diederik P. and Jimmy Ba (2015). 'Adam: A Method for Stochastic Optimization'. In: *International Conference on Learning Representations (ICLR)*.

Liu, Ziwei, Ping Luo, Xiaogang Wang and Xiaoou Tang (2015). 'Deep Learning Face Attributes in the Wild'. In: *Proceedings of International Conference on Computer Vision (ICCV)*.

Ren, Shaoqing, Kaiming He, Ross Girshick and Jian Sun (2015). 'Faster r-cnn: Towards real-time object detection with region proposal networks'. In: *Advances in neural information processing systems* 28, pp. 91–99.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky (2016). 'Domain-adversarial training of neural networks'. In: *The Journal of Machine Learning Research* 17.1, pp. 2096–2030.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2016). 'Inherent trade-offs in the fair determination of risk scores'. In: *arXiv preprint arXiv:1609.05807*.

Chouldechova, Alexandra (2017). 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments'. In: *Big data* 5.2, pp. 153–163.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr and Sepp Hochreiter (2017). 'Self-normalizing neural networks'. In: *Advances in neural information processing systems*, pp. 971–980.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin (2017). 'Attention is all you need'. In: *Advances in neural information processing systems*, pp. 5998–6008.

AlBadawy, Ehab A, Ashirbani Saha and Maciej A Mazurowski (2018). 'Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing'. In: *Medical physics* 45.3, pp. 1150–1158.

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

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

Hashimoto, Tatsunori B., Megha Srivastava, Hongseok Namkoong and Percy Liang (2018). 'Fairness Without Demographics in Repeated Loss Minimization'. In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1934–1943.

Ilse, Maximilian, Jakub Tomczak and Max Welling (2018). 'Attention-based deep multiple instance learning'. In: *International conference on machine learning*. PMLR, pp. 2127–2136.

Madras, David, Elliot Creager, Toniann Pitassi and Richard Zemel (2018). 'Learning adversarially fair and transferable representations'. In: *arXiv preprint arXiv:1802.06309*.

Mayson, Sandra G (2018). 'Bias in, bias out'. In: *YAle lJ* 128, p. 2218.

Saito, Kuniaki, Kohei Watanabe, Yoshitaka Ushiku and Tatsuya Harada (2018). 'Maximum classifier discrepancy for unsupervised domain adaptation'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732.

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

Ying, Rex, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton and Jure Leskovec (2018). 'Graph convolutional neural networks for web-scale recommender systems'. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 974–983.

Zhao, Han, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira and Geoffrey J Gordon (2018). 'Adversarial multiple source domain adaptation'. In: *Advances in neural information processing systems* 31, pp. 8559–8570.

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

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

Codevilla, Felipe, Eder Santana, Antonio M López and Adrien Gaidon (2019). 'Exploring the limitations of behavior cloning for autonomous driving'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338.

Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi and Richard Zemel (2019). 'Flexibly fair representation learning by disentanglement'. In: *arXiv preprint arXiv:1906.02589*.

De Haan, Pim, Dinesh Jayaraman and Sergey Levine (2019). 'Causal confusion in imitation learning'. In: *Advances in Neural Information Processing Systems* 32.

Kim, Byungju, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim and Junmo Kim (2019). 'Learning not to learn: Training deep neural networks with biased data'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, Chen-Yu, Tanmay Batra, Mohammad Haris Baig and Daniel Ulbricht (2019a). 'Sliced wasserstein discrepancy for unsupervised domain adaptation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295.

Lee, Juho, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi and Yee Whye Teh (2019b). 'Set transformer: A framework for attention-based permutation-invariant neural networks'. In: *International Conference on Machine Learning*. PMLR, pp. 3744–3753.

Quadrianto, Novi, Viktoriia Sharmanska and Oliver Thomas (2019). 'Discovering fair representations in the data domain'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8227–8236.

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

Sun, Yu, Eric Tzeng, Trevor Darrell and Alexei A Efros (2019). 'Unsupervised domain adaptation through self-supervision'. In: *arXiv preprint arXiv:1909.11825*.

Ashtiani, Hassan, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian and Yaniv Plan (Oct. 2020). 'Near-Optimal Sample Complexity Bounds for Robust Learning of Gaussian Mixtures via Compression Schemes'. In: *Journal of the ACM* 67.6. ISSN: 0004-5411. DOI: 10.1145/3417994. URL: https://doi.org/10.1145/3417994.

Bekker, Jessa and Jesse Davis (2020). 'Learning from positive and unlabeled data: a survey'. In: *Machine Learning* 109.4, pp. 719–760.

Castro, Daniel C, Ian Walker and Ben Glocker (2020). 'Causality matters in medical imaging'. In: *Nature Communications* 11.1, pp. 1–10.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge and Felix A Wichmann (2020). 'Shortcut learning in deep neural networks'. In: *Nature Machine Intelligence* 2.11, pp. 665–673.

Kehrenberg, Thomas, Myles Scott Bartlett, Oliver Thomas and Novi Quadrianto (2020). 'Null-sampling for Interpretable and Fair Representations'. In: *Computer Vision – ECCV 2020*. Glasgow, UK: Springer International Publishing. ISBN: 9783030586041. DOI: 10.1007/978-3-030-58604-1.

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

Nam, Jun Hyun, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee and Jinwoo Shin (2020). 'Learning from Failure: De-biasing Classifier from Biased Classifier'. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan and Hsuan-Tien Lin.

Sagawa, Shiori, Aditi Raghunathan, Pang Wei Koh and Percy Liang (2020). 'An Investigation of Why Overparameterization Exacerbates Spurious Correlations'. In: *CoRR* abs/2005.04345.

Seyyed-Kalantari, Laleh, Guanxiong Liu, Matthew McDermott, Irene Y Chen and Marzyeh Ghassemi (2020). 'CheXclusion: Fairness gaps in deep chest X-ray classifiers'. In: *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. World Scientific, pp. 232–243.

Sohoni, Nimit Sharad, Jared Dunnmon, Geoffrey Angus, Albert Gu and Christopher Ré (2020). 'No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems'. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan and Hsuan-Tien Lin.

Yu, Tianhe, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman and Chelsea Finn (2020). 'Gradient surgery for multi-task learning'. In: *Advances in Neural Information Processing Systems* 33, pp. 5824–5836.

Afrose, Sharmin, Wenjia Song, Charles B Nemeroff, Chang Lu and Danfeng Daphne Yao (2021). 'Overcoming Underrepresentation in Clinical Datasets for Accurate Subpopulation-specific Prognosis'. In: *medRxiv*.

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

Gulrajani, Ishaan and David Lopez-Paz (2021). 'In Search of Lost Domain Generalization'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=lQdXeXDoWtI.

Jumper, John et al. (2021). 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873, pp. 583–589.

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

Liang, Jian, Kaixiong Gong, Shuang Li, Chi Harold Liu, Han Li, Di Liu, Guoren Wang et al. (2021). 'Pareto domain adaptation'. In: *Advances in Neural Information Processing Systems* 34, pp. 12917–12929.

Liu, Evan Z, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang and Chelsea Finn (2021). 'Just train twice: Improving group robustness without

training group information'. In: *International Conference on Machine Learning*. PMLR, pp. 6781–6792.

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

Radford, Alec et al. (2021). 'Learning transferable visual models from natural language supervision'. In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.

Ravuri, Suman et al. (2021). 'Skillful Precipitation Nowcasting using Deep Generative Models of Radar'. In: *arXiv preprint arXiv:2104.00954*.

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

Tong, Shangyuan, Timur Garipov, Yang Zhang, Shiyu Chang and Tommi S. Jaakkola (2022). 'Adversarial Support Alignment'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=26gKg6x-ie.

5 | OKAPI: GENERALISING BETTER BY MAKING STATISTICAL MATCHES MATCH

Authors:
Myles Bartlett[1], Sara Romiti[1], Viktoriia Sharmanska[1,2] & Novi Quadrianto[1,3,4]
Affiliations:
[1]Predictive Analytics Lab (PAL), University of Sussex, Brighton, UK
[2]Imperial College London
[3]BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning
[4]Monash University, Indonesia
Conference: *Neural Information Processing Systems* (NeurIPS), 2022

Abstract

We propose *Okapi*, a simple, efficient, and general method for robust semi-supervised learning based on online statistical matching. Our method uses a nearest-neighbours-based matching procedure to generate cross-domain views for a consistency loss, while eliminating statistical outliers. In order to perform the online matching in a runtime- and memory-efficient way, we draw upon the self-supervised literature and combine a memory bank with a slow-moving momentum encoder. The consistency loss is applied within the feature space, rather than on the predictive distribution, making the method agnostic to both the *modality* and the *task* in question. We experiment on the WILDS 2.0 datasets (Sagawa et al., 2022), which significantly expands the range of modalities, applications, and shifts available for studying and benchmarking real-world unsupervised adaptation. Contrary to Sagawa et al. (2022), we show that it is in fact possible to leverage additional unlabelled data to improve upon empirical risk minimisation (ERM) results with the right method. Our method outperforms the baseline methods in terms of out-of-distribution (OOD) generalisation on the iWildCam (a multi-class classification task) and PovertyMap (a regression task) image datasets as well as the CivilComments (a binary classification task) text dataset. Furthermore, from a qualitative perspective, we show the matches obtained from the learned encoder are strongly semantically related. Code for our paper is publicly available at `https://github.com/wearepal/okapi/`.

5.1 Introduction

Machine learning models have been deployed for safety-critical applications such as disease diagnosis (Watson et al., 2019) and self-driving cars (Yu et al., 2020), and in socially-important contexts such as the allocation of healthcare, education, and credit (e.g. Hurley and Adebayo, 2017; Dunnmon et al., 2019). Many machine learning algorithms, however, rely on supervision

from a large amount of labelled data, and are typically trained to exploit complex relationships and distant correlations present in the training dataset. This strategy has proven to be effective in the setting when we have training (source) and test (target) data that are i.i.d..

In reality, machine learning models are often deployed on target data whose distribution is different from the source distribution they were trained on. For example, in the task of classifying animal species in a camera trap image, one aims to learn a model that can generalise to new camera trap locations despite variations in illumination, background, and label frequencies, given training examples from a limited set of camera trap locations. Exploiting correlations that only hold in these limited locations but not in the new locations can hurt OOD generalisation. While we only have a small subset of camera traps that have their images labelled, we have a large amount of unlabelled data from the other camera traps that capture diverse operating conditions. In general, unlabelled data is much more readily available than labelled data and can often be obtained from distributions beyond the source distribution. Taking advantage of these unlabelled data during training is a key element to build robust models that have good OOD performance without sacrificing in-distribution (ID) performance.

Our work is a direct response to the empirical conclusions of Sagawa et al. (2022) for the WILDS 2.0 dataset, which extends the WILDS benchmark datasets of Koh et al. (2021) through the addition of unlabelled data. In Sagawa et al. (2022) a variety of state-of-the-art methods leveraging unlabelled data, including domain-invariant, self-training, and self-supervised methods were evaluated for their ability to improve OOD generalisation. In the all but a few cases, however, these methods failed to outperform the combination of effective data-augmentation and standard empirical risk minimisation (ERM), and among those select cases none persisted across datasets.

We show that it is possible to make effective use of large volumes of unlabelled data as supplement to a smaller set of labelled data, from a limited set of domains, to achieve strong generalisation to data from domains outside the training distribution. We turn to a statistical matching (SM) framework Rubin, 1973; Rosenbaum and Rubin, 1985; Romiti et al., 2022, a model-based approach for providing joint information on variables and indicators collected through multiple sources. SM has been widely utilised to assess the effect of interventions in numerous fields, such as education, medical and community policies (e.g. Biglan et al., 2000; Christian et al., 2010). In SM, intervened units are paired with control units and those units without a sufficiently-good match according to a given statistical criteria are excluded when estimating the treatment effect. In the running example of animal-species classification, intervened units may correspond to the limited set of camera trap locations that are fully-annotated, while control units refer to the many more camera trap locations that are only partially annotated. Pairing is beneficial for capturing diverse operating conditions, yet the ability to drop unpaired units is crucial for mitigating the risk of statistically-poor matches corrupting the training signal.

By developing an online method for statistically matching samples from different domains (camera-trap locations) and using this to define a consistency loss, we arrive at our proposed semi-supervised method, *Okapi*. This consistency loss is predicated on the simple idea of pulling together similar samples from different domains within the latent space of the encoder, and using this to bootstrap said encoder such that the distributions become progressively more aligned over the course of training. Since matching samples using the full dataset at each step of training is computationally infeasible, we instead approximate it using a combination of momentum-

encoding and a memory-bank that has been well-proven in self-supervised learning (He et al., 2020; Koohpayegani et al., 2021). Compared with other consistency-based methods such as FixMatch (Sohn et al., 2020), Okapi has the advantage of being agnostic to both the task and the modality, in addition to being distributionally robust. Contrary, to Sagawa et al. (2022), we show that the supplementary unlabelled data and domain information can be leveraged by Okapi to improve upon standard ERM on datasets from the WILDS 2.0 benchmark.

5.2 Preliminaries

5.2.1 *Problem setting*

In the standard supervised setting, one is given a dataset, $\mathcal{D}_l \triangleq \{x_i, y_i\}_{i=1}^{N_l}$, and trains a model, parameterised by θ, to well-approximate the empirical distribution as $p_\theta(y|x)$. Labelled data is limited by the cost of annotation yet one often has access to a far larger corpus of unlabelled data, $\mathcal{D}_u \triangleq \{x_i\}_{i=1}^{N_u}$, which can be used to supplement $\mathcal{D}_l$. Semi-supervised learning (SemiSL) is motivated by the idea that this additional data can often be used to improve the ID and/or OOD performance of $p_\theta(y|x)$. We can view unsupervised domain adaptation (UDA) as a special case of SemiSL, where there is assumed to be some distribution shift (adverse to a naïvely-trained predictor) between $\mathcal{D}_l$ and $\mathcal{D}_u$. Here, $\mathcal{D}_u$ comes from the domain on which $p_\theta(y|x)$ is to be evaluated, such that we have $\mathcal{D}_u \triangleq \mathcal{D}_{\text{OOD}}$, where $\mathcal{D}_{\text{OOD}}$ denotes the target domain, that is OOD w.r.t. $\mathcal{D}_l$. In the most general sense, a *domain*, or *environment* (Arjovsky et al., 2019; Creager et al., 2021) describes some partitioning of the data according to its source or some secondary characteristic, such as time of day, weather, location, lighting, or the model of the device used to collect said data; one would hope that a predictor trained under one set of conditions (e.g. day) would perform with minimal degradation under another set of conditions (e.g. night) when those conditions are irrelevant to the task at hand.

Assuming the data follows the conditional generative distribution $x \sim p(x|s)$, where s is the domain label, one would ideally use $\mathcal{D}_{\text{OOD}}$ to learn invariance to the marginal distribution, $p(s)$, and thereby achieve the equivalence $p_\theta(y|x) = p_\theta(y|x, s)$. In practice, one typically does not have access to $\mathcal{D}_{\text{OOD}}$ but does have access to training data sourced from a mixture of domains which can be leveraged to learn a more general invariance that extends to those domains outside the training distribution (Arjovsky et al., 2019). Such a learning paradigm is referred to as domain generalisation (DG). While some DG works consider the more extreme case of s being unobserved (Creager et al., 2021), we follow the more conventional setup (Arjovsky et al., 2019; Krueger et al., 2021; Sagawa et al., 2022) in which the domain(s) associated with each sample (labelled and unlabelled) is indicated by the discrete label (set of labels) s. We denote the set of possible domains for the ID labelled and unlabelled data, as $\mathcal{S}_l$ and $\mathcal{S}_u$, respectively, and their union as $\mathcal{S} \triangleq S_l \cup S_u$. Following the setup established in Sagawa et al. (2022), $\mathcal{D}_u$ is assumed to be unlabelled only w.r.t. the targets and not w.r.t. the domain labels and thus that both $\mathcal{D}_l$ and $\mathcal{D}_u$ can be augmented with the latter to give the re-definitions $\mathcal{D}_l \triangleq \{x_i, y_i, s_i\}_{i=1}^{N_l}$ and $\mathcal{D}_u \triangleq \{x_i, s_i\}_{i=1}^{N_u}$.

5.2.2 *Statistical matching*

Statistical matching is a sampling strategy which aims to balance the distribution of the observed covariates in the *treated* and *control* groups. In general terms, observed covariates x are measured characteristics of the samples; in our work we refer to the encodings generated by a deep neural network as covariates instead of the original characteristics. The treated and control groups are two partitions of the data; specifically, the treated group is the set of samples having a specific value of a variable of interest (here, the domain indicator, s) and the control group is its complement.

In this work we utilise Nearest Neighbour (NN) matching, a distance-based matching method that pairs sample i of the treated group with the closest sample j belonging to the control group. A distance measure is used to define how close two samples, i and j, are, with *propensity score distance* (PSD) and *Euclidean distance* being two widely-used distances that we employ here – indirectly (as a means of filtering) and directly, respectively.

The propensity score distance is defined as the difference between propensity scores, e_i and e_j, of samples i and j, i.e. $\text{PSD}(e_i, e_j) \triangleq |e_i - e_j|$. In causal inference, the propensity score refers to the probability of sample i belonging to the treated group, given its covariates x_i (Rosenbaum and Rubin, 1983); in practice, this conditional probability is rarely known a priori and thus requires estimation, typically via logistic regression (Stuart, 2010). We generalise the notion of a propensity score to categorical domains simply by modelling the conditional probability for each domain, with e_i instead a $|\mathcal{S}|$-dimensional probability vector. The Euclidean-distance approach, in contrast, computes the distance between the covariates, x_i and x_j, of a given pair of samples. Despite PSD being the more prevalent of the two distances, it is ill-suited to cases in where pairs are close in value w.r.t. all covariates and in such cases Euclidean distance should be preferred (King and Nielsen, 2019). Nevertheless, propensity scores remain a relevant component of NN-based matching for defining *calipers* that can reduce the likelihood of false-positive matches.

We make use of two types of caliper, *fixed* and *standard deviation*. The fixed caliper (Crump et al., 2009), t_f, defines a region of common support between the estimated propensity score distribution of the two groups; only those samples within the feasible region are admissible for matching. For binary problems, the feasible region is symmetric such that we have $\{i \,|\, e_i \in (1 - t_f, t_f)\}$ whereas in the more general, categorical case the constraint is one-sided, i.e. $\{i \,|\, \|e_i\|_\infty < t_f)\}$. This selection rule helps by removing samples with extreme propensity scores. Rosenbaum and Rubin (1985) defines the maximum discrepancy permitted between paired two samples. The discrepancy is usually expressed in terms of estimated PSD as $|e_i - e_j| < \sigma \cdot t_\sigma$, where σ denotes the mean of the group-wise standard deviations of the propensity scores and t_σ controls the percentage bias-reduction of the covariates. In the categorical case, we can simply substitute the absolute value for the infinity norm: $\|e_i - e_j\|_\infty < \sigma \cdot t_\sigma$. In the following section, we describe how one can leverage this matching framework to define a consistency loss encouraging inter-domain robustness.

5.3 Method

Here, we introduce *Okapi*, a simple, efficient, and general (in the sense that it is applicable to any task *or* modality) method for robust semi-supervised learning based on online statistical matching. Our method belongs to the broad family of consistency-based methods, characterised by methods such as FixMatch (Sohn et al., 2020), where the idea is to enforce similarity between a model's outputs for two views of an unlabelled image. These semi-supervised approaches based on minimising the discrepancy between two views of a given sample are closely related with self-supervised methods based on instance discrimination (Chen et al., 2020) and self-distillation (Grill et al., 2020; Caron et al., 2021; Baevski et al., 2022). Many of the methods within this family, however, are limited in applicability due to their dependence on modality-specific transformations and only recently has research into self-supervision sought to redress this problem with modality-agnostic alternatives such as MixUp (Verma et al., 2021), masking (Baevski et al., 2022), and nearest-neighbours (Dwibedi et al., 2021; Koohpayegani et al., 2021; Van Gansbeke et al., 2021). Approaches such as FixMatch, AlphaMatch (Gong et al., 2021) and CSSL (Lienen and Hüllermeier, 2021) that enforce consistency between the *predictive* distributions suffer further from not being directly generalisable to tasks other than classification. Okapi addresses both of these aforementioned issues through 1) the use of a statistical matching procedure – that we call CaliperNN and detail in §5.3.2 – to generate multiple views for a given sample; 2) enforcing consistency between encodings rather than between predictive distributions.

We show that models trained to maximise the similarity between the encoding of a given sample and those of its CaliperNN-generated match are significantly more robust to real-world distribution shifts than the baseline methods, while having the advantage of being both computationally efficient and agnostic to the modality and task in question. Qualitatively speaking, we see that matches produced with the final model are related in semantically-meaningful ways. Furthermore, since the only constraint is that samples be from different domains, the method is applicable whether information about the domain is coarse or fine-grained.

In the following subsections, we begin by giving a general formulation of our proposed semi-supervised loss employing a generic cross-domain k-NN algorithm. We then explain how we can replace this algorithm with CaliperNN in order to mitigate the risk of poorly-matched samples, and how the loss may be computed in an online fashion to give our complete algorithm.

5.3.1 *Enforcing consistency between cross-domain pairs*

We consider our predictor as being composed of an encoder (or *backbone*) network, $f_\theta : \mathcal{X} \to \mathbb{R}^d$, generating intermediary outputs (features) $z \triangleq f_\theta(x)$, and a prediction head, g_ϕ, such that the prediction for sample x is given by $\hat{y} \triangleq g_\phi \circ f_\theta(x)$. We similarly consider the aggregate loss $\mathcal{L}$ as having a two-part decomposition given by

$$\mathcal{L} \triangleq \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}}, \tag{5.1}$$

where $\mathcal{L}_{\text{sup}}$ is the supervised component measuring the discrepancy (as computed, for example, by the MSE loss) between $\hat{y}$ and the ground-truth label y, $\mathcal{L}_{\text{unsup}}$ is the unsupervised component

based on some kind of pretext task, such as cross-view consistency, and λ is a positive pre-factor determining the trade-off between the two components. For our method, we do not assume any particular form for $\mathcal{L}_{\text{sup}}$ and focus solely on $\mathcal{L}_{\text{unsup}}$.

Given a pair of datasets $\mathcal{D}_l$ and $\mathcal{D}_u$, sourced from the labelled domain $\mathcal{S}_l$, and unlabelled domain $\mathcal{S}_u$ respectively, along with their union $\mathcal{D} \triangleq \mathcal{D}_l \cup \mathcal{D}_u$ our goal is to train a predictor that is robust (invariant) to changes in domain, including those unseen during training. To do this, we propose to regularise $z \triangleq f_\theta(x)$ to be smooth (consistent) within local, cross-domain neighbourhoods. At a high-level, for any given *query* sample x_q sourced from domain s_q, we compute $\mathcal{L}_{\text{unsup}}$ as the mean distance between its encoding $z_q \triangleq f_\theta(x_q)$ and that of its k-nearest neighbours, $V_k(z_q)$ with the constraint that $\{s_q\} \cap \mathbf{s}_n = \emptyset$, where $\mathbf{s}_n$ is the set of domain-labels associated with $V_k(z_q)$. The general form of this loss for a given sample can then be written as

$$V_k(z_q) \triangleq \text{NN}(z_q, \{f_\theta(x) \mid (x, s) \in \mathcal{D}, s \neq s_q\}, k), \tag{5.2}$$

$$\mathcal{L}_{\text{unsup}} \triangleq \frac{1}{k} \sum_{z_n \in V_k(z_q)} d(z_q, z_n) \tag{5.3}$$

where $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is some distance function. Here, we follow Grill et al. (2020) and define d to be the squared Euclidean distance between normalised encodings for our experiments. Allowing the NN algorithm to select pairs in an unconstrained manner, given the pool of queries and keys, however, can lead to poorly-matched pairs that are detrimental to the optimisation process. To address this, we replace the standard NN algorithm with a propensity-score-based variant, inspired by the statistical matching framework (Rosenbaum and Rubin, 1983).

5.3.2 *Cross-domain matching*

For the matching component of our algorithm, we propose to use a variant of k-NN which, in addition to incorporating the above cross-domain constraint, filters the queries and keys that represent probable outliers, according to their learned propensity scores. The initial stage of filtering employs a fixed caliper, where samples with propensity scores surpassing a fixed confidence threshold are removed; this is followed by a second stage of filtering wherein any two samples (from different domains) can only be matched if the Euclidean distance between their respective propensity scores is below a pre-defined threshold (std-caliper). See Fig 5.1 for a pictorial representation of these steps and Appendix 5.7.7 for reference pseudocode.

The propensity score, e, for a given sample x is estimated as $p(s|z)$ using a linear classifier f_θ, $h_\psi : \mathbb{R}^d \to \triangle^{|\mathcal{S}|}$ where $\triangle^{|\mathcal{S}|}$ is the probability simplex over possible domain labels, $\mathcal{S}$. h^d_ψ is trained via maximum (weighted) likelihood to predict the domain label of a given sample for all samples within the aggregate dataset $\mathcal{D}$, or (typically) a subset of it, encoded by f_θ. Since we apply both calipers to the learned propensity score, the shape of this distribution can have a significant effect on the outcome of matching. Accordingly, we apply temperature-scaling, with scalar $\tau \in \mathbb{R}^+_*$, to the logits of h_ψ (where $\triangle^{|\mathcal{S}|}$ is induced by the softmax function) to modulate the resulting propensity-score distribution. We denote the set of associated parameters ($\{t_f, t_\sigma, \tau\}$, as the threshold for the fixed-caliper, the threshold for the std-caliper, and the temperature,

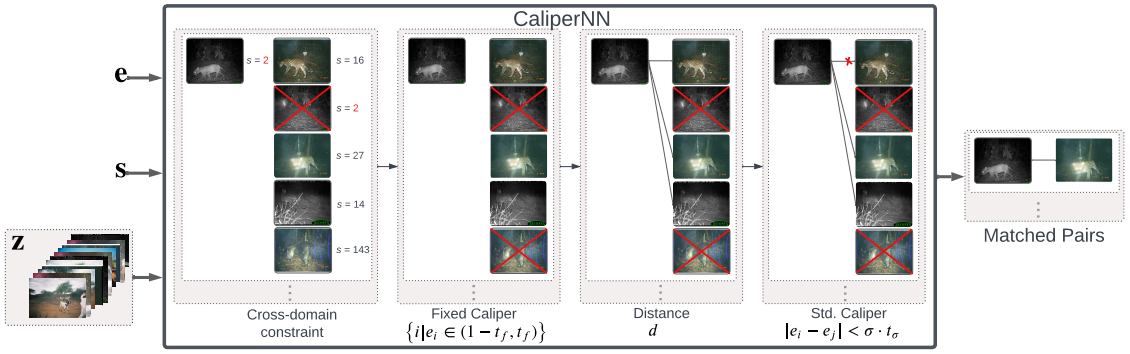


Figure 5.1: Illustration of our proposed statistical matching algorithm, CaliperNN. Given the anchor image encoding $\mathbf{z}$, the corresponding domain label s, and propensity score e, CaliperNN outputs the closest samples, according to distance d, subject to their being from different domains to the anchor.

respectively) as ξ and discuss in Appendix 5.7.4 how one can determine suitable values for these in practice.

For convenience we define the set of all encodings, given by f_θ, as $\mathbf{z} \triangleq \{f_\theta(x) \mid x \in \mathcal{D}\}$, the set of all associated propensity scores as $\mathbf{e} \triangleq \{h_\psi(x) \mid z \in \mathbf{z}\}$, and the set of associated domain labels as $\mathbf{s}$. In the offline case, the matches for $\mathcal{D}$ are then computed as

$$\text{MatchedSamples} \triangleq \{(z, \text{CaliperNN}_\xi(z, \mathbf{z}, \mathbf{e}, s, k)) \mid z \in \mathbf{z}, s \in \mathbf{s}\}, \tag{5.4}$$

with CaliperNN returning the set of k-nearest neighbours according to d, subject to the aforementioned cross-domain and caliper-based constraints. We allow for the fact that there may be no valid matches for some samples due to these constraints; in such cases we have $\emptyset$ as the second element of their tuples, indicating that $\mathcal{L}_{\text{unsup}}$ should be set to 0.

5.3.3 *Scaling up with Online Learning*

Re-encoding the dataset following each update of the feature-extractor, in order to recompute MatchedSamples, is prohibitively expensive, with cost scaling linearly with $N \triangleq N_l + N_u$. Moreover, CaliperNN requires explicit computation of the pairwise distance matrices, which can be prohibitive memory-wise for large values of N. We address these problems using a fixed-size memory bank, $\mathcal{M}_z^{N_\mathcal{M}}$ storing only the last $N_\mathcal{M}$ (where $N_\mathcal{M} \ll N$) encodings from a slow-moving momentum encoder (Grill et al., 2020; He et al., 2020), $f_{\theta'}$, which we refer to as the *target* encoder, in line with Grill et al. (2020), and accordingly refer to f_θ as the *online* encoder. Unlike Grill et al. (2020), however, we make use of neither a projector nor a predictor head (in the case of the target encoder) in order to compute the inputs to the consistency loss and simply use the output of the backbone as is – this is possible in our setting due to $\mathcal{L}_{\text{sup}}$ preventing representational collapse. More specifically, the target encoder's parameters, θ', are computed as a moving average of the online encoder's, θ, with decay rate $\zeta \in (0, 1)$, per the recurrence relation

$$\theta'_t = \zeta\theta'_{t-1} + (1 - \zeta)\theta_t, \tag{5.5}$$

As the associated domain labels are also needed both for matching and to compute the loss for the propensity scorer, we also store the labels associated with $\mathcal{M}_z^{N_\mathcal{M}}$ in a companion memory bank $\mathcal{M}_s^{N_\mathcal{M}}$. We initialise $\mathcal{M}_z^{N_\mathcal{M}}$ and $\mathcal{M}_s^{N_\mathcal{M}}$ to $\emptyset$, resulting in fewer than $N_\mathcal{M}$ samples being used during the initial stages of training when the memory banks are yet to be populated.

Each iteration of training, we sample a batch of size B from $\mathcal{D}$ consisting of inputs $\mathbf{x}$ and $\mathbf{s}$. During the matching phase, the inputs are passed through the *target* encoder to obtain $\mathbf{z}'_q \triangleq \{f_{\theta'}(x)|x \in \mathbf{x}\}$, serving as the queries for CaliperNN. We also experiment with a simpler variant where the *online* encoder is instead used for this query-generation step, such that we instead have $\mathbf{z}'_q \triangleq \{f_\theta(x)|x \in \mathbf{x}\}$, and find this can work equally well if ζ is sufficiently high. The keys are then formed by combining the current queries with the past queries contained in the memory bank: $\mathbf{z}_k \triangleq \mathbf{z}'_q \cup \mathcal{M}_z^{N_\mathcal{M}}$. The domain labels associated with $\mathbf{z}_k$ are likewise formed by concatenating the domain labels in the current batch with those stored in $\mathcal{M}_s^{N_\mathcal{M}}$: $\mathbf{s}_k \triangleq \mathbf{s}_q \cup \mathcal{M}_s^{N_\mathcal{M}}$. Once the matches for the current samples have been computed, the oldest B samples in $M_z^{N_\mathcal{M}}$ and $M_s^{N_\mathcal{M}}$ are overwritten with $\mathbf{z}_k$ and $\mathbf{s}_k$, respectively. The consistency loss is then enforced between each query $\mathbf{z}_q \triangleq \{f_\theta(x)|x \in \mathbf{x}\}$, according to the differentiable *online* encoder, and each of its matches, $V_k(z'_q) \triangleq \text{CaliperNN}_\xi(z_q, \mathbf{z}_k, h_\psi(\mathbf{z}_k), \mathbf{s}_k)$ providing that $V_k(z'_q) \neq \emptyset$ (that is, under the condition that the estimated propensity score for z'_q does not violate the caliper(s) and there are at least k valid matches whose estimated propensity scores also do not), with the loss simply 0 otherwise. Since $f_{\theta'}$ is frozen, $\mathbf{z}_k$ carries an implicit stop-gradient and gradients are computed only w.r.t. θ. These steps are illustrated pictorially in Fig 5.2 and as pseudocode in Appendix 5.7.7.

Similarly, rather than solving for the optimal parameters, $\psi^\star$ for the propensity scorer given the current values of $\mathbf{z}_k$, which is infeasible for the large values of $N_\mathcal{M}$ needed to well-approximate the full dataset, we resort to a biased estimate of $\psi^\star$. Namely, we train h_ψ in an online fashion to minimise the per-batch loss

$$\mathcal{L}_{\text{ps}} = \frac{1}{|\mathbf{z}_k|} \sum_{z \in \mathbf{z}_k, s \in \mathbf{s}_k} w_{\mathbf{s}_k}(s) \mathcal{H}(h_\psi(z), s), \tag{5.6}$$

where $\mathcal{H}$ is the standard cross-entropy loss between the predictive distribution and the (degenerate) ground-truth distribution, given by the one-hot encoded domain labels, and $w_{\mathbf{s}_k} : \mathcal{S} \to \mathbb{R}_*^+$ is a function assigning to each s an importance weight (Shimodaira, 2000) based on the inverse of its frequency in $\mathbf{s}_k$ to counteract label imbalance. In the special case in which the $\mathcal{D}_l$ and $\mathcal{D}_u$ are known to have disjoint support over S (that is, $\mathcal{S}_l \cap \mathcal{S}_u = \varnothing$), we can substitute their domain labels with 1 and 0, respectively (such that we have $\mathcal{D}_l \triangleq \{x_i, y_i, 1\}_{i=1}^{N_l}$ and $\mathcal{D}_u \triangleq \{x_i, 0\}_{i=1}^{N_u}$), thus reducing the propensity scorer and CaliperNN to their binary forms. Knowing whether this condition is satisfied a priori (and thus whether the use of domain labels can be forgone completely from our pipeline) is not unrealistic: one may, for example know that two sets of satellite imagery cover two different parts of the world (e.g. Africa and Asia) yet not know the exact coördinates underlying their respective coverage.

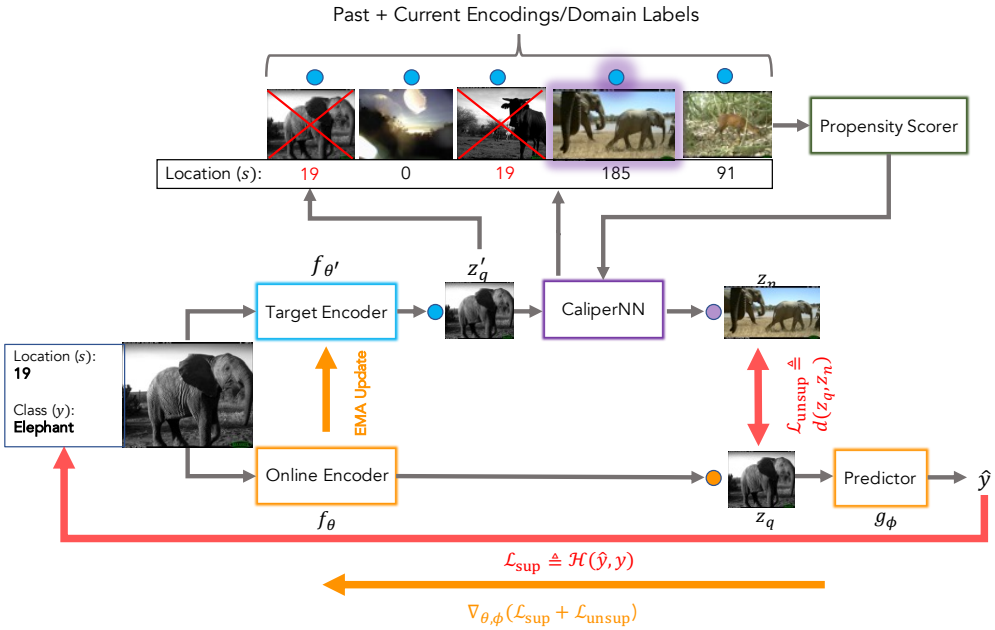


Figure 5.2: Overview of Okapi's online-learning pipeline based using the iWildCam dataset for the sake of illustration. For simplicity, we limit k to 1 so that the output of matching is a single vector rather than a set of vectors; for the same reason we illustrate the process for only a single sample taken from the labelled data set $\mathcal{D}_l$, annotated with both domain (s; in this case, *camera location*) and class (y) information. Inspired by recent advances in self-supervised learning, we maintain a copy (the target encoder) of the online encoder, f_θ, whose parameters, θ', are an exponential moving average (EMA) of θ. This EMA update is performed at the beginning of each training set at a rate governed by the decay coefficient, ζ. For a given sample, we first compute its embedding using the target encoder to produce the query vector, z'_q, and by the online encoder to produce z_q, which will serve as the 'anchor' in the consistency loss. This query vector is then used – alongside the output of the propensity scorer – by CaliperNN to compute its cross-domain nearest neighbour, z_n, where the keys are taken to be the current and past (stored in the Memory Bank) $N_{\mathcal{M}}$ encodings of the data. The cross-domain constraint, prohibiting matching of samples belonging to the same domain, is denoted through a red colouring of the location identifiers, the nearest sample obeying this constraint and the constraints of the calipers with purple highlighting. The consistency loss is the distance between z_q and z_n, defined by function some distance function d. Finally, the supervised loss, $\mathcal{L}_{\text{sup}}$ (here instantiated as the standard cross-entropy loss, $\mathcal{H}$), is computed using the output of the predictor acting on z_q and the ground-truth given by y.

5.4 Related Work

Domain Generalisation. The goal of domain generalisation (DG) is to produce models that are robust to a wide range of distribution shifts (including those outside the training distribution), given a training set consisting of samples sourced from multiple domains. Despite the various techniques (many well theoretically-motivated) designed to improve the generalisation of deep neural networks current methods continue to fall short in the face of natural distribution shifts (Gulrajani and Lopez-Paz, 2020; Koh et al., 2021). Indeed, ERM has repeatedly shown to be a strong baseline – frequently outperforming dedicated methods that leverage domain information or additional unlabelled data – for DG (Gulrajani and Lopez-Paz, 2020; Sagawa et al., 2022), despite the theoretical problems associated with using it when the training and test sets are misaligned. Until now, only pre-training on larger, more diverse datasets (with harder examples), has consistently proven to improve OOD generalisation, yet allowing pre-trained models to fit the ID data too closely can undo any such benefit conferred by the pre-training (Taori et al., 2020;

Andreassen et al., 2021; Kim et al., 2022; Wiles et al., 2022). Similar to Okapi, MatchDG (Mahajan et al., 2021) draws upon causal matching to tackle DG. Despite the surface-level similarity, there are a number of significant differences, principally in the respects that we consider semi-supervised DG (whereas MatchDG requires full-labelling w.r.t. y) and employ an augmented form of k-NN for bias-reduction in the absence of y.

SELF-SUPERVISED LEARNING. In self-supervised learning (SelfSL), models are trained to solve pretext tasks constructed from the input data. This learning paradigm has led to significant breakthroughs in unsupervised learning in recent years, with performance now approaching (or even surpassing, along some axes such as adversarial robustness) that of supervised methods for many tasks while requiring significantly less labelled data. Due to its generality, SelfSL has seen use across the complete spectrum of applications and modalities and underlies many of the foundation models (Bommasani et al., 2021) that have emerged in NLP (Devlin et al., 2018; Brown et al., 2020; Chowdhery et al., 2022), Computer Vision (Goyal et al., 2022), and at their intersection (Alayrac et al., 2022; Yu et al., 2022). Common pretext tasks include those based on the masked-language-modelling approach – originally popularised by BERT (Devlin et al., 2018) and recently generalised to other modalities (Bao et al., 2021; Baevski et al., 2022) – (Chen et al., 2020; He et al., 2020), contrastive captioning (Radford et al., 2021; Yu et al., 2022), and instance discrimination and self-distillation (Grill et al., 2020; Caron et al., 2021) which rely on transformations of the data to generate multi-view inputs. Approaches belonging to the latter two categories were originally limited by the fact that the transforms had to be tailored for a particular modality and for some modalities, such as tabular data, there is no obvious way to define them. A number of recent works have sought to obviate this problem through the use of MixUp (Verma et al., 2021), masking (He et al., 2021; Baevski et al., 2022), and k-NN (Dwibedi et al., 2021; Koohpayegani et al., 2021; Van Gansbeke et al., 2021), the latter of which is directly relevant to our work. Okapi bears closest resemblance to Koohpayegani et al. (2021) in combining momentum-encoding with nearest-neighbours lookup to generate the views for a BYOL-style (Grill et al., 2020) consistency loss. However, a key distinction lies in the use of an augmented form of nearest-neighbours, CaliperNN, which both constrains pairs of samples to being from *different* domains and filters out any queries or keys deemed outliers according to a learned *propensity score*.

SEMI-SUPERVISED LEARNING. Semi-supervised learning (SemiSL) encompasses a broad class of algorithms that combine unsupervised learning with supervised learning in order to improve the performance of the latter, especially when labelled data is limited. Many SemiSL methods are based on the self-training paradigm which can trace its roots back decades to the early work in pattern recognition by Scudder (1965) and continues to be relevant in the modern era due to its generality, both within SemiSL itself and in related fields such as domain adaptation (Ganin et al., 2016), and fledgling field of SelfSL (Caron et al., 2021) discussed above. Self-training applies to any framework predicated on using a model's own predictions to produce pseudo-labels for the unlabelled data which can either be used as targets for self-distillation (Xie et al., 2020) or enforcing consistency between predictions that themselves have been perturbed (Bachman et al., 2014; Xie et al., 2020) or that have been generated from perturbed/multi-view inputs (Sohn et al., 2020).

Table 5.1: A comparison between Okapi and different baselines on two benchmark image datasets. We include both the results of our re-run of the baselines and those of Sagawa et al. (2022). Both ID and OOD performances are reported. For iWildCam we average over results from 3 different seeds, for PovertyMap we do so over the 5 pre-defined CV folds. Standard deviations are shown in parentheses.

Method	**iWildCam**		**PovertyMap**			
	macro F1 ↑		worst U/R corr. ↑		worst U/R MSE ↓	
	ID	OOD	ID	OOD	ID	OOD
ERM (Sagawa et al., 2022)	47.0 (1.4)	32.2 (1.2)	0.66 (0.04)	0.49 (0.06)	-	-
FixMatch (Sagawa et al., 2022)	46.3 (0.5)	31.0 (1.3)	0.54 (0.10)	0.30 (0.11)	-	-
ERM	48.6 (1.1)	33.3 (0.3)	0.72 (0.03)	0.53 (0.09)	0.23 (0.03)	0.35 (0.12)
FixMatch	51.1 (1.0)	35.2 (0.7)	0.50 (0.13)	0.34 (0.12)	0.59 (0.42)	0.88 (0.61)
Okapi (ours)	50.6 (0.7)	36.1 (0.9)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)
Okapi (no calipers)	-	-	0.72 (0.02)	0.54 (0.12)	0.22 (0.02)	0.36 (0.14)

FixMatch (Sohn et al., 2020) is one example of a consistency-based method which has proven effective for semi-supervised classification, despite its simplicity, and various works (Gong et al., 2021; Lienen and Hüllermeier, 2021) have since built on the its framework prescribing the use of weakly- and strongly-augmented inputs to generate the targets and predictions, respectively. Like these methods, Okapi also makes use of a cross-view consistency loss, however, the alternative views for a given sample are generated not through data-augmentation but through statistical matching (Rosenbaum and Rubin, 1983), with the aim being to achieve invariance to the domain rather than a particular series of perturbations. Another example of particular relevance to our work is Tarvainen and Valpola (2017), which uses a copy of the model with exponentially-averaged weights to generate the targets for the unlabelled data. Okapi also uses such a model to produce the targets for its consistency loss, but is more akin to momentum-encoding (He et al., 2020) in the respect that the loss is imposed on the latent space.

5.5 Experiments

5.5.1 *Datasets*

We evaluate Okapi on three datasets taken from the WILDS 2.0 benchmark (Sagawa et al., 2022). These span a variety of modalities and tasks, allowing us to showcase the generality of our proposed method (Okapi): **iWildCam** (images, multiclass classification), **PovertyMap** (multispectral images, regression), and **CivilComments** (text, binary classification). Details of each dataset can be found in Appendix 5.7.1.

5.5.2 *Image experiments*

Results of our image-data experiments are summarised in Table 5.1. Due to spacial constraints, we defer the full set of results, including those for the 'offline' (w.r.t. the matching) version of Okapi to Appendix 5.7.3. For both datasets in question, we use the same metrics as Sagawa et al. (2022): macro-F1 for iWildCam and worst-group (with the group defined as urban (U) vs. rural (R)) Pearson correlation for Poverty Map. For completeness, we include mean squared error (MSE)

as a secondary metric for the latter dataset. Following Sagawa et al. (2022), we compute the mean and standard deviation (shown in parentheses) over multiple runs for both ID and OOD test sets, with these runs conducted with 3 different random seeds and 5 pre-defined cross-validation folds for iWildCam and PovertyMap, respectively.

We compare Okapi against two baselines, ERM and FixMatch (Sohn et al., 2020), both according to our re-implementation and according to the original implementation given in Sagawa et al. (2022). We note that since FixMatch, in its original form, is only applicable to classification problems due to its use of confidence-based thresholding, for the PovertyMap dataset, FixMatch represents a simplified variant (following (Sagawa et al., 2022)) without such thresholding, that is trained to simply minimise the MSE between *all* regressed values for the weakly- and strongly-augmented images. As described in Appendix 5.7.4, the main difference between the baselines run included in Sagawa et al. (2022) and our re-runs is in the backbone architecture, with us opting for a ConvNeXt (Liu et al., 2022) architecture over a ResNet one. For both datasets, and for both baselines we observe significant improvements stemming the change of backbone. Moreover, utilising ConvNeXt seems to be crucial in enabling FixMatch to surpass the ERM baseline in the classification task with 32.2 (ERM) vs 31.0 (FixMatch) and 33.3 (ERM) vs. 35.2 (FixMatch), with ResNet and ConvNeXt architecture respectively.

Okapi, convincingly outperforms the baselines, w.r.t. the OOD metric of interest, on both datasets. We observe an improvement of $+0.9$ macro F1, i.e. 36.1 vs 35.2 of Okapi and FixMatch (the best baseline for iWildCam) respectively. For the regression task in PovertyMap, Okapi achieves 0.55 and 0.33 on the OOD test set in terms of Pearson correlation and MSE, respectively, in contrast to the 0.53 and 0.33 of ERM. At the same time, we note that FixMatch fails to generalise well to this task, yielding by far the worst results amongst the evaluated methods.

5.5.3 *Text classification*

Method	**Civil Comments** worst-group acc ↑
	OOD
ERM (Sagawa et al., 2022)	66.6 (1.6)
ERM (fully-labelled; Sagawa et al., 2022)	69.4 (0.6)
ERM (reproduction)	68.5 (2.2)
Okapi (ours)	69.7 (2.0)

Table 5.2: Comparison between Okapi and the baselines methods on the Civil Comments dataset. We include both the original results of Sagawa et al. (2022) as well as those of our reproduction of their ERM baseline. Performance is measured in terms of worst-group accuracy and averaged over seeds; standard deviations are shown in parentheses.

We summarise in Table 5.2 the numerical results for the CivilComments dataset. Remaining consistent with Sagawa et al. (2022), we evaluate models according to the worst-group accuracy – the minimum of the conditional accuracies obtained by conditioning on each of the 8 dimensions of s – averaged over 5 replicates. Since there is no canonical ID test split available for this dataset, we report only the results only for the OOD split that is, rather than doing so for a custom split to avoid misrepresentation. We compare Okapi against both ERM variants featured in Sagawa et al. (2022) – one trained on only the official labelled data and one trained with annotated unlabelled data (fully-labelled) – as well as our re-implementation of the ERM variant trained on only the

Figure 5.3: Examples of input (labelled) images and their 1-NN matched (unlabelled) images retrieved using CaliperNN on iWildCam dataset. Here, we match images from the labelled-train set to images from the unlabelled-extra set, taking advantage the fact that their domains are disjoint.

labelled data with an identical hyperparameter configuration to the former. In contrast to the image datasets, we do not diverge in our choice of architecture, with all models trained with a pre-trained DistilBERT (Sanh et al., 2019) backbone.

We observe marked improvement in the worst-group accuracy of this baseline compared with that reported therein. We attribute this partly to the high variance of the model-selection procedure (inherited from Sagawa et al. (2022)) based on intermittently-computed validation performance (which does not consistently align with test performance) to determine the final model. This aside, we observe that Okapi outperforms the ERM baseline by a significant margin, to the point of parity with the fully-labelled baseline.

5.5.4 *Ablations and qualitatitive analysis*

In order to evaluate the importance of the caliper-based filtering to the performance of Okapi, we perform an ablation experiment on PovertyMap dataset (Okapi (no calipers)) with said filtering disabled (and all else constant), such that instead of CaliperNN we have standard k-NN, albeit with the cross-group constraint still in place (per Eq. 5.2). We observe that performance degrades according to both metrics of interest, and, crucially, that the standard deviation of the runs is significantly higher, in line with our expectation that filtering out poor matches should stabilise optimisation. We provide additional ablation experiments in Appendix 5.7.6, exploring the relative importance of the two (fixed and std-) calipers, the optimal number of neighbours to use for computing $\mathcal{L}_{\text{unsup}}$, and the feasibility of using the online encoder to generate the queries for CaliperNN.

Finally, in Fig. 5.3 we show samples of matched pairs retrieved by CaliperNN from the encodings of the learned encoder for the iWildCam dataset. Here, we see that semantic information (encoding the species of animal) is preserved across pairs, while nuisance factors such as illumination, background and contrast vary. Further examples from PovertyMap are shown in Appendix 5.7.5. In Appendix 5.7.8, we include matching results for the PACS (photo (P), art painting (A), cartoon (C), and sketch (S)) dataset Li et al., 2017 demonstrating how temperature scaling, in conjunction with the fixed caliper, can be used to control the filtering rate.

5.6 Conclusion

In this work, we introduced, Okapi, a semi-supervised method for training distributionally-robust models that is intuitive, effective, and is applicable to any modality or task. Okapi is based on the simple idea of supplementing the supervised loss with a cross-domain consistency loss that encourages the outputs of an encoder network to be similar for neighbouring (within the latent space of the encoder itself) samples belonging to different domains, which is made efficient using an online-learning framework. Rather than simply using k-NN with a cross-domain constraint, however, we propose an augmented form based on statistical matching (CaliperNN) that combines propensity scores with calipers to winnow out low-quality matches; we find this to be important for both the end-performance and consistency of Okapi. Our work serves as a response to Sagawa et al., 2022, in that we find that it is in fact possible to effectively incorporate unlabelled data and domain information into a training algorithm in order to improve upon ERM with respect to an OOD test set, assuming an appropriate choice of architecture. Namely, on three datasets from the WILDS 2.0 benchmark, representing two different tasks (classification and regression) and modalities (image and text), we show that Okapi outperforms both the ERM and FixMatch baselines according to the relevant OOD metrics.

Buoyed by these promising results, we intend to apply Okapi to other tasks (e.g. object detection and image segmentation) and other modalities (e.g. audio) to further establish its generality. Furthermore, one limitation of the current incarnation of the method is that the thresholds for the calipers are fixed over the course of training whereas it may be beneficial to set these adaptively with the view to optimise such measures of inter-domain balance as *Variance Ratio* and *Standard Mean Differences* that are commonly used to evaluate the the goodness of statistical matching procedures.

5.7 Appendix

5.7.1 *Datasets*

We evaluate Okapi using three datasets – iWildCam, PovertyMap, and CivilComments – taken from the WILDS 2.0 benchmark (Sagawa et al., 2022). These datasets were chosen specifically due to the poor performance reported by Sagawa et al. (2022) for semi-supervised and domain adaptation methods across the board, in relation to the ERM baselines. For PovertyMap in particular, ERM was found to vastly outperform any competing methods utilising the unlabelled data and/or domain labels.

iWildCam-WILDS is an extension of the iWildCam 2020 Competition Dataset (Beery et al., 2020). The task is multi-class species classification of animals in camera trap images. The dataset contains 1022K images of animals annotated with the domain, s, that identifies the camera trap that captured it. The target label, y, is one of 182 different animal species and it is provided solely for the 203K labelled data. The labelled training set contains 130K images taken by 243 camera traps. The OOD validation and target sets include images from 32 and 48 different camera traps which are disjoint from the 243 training domains. Additionally, 819K unlabelled images from 3215 new domains are available. Different cameras trap differ in characteristics such as illumination, background and relative animal frequency, models trained on the source domains might fail to generalise to images taken from new locations.

PovertyMap-WILDS is a variation of the dataset introduced in Yeh et al. (2020). The task is to predict the wealth index, y, from multispectral satellite images of 23 African countries. The country the image was taken in as well as whether it was taken in a rural or urban area represent the domain s. The dataset contains 5 cross-validation (CV) folds of roughly equal size, each one dividing the 23 countries differently across the source, OOD validation and OOD target splits. In each fold, the labelled training set contains 11K images from 14 different countries. The OOD validation and target sets include images from 5 different countries not represented in the source data. The dataset also includes 261K unlabelled images from the same 23 countries.

CivilComments-WILDS is an online-comment dataset adapted from Borkan et al. (2019), comprising 448K online comments annotated with both a binary indicator of toxicity ($\{$*toxic*, *not toxic*$\}$) – serving as the target label, y – and the demographic identities mentioned within them – serving as the domain s. Here, $s \in \{0,1\}^8$ is a binary vector rather than a scalar, with dimensions indicating membership (non-exclusively) to 8 demographic groups, spanning different genders, religions and ethnicities. For the WILDS 2.0 variant of the dataset, Sagawa et al. (2022) introduce an additional corpus of 1551K comments acting as the unlabelled training data belonging to extra domains. While the comments are completely unlabelled, w.r.t. both y and s and thus are not domain-separable at the sample level, the majority (92%) of the comments are known to be sourced from the same documents as those comments comprising the (OOD) labelled test data. As noted in Sagawa et al. (2022), CivilComments-WILDS exhibits label imbalance w.r.t. y; this is amended both therein and herein (as appertains all methods) through the use of class-balanced sampling, though with the minor distinction that for our experiments we ensure each batch is exactly balanced rather by sampling equally from each class, in contrast to Sagawa et al. (2022)

who sample hierarchically – sampling y_i uniformly from $\{0, \ldots, |\mathcal{Y}_l|\}$ and then uniformly from $\mathcal{D}_l$, conditioned on y_i – such that balance is achieved only in expectation.

5.7.2 *Relation to Algorithmic Fairness*

DG and AF overlap in their objective to train a model that yields predictions that are statistically independent of (and thus robust to variations in) domain, when for the latter the domain is taken to be some protected characteristic, such as age or gender, and fairness is measured according to invariance-driven notions of group fairness such as Demographic Parity (Feldman et al., 2015) and Equal Opportunity (Hardt et al., 2016). Indeed, methods that focus on equalising the empirical risk across subgroups – such as by importance weighting (Shimodaira, 2000; Idrissi et al., 2022) – have featured extensively in both DG (Arjovsky et al., 2019; Sagawa et al., 2019; Creager et al., 2021; Krueger et al., 2021) and fairness (Kamiran and Calders, 2012; Agarwal et al., 2018; Donini et al., 2018) and many approaches to FRL (Madras et al., 2018; Creager et al., 2019; Quadrianto et al., 2019; Kehrenberg et al., 2020; Oneto et al., 2020) have roots in the former (Muandet et al., 2013) and in the closely-related field of DA (Ganin et al., 2016). Beyond this more general equivalence, our work also has ties to notions of IF pioneered by Dwork et al. (2012) – broadly prescribing that similar individuals be treated similarly – in that our unsupervised loss involves maximising the similarity between inter-domain samples within representation space. This is reminiscent of the operationalisation of individual fairness proposed by Lahoti et al. (2019) that enforces similarity between a given representation and the representations of its neighbouring – in both the input space and according to a between-group (cross-domain) quantile graph – samples.

5.7.3 *Extended Results*

We present in Table 5.3 an extended version of the results for the iWildCam and PovertyMap datasets, relative to those found in Table 5.1 in the main text. This table includes additional results with the ResNet backbones per Sagawa et al. (2022) – justifying our decision to adopt a ConvNext backbone for our main set of image-dataset results – as well as those for an 'offline' version of Okapi (Okapi (offline)) where the matches are generated prior to training using features of the respective ERM baseline for each dataset. Since the target encoder is necessitated by the need for online match-retrieval, only a single encoder is involved in Okapi (offline); in binary cases, the algorithm is then identical to the one proposed by Romiti et al. (2022) with the exception that consistency is still enforced via distance in encoding space rather than with a JSD loss on the predictive distributions which fails to generalise to regression tasks such as PovertyMap.

5.7.4 *Implementation details*

DATA AUGMENTATION. We follow Sagawa et al. (2022) when defining the augmentations for the the WILDS datasets. In the case of PovertyMap-WILDS we corroborate the original finding that data-augmentation adversely affects performance, and, in light of this, elect only to use data-augmentation for FixMatch where it is needed to generate the weak and strong views used

Table 5.3: An extended comparison between Okapi and different baselines on two benchmark image datasets. We include both the results of our re-run of the baselines and those of Sagawa et al. (2022). Both ID and OOD performances are reported. For iWildCam we average over results from 3 different seeds, for PovertyMap we do so over the 5 pre-defined CV folds. Standard deviations are shown in parentheses. The additions relative to Table 5.1 include results with an offline variant of Okapi – where the matches are generated prior to training from the features of the trained ERM model and then fixed for the course of training – and with the ResNet backbones employed by Sagawa et al. (2022).

Method	**iWildCam**		**PovertyMap**			
	macro F1 ↑		worst U/R corr. ↑		worst U/R MSE ↓	
	ID	OOD	ID	OOD	ID	OOD
ERM (Sagawa et al., 2022)	47.0 (1.4)	32.2 (1.2)	0.66 (0.04)	0.49 (0.06)	-	-
FixMatch (Sagawa et al., 2022)	46.3 (0.50)	31.0 (1.3)	0.54 (0.10)	0.30 (0.11)	-	-
ERM (ConvNeXt)	48.6 (1.1)	33.3 (0.3)	0.72 (0.03)	0.53 (0.09)	0.23 (0.03)	0.35 (0.12)
FixMatch (ConvNeXt)	51.1 (1.0)	35.2 (0.7)	0.50 (0.13)	0.34 (0.12)	0.59 (0.42)	0.88 (0.61)
Okapi (ours; ConvNeXt)	50.6 (0.7)	36.1 (0.9)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)
Okapi (offline; ConvNeXt)	48.8 (0.8)	31.7 (0.2)	0.68 (0.02)	0.53 (0.07)	0.26 (0.02)	0.37 (0.13)
Okapi (no calipers; ConvNeXt)	-	-	0.72 (0.02)	0.54 (0.12)	0.22 (0.02)	0.36 (0.14)
ERM (ResNet)	46.5 (0.8)	29.7 (1.0)	0.69 (0.03)	0.53 (0.08)	0.24 (0.04)	0.34 (0.11)
FixMatch (ResNet)	43.0 (2.5)	25.5 (1.4)	0.70 (0.02)	0.53 (0.08)	0.24 (0.02)	0.35 (0.10)
Okapi (ResNet)	46.1 (0.7)	27.8 (0.3)	0.70 (0.04)	0.52 (0.07)	0.23 (0.02)	0.33 (0.10)

in computing the consistency loss. Since Okapi uses an NN-based approach for generating these views, it is decoupled from the augmentation strategy and problems that can arise from its misspecification.

ARCHITECTURE. For our image experiments, contrary to Sagawa et al. (2022), we opt to use the recently proposed ConvNeXt architecture (Liu et al., 2022), finding this change to provide large performance gains and to be crucial in enabling semi-supervised methods to surpass the ERM baseline. This is in line with Kim et al. (2022) who similarly found that a change of architecture (combined with large-scale pre-training) could greatly bolster performance on the iWildCam dataset. More precisely, we use the *tiny* variant of ConvNeXt, pre-trained on ImageNet 1k, as the initial backbone for our models. We compose this with a single fully-connected layer to construct the complete predictor both for the target and the propensity score. For our CivilComments experiments, in contrast, we do not diverge from Sagawa et al. (2022) in our choice of architecture, with all models trained with a pre-trained DistilBERT (Sanh et al., 2019).

OPTIMISATION. For optimising all models, we use the AdamW optimiser (Loshchilov and Hutter, 2018) coupled with a cosine annealing schedule without warm restarts (Loshchilov and Hutter, 2017). We set the initial learning to be 1×10^{-4} across the board, and forgo the use of weight decay. Models are trained for 120K, 30K, and 20K iterations for iWildCam, PovertyMap, and CivilComments, respectively. The decay coefficient, ζ, for the target encoder's exponential moving average is initialised to ζ_{start} and is linearly increased to ζ_{end} over the course of training. For PovertyMap we set ζ_{start} and ζ_{end} to be 0.996 and 1.0, respectively; for iWildCam, we set ζ_{start} and ζ_{end} to be 0.999 and 0.999, respectively, resulting in a fixed value of ζ; 1.0, respectively; for CivilComments, we set ζ_{start} and ζ_{end} to be 0.996 and 0.996, respectively, again resulting in a fixed value of ζ. We similarly warm up the pre-factor for the consistency loss, λ, according to a linear schedule during the first 10% of training to allow a period for the encoder to learn meaningful

relations between samples through the supervised loss before bootstrapping with the consistency loss, with a final value of 1.

MATCHING. In order to determine suitable hyperparameters, ξ, for CaliperNN, we perform a grid-search in the static setting, using a fixed model. Specifically, we use the backbone of an ERM-trained model as the encoder with which to generate the queries and keys for matching. The quality of matching with a given instantiation of ξ is measured using two metrics commonly used in the statistical matching literature: *Variance Ratio* (VR) and *Standard Mean Differences* (SMD) (Rubin, 2001). Both metrics operate on pairs of domains, but can be generalised to work when s is non-binary by simply aggregating over all pairwise results. For a given pair of domains, VR is defined as the ratio of the variances of the covariates between the two domains, with an ideal value of 1, while SMD is defined as the difference in their covariate means, normalised by the standard deviation for each covariate, and is to be minimised. While our proposed method is applicable whether s is binary or categorical, for the experiments in this paper we take advantage of the fact that the WILDS datasets specify splits with non-overlapping domains and match from $\mathcal{D}_l \rightarrow \mathcal{D}_u$ and in the reverse direction (from $\mathcal{D}_u \rightarrow \mathcal{D}_l$). This decision was based on preliminary experiments which found the binary variant generally enjoyed more stable optimisation, something which future work should seek to rectify. In the case of PovertyMap, however, the training splits themselves do not satisfy the aforementioned requirement of being sourced from mutually exclusive sets of domains and we instead treat the OOD validation set as $\mathcal{D}_u$ (and treat it as being unlabelled w.r.t. y, in that it is only used for $\mathcal{L}_{\text{unsup}}$).

5.7.5 *Additional Matching Examples*

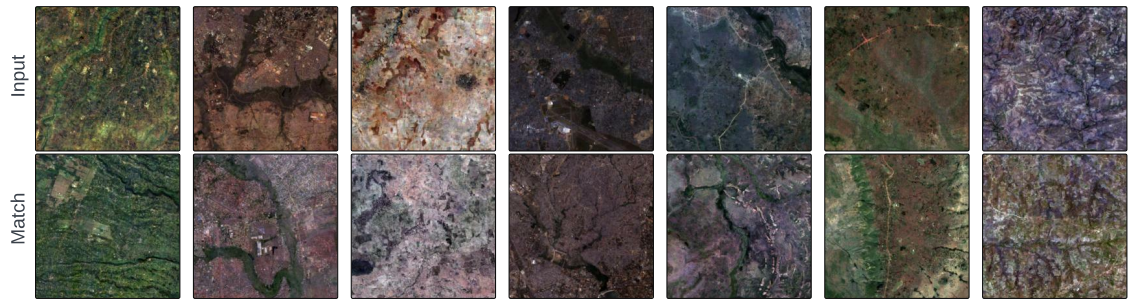


Figure 5.4: Examples of input (labelled) images and their 1-NN matched (unlabelled) images retrieved using CaliperNN from the PovertyMap-WILDS dataset. Here, we match images from the labelled-train set to images from the OOD-validation set, taking advantage the fact that their domains are disjoint.

5.7.6 *Ablations*

We supplement the ablation experiment on the use of calipers featured in the main text with additional experiments concerning effect of the number of nearest neighbours (k), the relative importance of the two (fixed and std) calipers, and the feasibility of using the online encoder as

Table 5.4: Ablation experiments for Okapi conducted using the PovertyMap-WILDS dataset. Specifically, we assess the importance of four elements of our proposed method: the number of nearest neighbours used in computing $\mathcal{L}_{\text{unsup}}$ (k), the use (enabled/disabled) of the fixed- and std-calipers in CaliperNN (considering these to be separate components), and which encoder (online or target) is used to generate the queries for statistical matching (with use of the target encoder 'TE queries' being the default and 'OE queries' denoting the alternative). Both ID and OOD performances are reported. The results are computed by aggregating over the results for each of the 5 pre-defined cross-validation folds. We report the average and standard deviation value across replicates of the metric of interest.

(a) CaliperNN ablations.

Method	**PovertyMap**			
	worst U/R corr. ↑		worst U/R MSE ↓	
	ID	OOD	ID	OOD
Okapi (k=1)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)
Okapi (k=5)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)
Okapi (k=10)	0.72 (0.02)	0.55 (0.09)	0.22 (0.02)	0.33 (0.10)
Okapi (k=5, no calipers)	0.72 (0.02)	0.54 (0.12)	0.22 (0.02)	0.36 (0.14)
Okapi (k=5, no std caliper)	0.72 (0.02)	0.54 (0.12)	0.22 (0.02)	0.35 (0.14)
Okapi (k=5, no fixed caliper)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)

(b) Query-generator ablation (target encoder (TE) vs. online encoder (OE)).

Method	**PovertyMap**			
	worst U/R corr. ↑		worst U/R MSE ↓	
	ID	OOD	ID	OOD
Okapi (TE queries)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)
Okapi (OE queries)	0.72 (0.02)	0.55 (0.10)	0.23 (0.02)	0.34 (0.10)

the query-generator instead of the target encoder. The results of these experiments are tabulated in 5.4 with the key takeaways being:

1. The number of neighbours used for computing the consistency loss has little impact – according to the given level of precision – on the performance of Okapi along all axes.

2. While disabling the calipers altogether considerably harmed performance, using only the std. caliper allows us to recover the performance of the complete algorithm, (`Okapi (k=5)`), whereas the same is not true for the fixed caliper which, while aiding performance compared to the no-caliper baseline, falls short of that benchmark. A caveat attached to these conclusions, however, is that the selected values for ξ are likely suboptimal in the online setting, given that they were optimised for the static setting: with improved selection of ξ, either by learning it jointly with the model's parameters (using, for instance, the perturbed maximum method (Berthet et al., 2020) to overcome the non-differentiability of the k-NN and thresholding operations), in an amortised fashion, or optimising it on a per-iteration basis.

3. While less appealing from a conceptual standpoint, due to the mismatch between the networks used to generate the queries and keys, from an empirical standpoint it is perfectly feasible to use the online encoder to generate the queries for statistical matching instead the

target encoder while experiencing minimal degradation in performance. This is particularly relevant when one wishes to perform the matching in only one direction (e.g. $\mathcal{D}_l \rightarrow \mathcal{D}_l$) due to the reduction in redundant encoding, with each encoder only encoding its respective subset of the data (e.g. f_θ only encodes samples from $\mathcal{D}_l$, f'_θ only encodes samples from $\mathcal{D}_u$

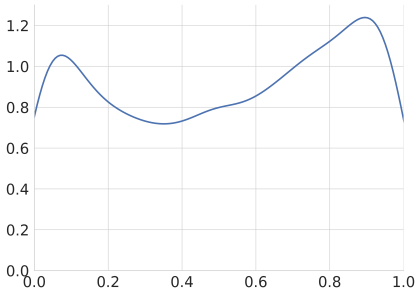
5.7.7 *Pseudocode*

We give PyTorch-style (Paszke et al., 2019) pseudocode for the CaliperNN (described in 5.3.2) and online-learning (described in 5.3.3) algorithms in Algorithm 2 and Algorithm 3, respectively. In both cases, we restrict the pseudocode to the special case of binary domains – practically achieved by using the labelled/unlabelled as a proxy for domain – for ease of illustration. The CaliperNN algorithm can be generalised freely to multiclass cases by considering pairwise interactions between the propensity scores for each domain for applying the calipers and by computing the pairwise inequalities between $\mathbf{s}_q$ and $\mathbf{s}_k$ (giving the connectivity matrix $(\mathbf{s}_q \cdot \mathbf{1}^T) \neq (\mathbf{1} \cdot \mathbf{s}_k^T)$, where $\mathbf{1}$ denotes the ones vector of the same shape as its multiplicand and mediates broadcasting) for enforcing the cross-domain constraint.

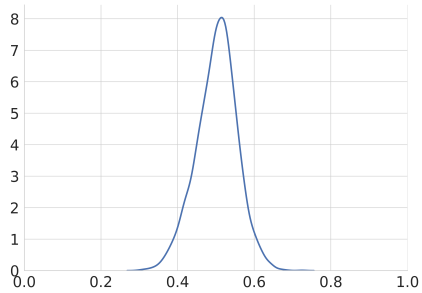
5.7.8 *Matching for PACS dataset*

In this section, we discuss results from initial experiments on the PACS dataset (Li et al., 2017) (using features extracted for a pre-trained CLIP (Radford et al., 2021) visual encoder) showing how the temperature scaling can be used to smooth the propensity score distribution to better control how many sample are discarded during matching. There are 1,670 *photo*, 2,048 *art painting*, 2,344 *cartoon*, and 3,929 *sketch* in the dataset. Here will evaluate the results of matching across the two domains *photo* and *art painting* as well as across *photo* and *sketch*. In Fig. 5.5 and Fig. 5.6 we compare the shape of the estimated propensity score with its scaled version using a temperature value of 10. As we can see, in the case of a distribution with extremely heavy tails (photo, sketch), the effect of smoothing the distribution is that when a fixed caliper is applied most of the samples are retained. On the other hand, when the initial distribution is smoother, a temperature of 10 is extreme, having the effect of transforming the bimodal distribution to a unimodal one. Additionally, we tabulate in Table 5.5 the number of matched pairs retrieved when matching across the two domains photo and sketch; here we can see that by increasing the temperature we smooth the estimated propensity score distribution and thereby retain more samples. Similarly, we can retrieve more pairs by reducing the fixed caliper threshold. We also analyse the case of matching across the two domains photo and art painting. Using a fixed caliper defined defined by a threshold $t_f = 0.1$ and no temperature scaling (i.e. $\tau = 1$) the algorithm retrieves 1,142 pairs matching in the direction photo $\rightarrow$ art and 1,501 in the direction art $\rightarrow$ photo.

In Fig. 5.7 and Fig. 5.8 we show examples of matching pairs found using our CaliperNN algorithm. Although the features were not fine-tuned on PACS, we can see a few examples of intraclass matching. For the photo-art painting application we can see preservation in colour and background; while in the photo-sketch case shape and pose.

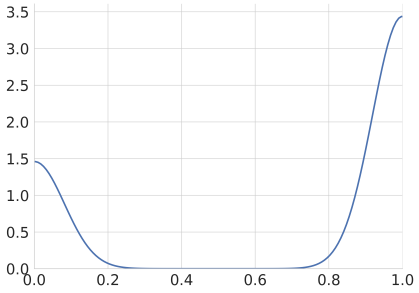


(a) No temperature scaling ($\tau = 1$).

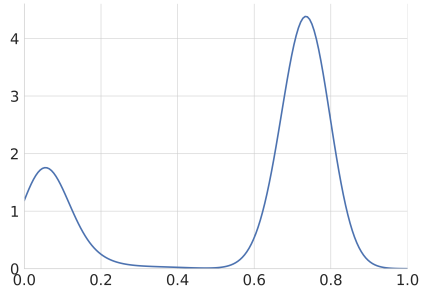


(b) Temperature scaling, with $\tau = 10$.

Figure 5.5: Estimated propensity score distribution of *photo* and *art painting* on the PACS dataset. We compare (a) the original distribution ($\tau = 1$) and (b) the temperature-scaled distribution ($\tau = 10$). Here, the large temperature has the effect of transforming a bimodal distribution into a unimodal one.



(a) No temperature scaling ($\tau = 1$).



(b) Temperature scaling, with $\tau = 10$.

Figure 5.6: Estimated propensity score distribution of *photo* and *sketch* on the PACS dataset. We compare (a) the original distribution ($\tau = 1$) and (b) the temperature-scaled distribution ($\tau = 10$). Here, the large temperature has the effect of smoothing the distribution.

Table 5.5: Analysis of the number of the retrieved matched pairs when matching across the two domain *photo* and *sketch* on the PACS dataset. The fixed caliper threshold and temperature scaling can be used to smooth the propensity score distribution and effect the number of pairs.

Fixed Caliper (t_f)	**Temperature (τ)**	**photo → sketch**	**sketch → photo**
0.1	1	0	0
0	1	1540	3929
0.01	1	6	9
0.01	1.3	14	56
0.01	1.8	25	574
0.01	2.5	41	3082
0.01	10	1540	3929
0.1	10	298	3929

5.7.9 *Energy and Carbon Footprint Estimates*

To highlight the efficiency of Okapi, we provide estimates in Table 5.6 of the carbon footprint associated with the running of it and of the ERM and FixMatch baselines on the iWildCam dataset, using the same hyperparameter configuration used to generate the results in the main text. The runs were conducted in a controlled fashion, using the computing infrastructure and device count in all cases.

Figure 5.7: Examples of input (photo) images and their 1-NN matched (art paint) images retrieved using CaliperNN from the PACS dataset.

Figure 5.8: Examples of input (photo) images and their 1-NN matched (sketch) images retrieved using CaliperNN from the PACS dataset.

Table 5.6: Comparison of the estimated carbon footprint (kgCoeq) of Okapi with the ERM and FixMatch baselines per replicate of the iWildCam dataset. For the controlled training conducted to enable fair computation of these estimates, we used a private infrastructure with an estimated carbon efficiency of 0.432 kgCOeq/kWh and RTX 3090 GPUs, each job being run on a single GPU, coupled with four data-loading workers.

Method	**kgCOeq ↓**
ERM	1.36
FixMatch	2.12
Okapi (ours)	1.97

Algorithm 2 PyTorch-style pseudocode for the CaliperNN matching algorithm for the special case where the domain is binary. The algorithm generalises freely to arbitrary numbers of domains however we restrict ourselves to the binary version here for illustrative purposes.

```
def binary_caliper_nn(
    x_query, # samples to be used as the queries for matching
    s_query, # binary labels indicating the domain of x_query
    x_key, # samples to which the query samples may be matched to.
    s_key, # binary labels indicating the domain of x_key
    ps_query, # propensity scores associated with x_query
    ps_key, # propensity scores associated with x_key
    t_f, # threshold for the fixed caliper
    t_sigma, # number of standard deviations at which to threshold
    k # number of neighbours to attempt to retrieve per query
):
    anchor_inds, positive_inds = [], []
    for direction in (0, 1): # which domain (0 or 1) to treat as the 'anchor'
        key_mask = s_key != direction
        # exclude samples with propensity scores outside the valid range
        # determined by t_f: (1 - t_f, t_f)
        fc_mask = (ps_query > (1 - t_f)) &  (ps_query < t_f)
        anchor_mask = fc_mask & (s_query == direction)
        queries_x_filtered = queries.x[anchor_mask]
        ps_query_filtered = ps_query[anchor_mask]
        fc_mask = (ps_key > (1 - t_f)) & (ps_key < t_f)
        key_mask &= fc_mask
        ps_key_filtered = ps_key[key_mask]
        # 2-norm distance between unfiltered propensity scores
        dists_ps = cdist(ps_query_filtered, ps_key_filtered, p=2)
        # 2-norm distance between the filtered anchors and keys
        dists_x = cdist(queries_x_filtered, x_key[key_mask], p=2)
        # compute sigma as the mean of the per-domain standard deviations
        std_ps = (0.5 * (ps_query_filtered.var() + ps_key_filtered.var())).sqrt()
        std_threshold = t_sigma * std_ps
        # filter out any samples that violate the std-caliper
        dists_x[dists_ps > std_threshold] = float("inf")
        nbr_dists, nbr_inds = dists_x.topk(dim=1, largest=False, k=k)
        # filter out queries not yielding the requisite number of matches (k)
        is_matched = ~nbr_dists.isinf().any(dim=1)

        anchor_inds.append(anchor_mask.nonzero()[is_matched])
        positive_inds.append(key_mask.nonzero()[nbr_inds[is_matched]])

    return cat(anchor_inds, dim=0), cat(positive_inds, dim=0)
```

Algorithm 3 PyTorch-style pseudocode for the online learning algorithm for the special case where the labelled and unlabelled datasets are treated as the domains. The algorithm generalises freely to arbitrary numbers of domains however we restrict ourselves to the binary version here for illustrative purposes.

```
# online_encoder: online encoder
# predictor_head: online predictor head
# propensity_scorer: online propensity scorer
# target_encoder momentum encoder (frozen)
# n_m: memory-bank capacity
# zeta: decay rate of the EMA updates
# tau: temperature-scaling parameter for the propensity scores.
# t_f: fixed caliper threshold for CaliperNN
# t_sigma: number of standard deviations at which to threshold in CaliperNN
# l_sup: supervised loss function
# k: number of matches to retrieve per query
# lambda_: loss pre-factor for the unsupervised loss
# D: Dimensionality of the encodings.
feature_mb, label_mb = Mb(empty(n_m, D)), Mb(empty(n_m)) # initialise memory banks
# load minibatches with B_l labelled (x-y tuples) samples and B_u unlabelled samples
for x_l, y, x_u in train_loader:
    # EMA update: \theta^\prime_t = \zeta \theta^\prime_{t - 1} + (1 - \zeta) \theta_t
    ema_update(target_encoder, online_encoder, zeta)
    features_o_l = online_encoder(x_l) # f_\theta(x_l) -> z_l
    features_t = target_encoder(cat([x_l, x_u])) # f_\theta(x_l \cup x_u) -> z_q^\prime
    y_hat = predictor_head(features_o_l) # g_\phi(z_l) -> \hat{y}
    features_o_u = online_encoder(x_u) # f(x_u) -> z_u
    features_o = cat([features_o_l, features_o_u]) # z_q := z_l \cup z_u
    # normalize the encodings to unit vectors.
    features_o_n = normalize(features_o, p=2, dim=1)
    queries = normalize(features_t, p=2, dim=1)
    # we treat x_l and x_u as coming from domains indexed by 0 and 1, respectively
    labels_l_q, labels_s_u_q = ones(len(x_l)), zeros(len(x_u) # ones and zeros vectors
    labels_q = cat([labels_l_q, labels_u_q])
    mb_mask = is_empty(label_mb) # mask indicating which elements of the MB are filled
    labels_k = cat([labels_q, label_mb[mb_mask].clone()])
    # keys are the union of the queries and the memory-bank-stored features
    keys = cat((queries, feature_mb[mb_mask].clone()), dim=0)
    feature_mb.push(queries) # update the feature memory bank
    label_mb.push(labels_q) # update the label memory bank
    logits_ps_k = propensity_scorer(keys) # h_\psi(z_k) -> e_k
    loss_ps = xent(logits_ps_k, labels_k) # (binary) cross-entropy loss
    logits_ps_k = sigmoid(logits_ps_k / tau) # tempered logistic function
    logits_ps_q = logits_ps_k[:len(queries)]
    inds_a, inds_p = binary_caliper_nn( # compute matches with CaliperNN
        features_t_n, labels_q, keys, labels_k,
        logits_ps_q, logits_ps_k, t_f, t_sigma, k
    )
    z_q, v_k = features_o[inds_a], keys[inds_p] # extract the queries and matches
    match_rate = len(z_q) / len(features_o) # used as an adaptive weight
    loss_u = match_rate * (z_q.unsqueeze(1) - v_q).pow(2).sum(-1).mean()
    loss = l_sup(y_hat, y) + lambda_ * loss_u + loss_ps # aggregate loss
    loss.backward() # compute gradients
    update(online_encoder, predictor_head, propensity_scorer) # optimizer updates
```

5.8 Authorial contributions

I conceived of the overall Okapi framework, performed a preliminary literature review, wrote the core code for it and the baseline methods, configured and ran all non-CaliperNN-specific experiments, and wrote the lion's share of the paper.

S. Romiti was responsible for developing the CaliperNN algorithm – as published in Romiti et al. (2022) – central to the Okapi algorithm, and naturally for the writing of its reference code (adapted for the current work for increased generality and scalability). Within the scope of the current work, she performed, and analysed results of (quantitatively and visually), experiments with the aforementioned CaliperNN algorithm, helped analyse the results of experiments generally, and contributed to figure-making, tabulation and the writing of significant written portions of the text (notably, those sections dedicated to statistical matching, analysis of results, and, in the appendix, analysis of the learned matches and CaliperNN ablations).

V. Sharmanska contributed to the conceptual development of the proposed method and its predecessor, to the analysis of results, and to the guidance of the project as a whole and the paper it begot (participating in regular discussion, giving constructive feedback on the paper and the rebuttal, etc.).

N. Quadrianto proposed the problem setting and the initial roadmap, coördinated and supported team members, and supervised, and led discussion on, the project in all respects. He also contributed to the text itself in the drafting of the introduction and in providing comprehensive feedback on the state of the paper.

BIBLIOGRAPHY

Scudder, Henry (1965). 'Probability of error of some adaptive pattern-recognition machines'. In: *IEEE Transactions on Information Theory* 11.3, pp. 363–371.

Rubin, Donald B (1973). 'Matching to remove bias in observational studies'. In: *Biometrics*, pp. 159–183.

Rosenbaum, Paul R and Donald B Rubin (1983). 'The central role of the propensity score in observational studies for causal effects'. In: *Biometrika* 70.1, pp. 41–55.

– (1985). 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score'. In: *The American Statistician* 39.1, pp. 33–38.

Biglan, Anthony, Dennis Ary and Alexander C Wagenaar (2000). 'The value of interrupted time-series experiments for community intervention research'. In: *Prevention Science* 1.1, pp. 31–49.

Shimodaira, Hidetoshi (2000). 'Improving predictive inference under covariate shift by weighting the log-likelihood function'. In: *Journal of statistical planning and inference* 90.2, pp. 227–244.

Rubin, Donald B (2001). 'Using propensity scores to help design observational studies: application to the tobacco litigation'. In: *Health Services and Outcomes Research Methodology* 2.3, pp. 169–188.

Crump, Richard K, V Joseph Hotz, Guido W Imbens and Oscar A Mitnik (2009). 'Dealing with limited overlap in estimation of average treatment effects'. In: *Biometrika* 96.1, pp. 187–199.

Christian, Parul, Laura E Murray-Kolb, Subarna K Khatry, Joanne Katz, Barbara A Schaefer, Pamela M Cole, Steven C LeClerq and James M Tielsch (2010). 'Prenatal micronutrient supplementation and intellectual and motor function in early school-aged children in Nepal'. In: *Jama* 304.24, pp. 2716–2723.

Stuart, Elizabeth A. (2010). 'Matching methods for causal inference: A review and a look forward'. In: *Stat Sci* 25.1, pp. 1–21. ISSN: 0883-4237.

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

Kamiran, Faisal and Toon Calders (2012). 'Data preprocessing techniques for classification without discrimination'. In: *Knowledge and information systems* 33.1, pp. 1–33.

Muandet, Krikamol, David Balduzzi and Bernhard Schölkopf (2013). 'Domain generalization via invariant feature representation'. In: *International Conference on Machine Learning*. PMLR, pp. 10–18.

Bachman, Philip, Ouais Alsharif and Doina Precup (2014). 'Learning with pseudo-ensembles'. In: *Advances in neural information processing systems* 27.

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

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky (2016). 'Domain-adversarial training of neural networks'. In: *The journal of machine learning research* 17.1, pp. 2096–2030.

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

Hurley, Mikella and Julius Adebayo (2017). 'Credit scoring in the era of big data'. In: *Yale Journal of Law and Technology* 18, pp. 148–216.

Li, Da, Yongxin Yang, Yi-Zhe Song and Timothy M Hospedales (2017). 'Deeper, broader and artier domain generalization'. In: *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550.

Loshchilov, Ilya and Frank Hutter (2017). 'SGDR: Stochastic Gradient Descent with Warm Restarts'. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=Skq89Scxx.

Tarvainen, Antti and Harri Valpola (2017). 'Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results'. In: *Advances in neural information processing systems* 30.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford and Hanna Wallach (2018). 'A reductions approach to fair classification'. In: *International Conference on Machine Learning*. PMLR, pp. 60–69.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). 'Bert: Pre-training of deep bidirectional transformers for language understanding'. In: *arXiv preprint arXiv:1810.04805*.

Donini, Michele, Luca Oneto, Shai Ben-David, John S Shawe-Taylor and Massimiliano Pontil (2018). 'Empirical risk minimization under fairness constraints'. In: *Advances in Neural Information Processing Systems* 31.

Loshchilov, Ilya and Frank Hutter (2018). 'Decoupled Weight Decay Regularization'. In: *International Conference on Learning Representations*.

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

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

Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain and Lucy Vasserman (2019). 'Nuanced metrics for measuring unintended bias with real data for text classification'. In: *Companion proceedings of the 2019 world wide web conference*, pp. 491–500.

Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi and Richard Zemel (2019). 'Flexibly fair representation learning by disentanglement'. In: *International conference on machine learning*. PMLR, pp. 1436–1445.

Dunnmon, Jared A, Darvin Yi, Curtis P Langlotz, Christopher Ré, Daniel L Rubin and Matthew P Lungren (2019). 'Assessment of convolutional neural networks for automated classification of chest radiographs'. In: *Radiology* 290.2, pp. 537–544.

King, Gary and Richard Nielsen (2019). 'Why propensity scores should not be used for matching'. In: *Political Analysis* 27.4, pp. 435–454.

Lahoti, Preethi, Krishna Gummadi and Gerhard Weikum (2019). 'Operationalizing Individual Fairness with Pairwise Fair Representations'. In: *Proceedings of the VLDB Endowment* 13.4, pp. 506–518.

Paszke, Adam et al. (2019). 'Pytorch: An imperative style, high-performance deep learning library'. In: *Advances in neural information processing systems* 32.

Quadrianto, Novi, Viktoriia Sharmanska and Oliver Thomas (2019). 'Discovering fair representations in the data domain'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8227–8236.

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

Sanh, Victor, Lysandre Debut, Julien Chaumond and Thomas Wolf (2019). 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter'. In: *arXiv preprint arXiv:1910.01108*.

Watson, David S, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes and Luciano Floridi (2019). 'Clinical applications of machine learning algorithms: beyond the black box'. In: *Bmj* 364.

Beery, Sara, Elijah Cole and Arvi Gjoka (2020). 'The iWildCam 2020 competition dataset'. In: *arXiv preprint arXiv:2004.10340*.

Berthet, Quentin, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert and Francis Bach (2020). 'Learning with differentiable pertubed optimizers'. In: *Advances in neural information processing systems* 33, pp. 9508–9519.

Brown, Tom et al. (2020). 'Language models are few-shot learners'. In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Chen, Ting, Simon Kornblith, Mohammad Norouzi and Geoffrey Hinton (2020). 'A simple framework for contrastive learning of visual representations'. In: *International conference on machine learning*. PMLR, pp. 1597–1607.

Grill, Jean-Bastien et al. (2020). 'Bootstrap your own latent-a new approach to self-supervised learning'. In: *Advances in Neural Information Processing Systems* 33, pp. 21271–21284.

Gulrajani, Ishaan and David Lopez-Paz (2020). 'In Search of Lost Domain Generalization'. In: *International Conference on Learning Representations*.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie and Ross Girshick (2020). 'Momentum contrast for unsupervised visual representation learning'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738.

Kehrenberg, Thomas, Myles Bartlett, Oliver Thomas and Novi Quadrianto (2020). 'Null-sampling for interpretable and fair representations'. In: *European Conference on Computer Vision*. Springer, pp. 565–580.

Oneto, Luca, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer and Massimiliano Pontil (2020). 'Exploiting mmd and sinkhorn divergences for fair and transferable representation learning'. In: *Advances in Neural Information Processing Systems* 33, pp. 15360–15370.

Sohn, Kihyuk et al. (2020). 'Fixmatch: Simplifying semi-supervised learning with consistency and confidence'. In: *Advances in Neural Information Processing Systems* 33, pp. 596–608.

Taori, Rohan, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht and Ludwig Schmidt (2020). 'Measuring robustness to natural distribution shifts in image classification'. In: *Advances in Neural Information Processing Systems* 33, pp. 18583–18599.

Xie, Qizhe, Minh-Thang Luong, Eduard Hovy and Quoc V Le (2020). 'Self-training with noisy student improves imagenet classification'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon and Marshall Burke (2020). 'Using publicly available satellite imagery and deep learning to understand economic well-being in Africa'. In: *Nature communications* 11.1, pp. 1–11.

Yu, Fisher, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan and Trevor Darrell (2020). 'Bdd100k: A diverse driving dataset for heterogeneous multitask learning'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645.

Andreassen, Anders, Yasaman Bahri, Behnam Neyshabur and Rebecca Roelofs (2021). 'The evolution of out-of-distribution robustness throughout fine-tuning'. In: *arXiv preprint arXiv:2106.15831*.

Bao, Hangbo, Li Dong and Furu Wei (2021). 'Beit: Bert pre-training of image transformers'. In: *arXiv preprint arXiv:2106.08254*.

Bommasani, Rishi et al. (2021). 'On the opportunities and risks of foundation models'. In: *arXiv preprint arXiv:2108.07258*.

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski and Armand Joulin (2021). 'Emerging properties in self-supervised vision transformers'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660.

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

Dwibedi, Debidatta, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet and Andrew Zisserman (2021). 'With a little help from my friends: Nearest-neighbor contrastive learning of visual representations'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597.

Gong, Chengyue, Dilin Wang and Qiang Liu (2021). 'AlphaMatch: Improving Consistency for Semi-supervised Learning with Alpha-divergence'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13683–13692.

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár and Ross Girshick (2021). 'Masked Autoencoders Are Scalable Vision Learners'. In: *arXiv:2111.06377*.

Koh, Pang Wei et al. (2021). 'Wilds: A benchmark of in-the-wild distribution shifts'. In: *International Conference on Machine Learning*. PMLR, pp. 5637–5664.

Koohpayegani, Soroush Abbasi, Ajinkya Tejankar and Hamed Pirsiavash (2021). 'Mean shift for self-supervised learning'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10326–10335.

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

Lienen, Julian and Eyke Hüllermeier (2021). 'Credal Self-Supervised Learning'. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 14370–14382. URL: https://proceedings.neurips.cc/paper/2021/file/7866c91c59f8bffc92a79a7cd09f9af9-Paper.pdf.

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

Radford, Alec et al. (2021). 'Learning transferable visual models from natural language supervision'. In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.

Van Gansbeke, Wouter, Simon Vandenhende, Stamatios Georgoulis and Luc V Gool (2021). 'Revisiting contrastive methods for unsupervised learning of visual representations'. In: *Advances in Neural Information Processing Systems* 34.

Verma, Vikas, Thang Luong, Kenji Kawaguchi, Hieu Pham and Quoc Le (2021). 'Towards domain-agnostic contrastive learning'. In: *International Conference on Machine Learning*. PMLR, pp. 10530–10541.

Alayrac, Jean-Baptiste et al. (2022). 'Flamingo: a Visual Language Model for Few-Shot Learning'. In: *arXiv preprint arXiv:2204.14198*.

Baevski, Alexei, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu and Michael Auli (2022). 'Data2vec: A general framework for self-supervised learning in speech, vision and language'. In: *arXiv preprint arXiv:2202.03555*.

Chowdhery, Aakanksha et al. (2022). 'Palm: Scaling language modeling with pathways'. In: *arXiv preprint arXiv:2204.02311*.

Goyal, Priya et al. (2022). 'Vision models are more robust and fair when pretrained on uncurated images without supervision'. In: *arXiv preprint arXiv:2202.08360*.

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

Kim, Donghyun, Kaihong Wang, Stan Sclaroff and Kate Saenko (2022). 'A Broad Study of Pre-training for Domain Generalization and Adaptation'. In: *arXiv preprint arXiv:2203.11819*.

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell and Saining Xie (2022). 'A ConvNet for the 2020s'. In: *arXiv preprint arXiv:2201.03545*.

Romiti, Sara, Christopher Inskip, Viktoriia Sharmanska and Novi Quadrianto (2022). 'RealPatch: A Statistical Matching Framework for Model Patching with Real Samples'. In: *CoRR* abs/2208.02192.

Sagawa, Shiori et al. (2022). 'Extending the WILDS Benchmark for Unsupervised Adaptation'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=z7p2V6KROOV.

Wiles, Olivia, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham and Ali Taylan Cemgil (2022). 'A Fine-Grained Analysis on Distribution Shift'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Dl4LetuLdyK.

Yu, Jiahui, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini and Yonghui Wu (2022). 'CoCa: Contrastive Captioners are Image-Text Foundation Models'. In: *arXiv preprint arXiv:2205.01917*.

Part III

END

This part marks the end of this thesis and the thread of its tale. Herein, I discuss the works presented in Part II retrospectively, including their merits and demerits, their broader context (both contemporaneous and present), and what the future may yet hold for the avenues explored given current trends.

6 | DISCUSSION

"I do not care what comes after; I have seen the dragons on the wind of morning."

The Farthest Shore
Ursula K. Le Guin

Of what was said and the silence between

Synopsis

I have introduced three methods, each a solution to a different problem, though with all problems conjoined by a notion of distributional robustness. To briefly recount, and thereby set the scene for the discussion to follow: in Chapter 3 we proposed an INN-based transfer-learning approach – transferring invariance from a partially-labelled representative set to the training set – for solving a SC problem where for training samples the target is completely determined by the sensitive attribute and because the latter is easier to learn it constitutes a shortcut; in Chapter 4, we proposed to match in representation-space the support (of intersectional groups) of the training and deployment sets in order to overcome a relaxed version of aforesaid SC problem where the deployment set is a dataset representative (in support) of the test set, but for which no annotations are required, in contrast to the representative set featured in Chapter 3; finally, in Chapter 5 we grappled with the problem of semi-supervised DG – by which I mean the problem of how to make effective use of unlabelled data drawn from extra domains to bolster OOD performance – and proposed a consistency-regularised approach employing a robust, causally-inspired algorithm as a match-generation engine, with the matches bootstrapped from encodings guided (in optimisation) by past matches. Again, all of these works are mine but not unshared, for I owe much gratitude to all my coäuthors – my colleagues, my advisors, my friends – for allowing this thesis to become what it has thus become. I shall begin by discussing, with the gift of hindsight and the wisdom which comes with being humbled by experience, the limitations of the works presented, and I will do so candidly: for as I said at the outset, I do not fear the 'lesser elements' – for every candle lit there is a shadow cast, but it is also because of the shadow that we may see the light. Among these limitations, I shall begin with the most fundamental one that pervades all the works and is itself bipartite, comprising the assumed availability of subgroup (I, arbitrarily, use this term to cover the myriad names for S here) annotations, and the assumed discrete property of these annotations.

The gift of hindsight

In journeying through this thesis, we have witnessed a progressive relaxation of the first part, such that in Chapter 4 we need only know in advance all possible (w.r.t. the test set) sources (and know that the deployment set contains them); by Chapter 5 we need only be able to partition

the data into two disjoint sets, across which the matching is performed. We argued in Chapter 3 that partially-labelled data, for which the subgroup- but not the target-attribute is provided, is generally more 'readily available'. While this may be true for certain domains and applications, it is not true for others, and it is largely contingent on what the target and subgroup attributes have been determined to be, and, moreover, how they interact (their relative complexity). It is reasonable in the case of a face dataset like CelebA to assume that gender information can be explicitly, or implicitly, gleaned – for instance, by virtue of the images appearing in gender-specific catalogues – whereas 'Smiling' is not a feature to innately partition by (and we would assume that in the forgoing catalogue case that most models will be smiling), and these two attributes would indeed realistically serve well as the subgroup and target attributes, respectively, for Chapter 3's framework; it is less easy to intuit whether a shortcut would emerge though we contend there is no harm in taking the precaution if using a lossless encoder. If it were 'Hair Colour' that we sought invariance to and 'Age' we were targeting (this combination plausible enough), however, we would not expect things to pan out nearly so neatly.

The problem of learning distributional-robust models from biased data, and circumventing SCs thereof, when the distributions (subgroups) in question are unknown has attracted considerable attention as of late (Hashimoto et al., 2018; Sohoni et al., 2020; Creager et al., 2021; Liu et al., 2021a; Pezeshki et al., 2021; Kim et al., 2022; Taghanaki et al., 2022, inter alia). However, one must unavoidably rely on certain assumptions (inductive biases derived from prior knowledge of the task/domain) to compensate for the lack of information and the referenced methods can fall flat (in the sense of underperforming the ERM baseline) should such not be satisfiable. Chapter 4 entails this to a degree, in that the sources contained in the deployment set need be discovered for constructing support-representative batches, although the (annotated) sources in the training set can subserve this process of discovery, for it is not entire subgroups or classes that are excluded from annotation but their intersections. Nonetheless, this discovery can fail to align with expectation, and we assume in Chapter 4 that the subgroups/targets are sufficiently salient to be well-clustered by un-/semi-supervised means – one would not expect a source formed from gender/pulmonary infiltration (subgroup and target attributes, respectively) to lend itself to natural clusters.

Speaking of the discreteness assumption, we have throughout assumed that the subgroup can be represented, innately or by simple preprocessing, by some index; while this accords with much of the literature (because of its simplicity and prevalence) there are nonetheless cases where this assumption is untenable. This is particularly germane to the method proposed in Chapter 4 which expects the data to both be clusterable and for the sources to be finite (and, practically, small enough to be computationally tractable) sets, such that one can balance batches w.r.t. them – how one might extend the method to continuous subgroups is unclear – the notion perhaps not even sensible – and in order to maintain tractability in pursuit of generality it would seem necessary to cede some, or all, of the theoretical guarantees established. The adversarial-infomin approach adopted in Chapter 3, on the other hand, can be readily adapted to continuous subgroups, for instance, by substituting cross-entropy with HGRMC (recalling that we use HGRMC in said chapter not as an objective but as a fairness metric for tasks involving *categorical* subgroups).

I have spoken before of the practical deficiencies of the AdvL paradigm in a DL context – namely the fragility of optimisation, disposition to cyclic dynamics, architectural dependency, and

The idiosyncrasies of adversarial infomin

the loss of guarantees incurred by estimating the best-response dynamics with a finite (typically small) number of steps – but it is apposite that we revisit these points again here, retrospectively instead of prospectively. Indeed – as again spoken of before but of which there is again no harm recalling – said deficiencies have been well noted in the infomin-related literature, motivating attempts to develop non-adversarial approaches, taking advantage of, for instance, the exact-density-estimation afforded by NFs (Balunović et al., 2021), sliced mutual information (Goldfeld and Greenewald, 2021) to scalably target the infomin objective directly (Chen et al., 2022), or the information-bottleneck principle (Tishby and Zaslavsky, 2015; Moyer et al., 2018) for which the subgroup attribute plays the part of the *nuisance factor*, to reconcile the parlance. While cognisant of, and taking measures (e.g. ensembling, limiting the volume of the latent space) to ameliorate, these deficiencies, the making of the aforesaid chapters was, in no small amount, harried by them, often demanding careful and exhaustive tuning to coax the respective methods into working as desired, the 'tuning' itself problematic due to the underspecified nature of the problem setups. We found the inherent instability of AdvL particularly pronounced in the case of Chapter 3, specifically, owing to the compounding instabilities imparted by the invertible architectures, a problem which itself would not be addressed *in toto* until Behrmann et al., 2021. In light of the forgoing, exploring non-adversarial methods, of the kind referenced, as alternative infomin engines for the higher-level methods proposed in said chapters is well-founded (for those wishing to apply, or undertake further research, on said methods), though I cannot – despite their theoretical appeal – attest to their practical efficacy in these contexts, *a priori*; I would note in defence of the chapters, however, that for neither higher-level method is the form of said engine integral to its identity, and is in fact modular, such that we may conceivably freely interchange engines subject to their respective requirements. In Chapter 5 we induce invariance by non-adversarial means ourselves, namely by enforcing similarity between matched samples from different domains within representation space – the adversary substituted with a non-parametric match-generator – and we thereby avert many of the optimisation difficulties plaguing the aforementioned chapters, though the problem of judicious hyperparameter-selection lingers, perhaps even amplified by the non-parametricity – we point in the paper to adaptive-configuration of the calipers being an obvious (in motivation but not implementation) avenue of extension.

On the problem of identifiability

A central question that I have perhaps given shorter shrift to than due, is that of the identifiability of bias – a question of two parts: 1) diagnosing those cases warranting methodological (or data-sided) interventions, such as those delineated; 2) and evaluating the success of those interventions – given the element of underspecification (Semenova et al., 2019). The obtuse answer would be that it need not be answered, or answerable, presupposing that the model in question (figuratively) is to be deployed regardless with or without intervention (the control), for comparisons should be made w.r.t. the latter rather than w.r.t. an ideal, a gold-standard that may (and often will not be) practically realisable – one need only ensure that performance does not degrade based on what is evaluable. This cavalier ethos of 'do the best we can with what's available [short of bringing humans directly into the fold], regardless of the consequences', however is patently not an admirable one, however, and realistically one (a practitioner or collective) does, or should, aim to deploy models subject to their being sensible/unbiased (if only for reputation's sake) and to thoroughly diagnose, and attempt to remedy failure cases via well-measured and iterative processes. The problem of underspecification – of validating models without access to data

representative of that to-be-encountered at deployment time, as defines DG – that encapsulates the second aspect of the question remains an outstanding one, that has been discussed broadly and in the context of DG specifically, posing a threat to the validity of inter-method comparisons (Gulrajani and Lopez-Paz, 2020). In absence of a such validation set by which to quantify biases and generalisation-failures, one may draw upon methods from the explainability/interpretability literature (Gunning et al., 2019) to determine whether the learned solutions align with the intended solutions; one may, for instance, readily diagnose the use of background as a shortcut, per the now-canonical example from Beery et al. (2018), with a standard-method-in-that-literature in Grad-cam (Selvaraju et al., 2017), allowing for a targeted intervention (e.g. by augmentation) – one may even use the resulting attribution maps directly for this purpose as proposed by Taghanaki et al. (2022). It is for this reason that we emphasise the interpretability aspect in Chapter 3, for even if we cannot sufficiently debias a model (if it need be debiased), we might glean when this is the case and for what reason without need for quantification. This is, of course, easier for some domains than others – images being naturally interpretable due to their underlying structure and familiarity (the window through which we, quite literally, see the world), whereas tabular data generally affords no such luxuries – but the bottom line of this excursis is my advocacy (which I am certainly not unique for) for human oversight, and the rigorous model-vetting it should beget, irrespective of the theoretical trustability of the methods/data.

Interpretability and oversight

As far as limitations go, I shall last speak expressly of empirical and theoretical claims, and the desire to couple the two. In Chapter 4 we provide theoretical guarantees regarding when the proposed support-matching should succeed, subject to certain (relatively-loose) assumptions about the data-generating distributions. Chapter 5, however, features no such proofs – only intuition – for the efficacy of its respective method, that being a notable weakness of the current incarnation of the paper – having such may serve to guide us regarding the configuration of the matching algorithm and thereby obviate the somewhat-lengthy hyperparameter-selection procedure went through to obtain the presented results. In Chapter 3, on the other hand, we leverage an established infomin framework, around which (or in the vicinity of which) there is a significant body of existing theoretical work – both within the context of the various subfields ML concerned with it (DA, DG, AF) and within the broader context of game theory, dynamical systems theory, and information theory – however, we again furnish only intuition and empirical claims for the stabilisation techniques, for which (proof-driven) theoretical-grounding would be demonstrably desirable.

Theoretical-groundedness

Of what has since become and might yet be

Recent work on AI-alignment has proposed 'scaling supervision' (understanding 'supervision' to mean what I called before 'oversight') of generative-language models by using human-aligned LLMs not only as the model-to be-supervised, but also as the supervisor, acting according to a set of values (in the form of prompts), or 'constitution', specified by the practitioner (Bai et al., 2022a; Bowman et al., 2022). Such AI supervisors exhibit the ability to accurately, and well-calibratedly, detect biases violating said constitution in generated responses and provide feedback for redressing them; this feedback may be used to further align the supervised model,

Learning from human preferences and scaling supervision

as substitute for the human-generated feedback fuelling the eponymous reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020; Bai et al., 2022b), analogously termed reinforcement learning from AI feedback (RLAIF). While RLHF does not allow for sidestepping of annotations (which has been ameliorated by unprecedented levels of crowdsourcing), tuning based on preferences (given for pairs of responses) allows for increased expressivity compared with traditional supervised approaches, pertinent when the bias exists on a (difficult-to-quantify) spectrum or is fundamentally subjective in nature. RLXF (to coin a catch-all initialism for reinforcement learning from some kind of feedback) affords a spectacularly simple and general paradigm in terms of alignment and distributional robustness (that we view throughout this thesis in terms of invariance and worst-group performance) yet it and fairness should not be confused as one and the same though despite their frequent overlap – for some tasks there is no subjective element to engender the notion of 'preference', i.e. the value system is fixed (inherently or by legislation), and so may be the answer-set (e.g. the label-set for closed-set classification). Thus, while one may use RLXF to obtain generative models that are more helpful/harmless/honest (the 'HHH triad') – a sense of 'debiasing' – it is no panacea for problems of the nature discussed in this thesis, especially so for specialised tasks for which there is little prior knowledge – acquired from large-scale pre-training – to be leveraged; methods like those presented herein might then still have their place in this brave new world of ML where multi-modal foundation models (Driess et al., 2023; Huang et al., 2023; Katz et al., 2023; OpenAI, 2023) and RLXF putatively rule the roost (though most definitely the headlines). The premise of scaling supervision, through this process of self-review, as in Bai et al. (2022a), may have broader application, however: for image-classification, one could conceive of a model that classifies based on natural-language descriptions of a scene, providing a reasoning mechanism that is evaluable by 'constitutional' critics and human auditors (who, again, should very much remain part of the equation – they are to be assisted not superseded).

Beyond this alignment-centric perspective, it is also appropriate to speak of the merits of large-scale pre-training, and subsequent fine-tuning, from a non-generative perspective and rather in terms of the distributionally-robustness representations it may give rise to; whereas the focus before was on the language domain, here it will be on the vision one which preponderated in this thesis's works, due foremost to its interpretability. A host of prominent works attest that large-scale (primarily self-supervised) training improves downstream robustness for visual tasks, in both covariate (Hendrycks et al., 2019, 2020; Radford et al., 2021) and target/subgroup senses (Liu et al., 2021b; Goyal et al., 2022), with Goyal et al. (2022) showing that the pre-training corpus, if of sufficient scale (here being on the order of millions of samples), need not be curated for this to apply – curation being the natural enemy of scalability, which begets diversity which itself subserves robustness, at least in the OOD sense (for it cannot necessarily imply invariance in general). Thus, for some OOD/DG tasks, one may not need to resort to dedicated algorithms, but instead simply fine-tune, or use as is, the representations of such foundation models – one may even be able to obviate the need to perform any manner of fine-tuning if the task in question is sufficiently general as to admit zero-shot solutions by vision-language models (Radford et al., 2021; Alayrac et al., 2022). There are two issues that prevent or hinder such foundation-model-based approaches from providing general-purpose solutions, however. The first I have alluded to before, that being that for tasks of a specialised nature – such as those found in medical imaging – we

would expect the degree of positive transfer from large-scale web-derived datasets to be limited, on the representation front, and zero-shot approaches all the more inauspicious; the second is that the robustness of pre-training models is known to degrade (become 'distorted') as the result of fine-tuning processes (Andreassen et al., 2021; Kumar et al., 2022), a phenomenon which can be mitigated but not altogether averted, at present – the development of more-robust, less-distorting fine-tuning routines remains an active area of research (Lee et al., 2022; Trivedi et al., 2023). An additional concern, of a more practical, but not innegligibly-niche nature, is that running the germane foundation model might demand more compute or time than can be afforded, even when run in inference mode (simply loading larger foundation models into memory can be challenging, requiring model-sharding), this being most-obviously applicable to low-compute edge devices. These issues aside, while foundation models may pave some of the way to robustness, in certain respects, it will often be the case that one can do better given the data available, and the remainder of the way need be paved with dedicated distributionally-robust methods; in cases where no annotations for the attribute to-be-invariant to are given or obtainable, however, they do provide a promising recourse, one that requires few, if no, assumptions – one may obtain robustness for 'free', so to speak – and little to no (in the zero-shot case) compute be spent on training, albeit with the aforementioned caveat on compute standing.

BIBLIOGRAPHY

Tishby, Naftali and Noga Zaslavsky (2015). 'Deep learning and the information bottleneck principle'. In: *2015 ieee information theory workshop (itw)*. IEEE, pp. 1–5.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg and Dario Amodei (2017). 'Deep reinforcement learning from human preferences'. In: *Advances in neural information processing systems* 30.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra (2017). 'Grad-cam: Visual explanations from deep networks via gradient-based localization'. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

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

Hashimoto, Tatsunori B., Megha Srivastava, Hongseok Namkoong and Percy Liang (2018). 'Fairness Without Demographics in Repeated Loss Minimization'. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 1934–1943.

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

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf and Guang-Zhong Yang (2019). 'XAI—Explainable artificial intelligence'. In: *Science robotics* 4.37, eaay7120.

Hendrycks, Dan, Kimin Lee and Mantas Mazeika (2019). 'Using pre-training can improve model robustness and uncertainty'. In: *International Conference on Machine Learning*. PMLR, pp. 2712–2721.

Semenova, Lesia, Cynthia Rudin and Ronald Parr (2019). 'A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning'. In: *arXiv preprint arXiv:1908.01755*.

Gulrajani, Ishaan and David Lopez-Paz (2020). 'In Search of Lost Domain Generalization'. In: *International Conference on Learning Representations*.

Hendrycks, Dan, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan and Dawn Song (2020). 'Pretrained transformers improve out-of-distribution robustness'. In: *arXiv preprint arXiv:2004.06100*.

Sohoni, Nimit Sharad, Jared Dunnmon, Geoffrey Angus, Albert Gu and Christopher Ré (2020). 'No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems'. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Stiennon, Nisan et al. (2020). 'Learning to summarize with human feedback'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33, pp. 3008–3021.

Andreassen, Anders, Yasaman Bahri, Behnam Neyshabur and Rebecca Roelofs (2021). 'The evolution of out-of-distribution robustness throughout fine-tuning'. In: *arXiv preprint arXiv:2106.15831*.

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

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

Creager, Elliot, Jörn-Henrik Jacobsen and Richard Zemel (2021). *Environment Inference for Invariant Learning*. arXiv: 2010.07249 [cs.LG].

Goldfeld, Ziv and Kristjan Greenewald (2021). 'Sliced mutual information: A scalable measure of statistical dependence'. In: *Advances in Neural Information Processing Systems* 34, pp. 17567–17578.

Liu, Evan Zheran, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang and Chelsea Finn (2021a). *Just Train Twice: Improving Group Robustness without Training Group Information*. arXiv: 2107.09044 [cs.LG].

Liu, Hong, Jeff Z HaoChen, Adrien Gaidon and Tengyu Ma (2021b). 'Self-supervised learning is more robust to dataset imbalance'. In: *arXiv preprint arXiv:2110.05025*.

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

Radford, Alec et al. (2021). 'Learning transferable visual models from natural language supervision'. In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.

Alayrac, Jean-Baptiste et al. (2022). 'Flamingo: a Visual Language Model for Few-Shot Learning'. In: *arXiv preprint arXiv:2204.14198*.

Bai, Yuntao et al. (2022a). 'Constitutional AI: Harmlessness from AI Feedback'. In: *arXiv preprint arXiv:2212.08073*.

Bai, Yuntao et al. (2022b). 'Training a helpful and harmless assistant with reinforcement learning from human feedback'. In: *arXiv preprint arXiv:2204.05862*.

Bowman, Samuel R et al. (2022). 'Measuring progress on scalable oversight for large language models'. In: *arXiv preprint arXiv:2211.03540*.

Chen, Yanzhi, Weihao Sun, Yingzhen Li and Adrian Weller (2022). 'Scalable Infomin Learning'. In: *Advances in Neural Information Processing Systems*.

Goyal, Priya et al. (2022). 'Vision models are more robust and fair when pretrained on uncurated images without supervision'. In: *arXiv preprint arXiv:2202.08360*.

Kim, Nayeong, Sehyun Hwang, Sungsoo Ahn, Jaesik Park and Suha Kwak (2022). 'Learning Debiased Classifier with Biased Committee'. In: *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Kumar, Ananya, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma and Percy Liang (2022). 'Fine-Tuning Distorts Pretrained Features and Underperforms Out-of-Distribution'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=UYneFzXSJWh.

Lee, Yoonho, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang and Chelsea Finn (2022). 'Surgical fine-tuning improves adaptation to distribution shifts'. In: *arXiv preprint arXiv:2210.11466*.

Taghanaki, Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri and Ghassan Hamarneh (2022). *MaskTune: Mitigating Spurious Correlations by Forcing to Explore*. arXiv: 2210.00055 [cs.LG].

Driess, Danny et al. (2023). 'Palm-e: An embodied multimodal language model'. In: *arXiv preprint arXiv:2303.03378*.

Huang, Shaohan et al. (2023). 'Language is not all you need: Aligning perception with language models'. In: *arXiv preprint arXiv:2302.14045*.

Katz, Daniel Martin, Michael James Bommarito, Shang Gao and Pablo Arredondo (2023). 'Gpt-4 passes the bar exam'. In: *Available at SSRN 4389233*.

OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Trivedi, Puja, Danai Koutra and Jayaraman J. Thiagarajan (2023). *A Closer Look at Model Adaptation using Feature Distortion and Simplicity Bias*. arXiv: 2303.13500 [cs.LG].