Université du Québec
Institut national de la recherche scientifique
Centre Énergie Matériaux Télécommunications

On Robust and Generative Neural Networks with Applications to Brain-Computer Interfaces and Object Recognition

By

Isabela Albuquerque

A thesis submitted in fulfillment of the requirements for the degree of
Doctorate of Sciences, Ph.D
in Telecommunications

Evaluation Committee

Internal evaluator and committee president:	Prof. Kulbir Ghuman INRS-EMT
External evaluator 1:	Prof. Éric Granger École de Technologie Supérieure
External evaluator 2:	Prof. Hansenclever Bassani Universidade Federal de Pernambuco
Research advisor:	Prof. Tiago H. Falk INRS-EMT

I acknowledge that the majority of this work took place on the traditional and unceded territory of the Kanien'kehá:ka (Mohawk) Nation.

Acknowledgements

This work was only possible due to the contribution of countless people who supported, challenged, and helped me in the past years. I am fully aware of how privileged I am for having the opportunity to follow my dreams and I am deeply grateful to everyone that allowed me to go through this journey.

I would like to thank Prof. Tiago Falk for his support, kindness, and guidance during the last years. I am also grateful for sharing this time with my colleagues from the MuSAE Lab: Anderson Ávila, Raymundo Cassani, Abhishek Tiwari, Shruti Kshirsagar, Marília Soares, Olivier Rosanne, Belmir Jesus, Marc-Antoine Moinnereau, Liviu Ivanescu, Stefany Bedoya, João Felipe Santos, and Diana Tobón-Vallejo.

I was very fortunate to collaborate with amazing people from whom I learned so much. I am grateful to all my co-authors and collaborators during this work, especially Prof. Ioannis Mitliagkas for being so uplifting and a constant source of motivation. During my Ph.D., I was lucky to have the opportunity of doing industry internships where I could pursue research directions that would not be possible otherwise. I am grateful to all the amazing people I met, especially my hosts Nikhil Naik (Salesforce Research), Pascal Vincent (Facebook AI Research), and David Warde-Farley (DeepMind) for their enthusiasm, guidance, and for helping me to become a better researcher.

Finally, I am deeply grateful to my family, Patricia, Antonio, Carolina, Nair, Ary, Pompeia, Ana, and Julia for their love, trust, and constant care despite the 7000 kilometers geographically separating us. I would also like to thank my friends Eynat, Marília Costa, Renata, Marília Paranhos, and Laíza, for their love and for incessantly rooting for me in the past 18 years. Last, but certainly not least, I would like to thank my partner João for his love and endless support. Thank you for sharing this journey with me, for making sure we celebrated all the wins, even the smallest ones, and for always making the bad days better.

Abstract

Standard assumptions for supervised learning under the risk minimization framework are too strict and likely unrealistic. As a consequence, it is often the case that theoretically justified methods fail in practice or require extra engineering in order to work once such assumptions are not satisfied. A well known example of this issue is the requirement that training and testing datasets are sampled independently from a fixed distribution, even though real-world applications of machine learning are susceptible to distribution shifts. Such mismatches between training and testing conditions might be due to natural perturbations to the data distribution induced by, for example, changes in the collection conditions, or synthetic distortions such as adversarial attacks. In this thesis, we make contributions towards improving the applicability of machine learning, especially neural networks, by better understanding the out-of-distribution generalization problem and developing versatile and robust learning systems. Our main goals consist of proposing approaches to assess the existence of distribution shifts, defining to which kinds of distributions beyond the training domains it is possible to expect a model will generalize to, and devising algorithms capable of mitigating the effects of distribution shifts. Moreover, we aim at proposing versatile and general approaches which can also be applied to other settings and problems. We achieve such goals by introducing a new dataset, generalization guarantees, and algorithms. Considering the current wide range of machine learning applications, we evaluate the proposed contributions on different domains where distribution shifts are ubiquitous and potentially harmful: computer vision tasks and brain-computer interfaces. We start our contributions by introducing WAUC, a new multi-modal dataset for the assessment of mental workload in real-world conditions such as varying levels of physical strain containing recordings from 48 subjects. Next, we propose a strategy to estimate two types of discrepancies between the data collected from different domains and evaluate it on the WAUC dataset consider different subjects as distinct domains. We show that the estimates of statistical shifts obtained with the proposed approach can be used for investigating other aspects of a machine learning pipeline, such as quantitatively assessing the effects of different normalization strategies commonly used to mitigate cross-subject variability. Furthermore, we investigate the relationship between the estimated shifts and the accuracy of mental workload prediction models.

We then focus on the domain generalization setting: a formalization where the data generating process at test time may yield samples from never-before-seen domains. We prove a generalization bound for this setting and show that representing the data in a space that yields predictive power for a particular task and where training distributions are indistinguishable, induces low risk over unseen domains. Minimizing the terms of the bound yields an adversarial approach in which pairwise domain divergences are estimated and minimized. In addition, we show that the proposed algorithmic innovations are versatile and can be employed in other machine learning applications where learning can also be formulated as a minimax optimization problem. We consider the training of Generative Adversarial Networks (GANs) and revisit the multiple-discriminator setting by framing the simultaneous minimization of losses provided by different models as a multi-objective optimization problem. We introduce the use of gradient-based multi-objective optimization for training GANs and compare the multiple gradient descent algorithm with hypervolume maximization on a number of datasets. Moreover, we argue that the previously proposed methods and hypervolume maximization can all be seen as variations of multiple gradient descent in which the update direction can be computed more efficiently. We finish our contributions to the development of general and robust learning systems by proposing a unified and versatile approach to mitigate both natural and artificial domain shifts via the use of random projections. We show that such projections, implemented as convolutional layers with random weights placed at the input of a model, are capable of increasing the overlap between the different distributions that may appear at training/testing time. We evaluate the proposed approach on settings where different types of distribution shifts occur, and show it provides gains in terms of improved out-of-distribution performance under the domain generalization setting, as well as increased robustness to adversarial perturbations.

Keywords: Out-of-distribution generalization, robust machine learning, generative modeling, multi-objective optimization, brain-computer interfaces, object recognition

Résumé

Les hypothèses standard pour l'apprentissage supervisé dans le cadre de la minimisation des risques sont trop strictes et probablement irréalistes. En conséquence, il arrive souvent que des méthodes théoriquement justifiées échouent dans la pratique ou nécessitent une ingénierie supplémentaire pour fonctionner une fois que ces hypothèses ne sont pas satisfaites. Un exemple bien connu de ce problème est l'exigence selon laquelle les ensembles de données d'entraînement et de test sont échantillonnés indépendamment d'une distribution fixe, même si les applications réelles de l'apprentissage automatique sont sensibles aux changements de distribution. De telles discordances entre les conditions d'entraînement et de test peuvent être dues à des perturbations naturelles de la distribution des données induites, par exemple, par des changements dans les conditions de collecte, ou à des distorsions synthétiques telles que des attaques contradictoires. Dans cette thèse, nous contribuons à améliorer l'applicabilité de l'apprentissage automatique, en particulier les réseaux de neurones, en comprenant mieux le problème de généralisation hors distribution et en développant des systèmes d'apprentissage polyvalents et robustes. Nos principaux objectifs consistent à proposer des approches pour évaluer l'existence de décalages de distribution, définir à quels types de distributions au-delà des domaines d'apprentissage il est possible de s'attendre à ce qu'un modèle se généralise, et concevoir des algorithmes capables d'atténuer les effets des décalages de distribution. De plus, nous visons à proposer des approches polyvalentes et générales qui peuvent également être appliquées à d'autres contextes et problèmes. Nous atteignons ces objectifs en introduisant un nouvel ensemble de données, des garanties de généralisation et des algorithmes. Compte tenu du large éventail actuel d'applications d'apprentissage automatique, nous évaluons les contributions proposées dans différents domaines où les changements de distribution sont omniprésents et potentiellement nocifs : tâches de vision par ordinateur et interfaces cerveau-ordinateur. Nous commençons nos contributions en introduisant WAUC, un nouvel ensemble de données multimodales pour l'évaluation de la charge de travail mentale dans des conditions réelles telles que des niveaux variables d'effort physique contenant des enregistrements de 48 sujets. Ensuite, nous proposons une stratégie pour estimer deux types d'écarts entre les données collectées dans différents domaines et les évaluer sur l'ensemble de données WAUC en considérant différents sujets comme des domaines distincts. Nous montrons que les estimations des décalages statistiques obtenues avec l'approche proposée peuvent être utilisées pour étudier d'autres aspects d'un pipeline d'apprentissage automatique, tels que l'évaluation quantitative des effets de différentes stratégies de normalisation couramment utilisées pour atténuer la variabilité inter-sujets. De plus, nous étudions la relation entre les changements estimés et la précision des modèles de prédiction de la charge de travail mentale.

Nous nous concentrons ensuite sur le paramètre de généralisation de domaine: une formalisation où le processus de génération de données au moment du test peut produire des échantillons de domaines jamais vus auparavant. Nous prouvons une généralisation liée à ce paramètre et montrons que la représentation des données dans un espace qui donne un pouvoir prédictif pour une tâche particulière et où les distributions d'entraînement sont indiscernables, induit un faible risque sur des domaines invisibles. La minimisation des termes de la borne donne une approche contradictoire dans laquelle les divergences de domaine par paires sont estimées et minimisées. De plus, nous montrons que les innovations algorithmiques proposées sont polyvalentes et peuvent être utilisées dans d'autres applications d'apprentissage automatique où l'apprentissage peut également être formulé comme un problème d'optimisation minimax. Nous considérons l'entraînement de *Generative Adversarial Networks* (GANs) et revisitons le paramètre de discriminateur multiple en encadrant la minimisation simultanée des pertes fournies par différents modèles comme un problème d'optimisation multi-objectifs. Nous introduisons l'utilisation de l'optimisation multi-objectifs basée sur le gradient pour l'entraînement des GANs et comparons l'algorithme de descente de gradient multiple avec la maximisation de l'hypervolume sur un certain nombre d'ensembles de données. De plus, nous soutenons que les méthodes proposées précédemment et la maximisation de l'hypervolume peuvent toutes être considérées comme des variations de descente à gradient multiple dans lesquelles la direction de mise à jour peut être calculée plus efficacement. Nous terminons nos contributions au développement de systèmes d'apprentissage généraux et robustes en proposant une approche unifiée et polyvalente pour atténuer les changements de domaine naturels et artificiels via l'utilisation de projections aléatoires. Nous montrons que de telles projections, mises en œuvre sous forme de couches convolutives avec des poids aléatoires placés en entrée d'un modèle, sont capables d'augmenter le chevauchement entre les différentes distributions pouvant apparaître au moment de l'apprentissage/du test. Nous évaluons l'approche proposée sur des paramètres où différents types de changements de distribution se produisent, et montrons qu'elle fournit des gains en termes d'amélioration des performances hors distribution dans le cadre de la généralisation du domaine, ainsi qu'une robustesse accrue aux perturbations contradictoires.

Mots-clés: Généralisation hors distribution, apprentissage automatique robuste, modélisation générative, optimisation multi-objectifs, interfaces cerveau-ordinateur, reconnaissance d'objets

Contents

List of Figures

List of Tables

List of Abbreviations

ALP	Adversarial Logit Pairing
AT	Adversarial Training
BCI	Brain-Computer Interface
CIDDG	Conditional Invariant Deep Domain Generalization
DA	Domain Adaptation
DAN	Deep Adaptation Network
DANN	Domain Adversarial Neural Network
DCGAN	Deep Convolutional Generative Adversarial Network
DG	Domain Generalization
EEG	Electroencephalography
ERM	Empirical Risk Minimization
FGSM	Fast Gradient Sign Method
FID	Fréchet Inception Distance
G2DM	Generalizing to unseen Domains via Distribution Matching
GAN	Generative Adversarial Network
GMAN	Generative Multi-Adversarial Networks
HV	Hypervolume maximization
IID	Independent and Identically Distributed
IRM	Invariant Risk Minimization
IS	Inception Score
LOSO	Leave-One-Subject-Out
LSGAN	Least-square Generative Adversarial Network
MDAN	Multi-source Domain Adversarial Network
MDMN	Multiple Domain Matching Network
MGD	Multiple gradient descent
MMD-AAE	Maximum Mean Discrepancy Adversarial Autoencoder
PGD	Projected Gradient Descent
RPODS	Randomly Projecting Out Distribution Shifts
SEED	SJTU Emotion EEG Dataset
SGD	Stochastic Gradient Descent
SNGAN	Spectral Normalized Generative Adversarial Network

TLA	Triplet Loss Adversarial Training
TRADES	TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization
WAUC	Workload Assessment Under physical aCtivity
WGAN	Wasserstein Generative Adversarial Network
WGAN-GP	Wasserstein Generative Adversarial Network with Gradient Penalty

Chapter 1

Introduction

Machine learning techniques have been shown to allow the automation of several tasks [1, 2, 3, 4] by learning from collections of data points that are related to the desired goal. In recent years, the rapid growth in the amount of available data made possible by the exploitation of tools such as Amazon Mechanical Turk and platforms such as social networks [5], facilitated the development of techniques that are capable of increasing the degrees of automation of learning systems[1]. Such big data, coupled with advances in hardware to run computer programs in parallel, has enabled the re-emergence of neural networks [8, 9]. Such learning systems showed to require minimal pre-processing of the inputs [8] and to achieve unprecedented levels of performance on tasks within a wide variety of application domains [10, 11, 12, 13].

The training of neural networks, and of many machine learning systems, is usually performed following the empirical risk minimization (ERM) setting [14]. The main assumption within this framework is that all examples used for training and testing predictors are independently drawn from a fixed distribution, i.e. the i.i.d. assumption. A number of generalization guarantees were derived upon this assumption, resulting in several supervised learning algorithms [15]. Despite the success of machine learning applications relying on the ERM framework, important limitations in this setting can be highlighted: i) the i.i.d. property is *unverifiable* [16] given that one does not have access to the data distribution, and ii) it doesn't account for distribution shifts, which often occur in practice. Representative examples of these distribution shifts include changes in data acquisition

[1] Please refer to the work of Birhane et al. [6, 7] and Crawford [5] for an in-depth study of the harms caused by such approaches for acquiring datasets and how they reinforce current power imbalance relationships and negatively affect underrepresented populations.

conditions, such as illumination in images for object segmentation, or new data sources such as unseen speakers when performing speech recognition.

A number of alternative settings were then introduced to better cope with more realistic cases where *out-of-distribution generalization*[2] is required. Risk minimization under the *domain adaptation* setting, for instance, relaxes part of the i.i.d. assumption by allowing a source distribution (or domain)[3] as well as a different target distribution observed at test time. The domain adaptation results introduced in [17] showed that the generalization gap in terms of risk difference across the two considered distributions for a fixed predictor is upper bounded by a notion of distance measured between the training and testing domains. While less restrictive than the previous setting, the domain adaptation case is still limited in that only pairs of distributions seen during training are expected to yield low risk, and shifts beyond those domains will likely induce poor performance. Moreover, algorithms devised for this setting often *rely on access at training time to an unlabeled sample from the target distribution* so that representations can be learned inducing invariance across train and target domains [18]. This is a limiting factor for practical applications where target domain data may be inaccessible; for example, a speech recognition service cannot be (re)trained on data obtained from every new speaker it observes.

Despite the success of domain adaptation strategies in several application scenarios [19, 20, 21], we take a step further from this setting and consider a more general framework which is often referred to in the literature as *domain generalization* [22, 23]. In this case, it is assumed that a set of distributions over the input space is available at training time. At test time, however, both observed distributions, as well as unseen novel domains might appear, and low risk should be obtained regardless of the underlying domain. Hence, *domain generalization strategies aim at finding a representation space that yields good performance on novel distributions, unknown at training time.*

In addition to natural distribution shifts which are accounted for by domain adaptation and domain generalization strategies, previous work identified that neural networks are also vulnerable to artificially, hand-crafted perturbations, known as adversarial examples [24]. Such type of perturbation is particularly harmful to safety-critical real-world applications, such as self-driving cars

[2] Generalizing "outside" of the training distributions, i.e., maintaining the performance level on unseen distributions close to the level observed on the available distributions at training time.

[3] We use the terms *domain*, *data distribution*, and *data source* interchangeably throughout the text.

[25]. Perturbations are designed to be imperceptible by humans, while making a learned model misclassify the attacked input [24] with very high confidence of its decision. Despite the awareness of the machine learning community about these limitations, the existing literature still lacks contributions that aim at simultaneously treating both distribution shifts and adversarial perturbations. Moreover, another limitation of current machine learning systems is their specificity to tackle a very particular setting and task [26]. For example, minimax formulations, where multiple subsets of the parameters of a model are learned by optimizing different loss functions, have been involved in the development of models to target several different tasks [27, 28, 29, 30]. However, few algorithms have considered the use of multi-objective training in order to build more “generic” models applicable across multiple such applications. This could have crucial engineering benefits in everyday practical applications.

As mentioned above, such vulnerable models could have severe outcomes in certain application domains. In this thesis, we aim to tackle these challenges and we place focus on two particular application domains of increasing importance and popularity with the machine learning community, namely brain-computer interfacing and computer vision.

1.1 Application domains addressed in the thesis

1.1.1 EEG-based Brain-Computer Interfaces

With recent innovations in wireless bioamplifiers and dry electrode technologies, portable electroencephalography (EEG) based applications are on the rise with applications across a wide range of domains, ranging from diagnostics to human-machine interaction. By recording and processing EEG signals, it is possible to translate neuronal activity and employ it to, for example, prosthetic control [31]. Systems capable of recording, processing, and decision-making based on neuronal information are called brain-computer interfaces (BCIs). In the past decades, the interest in EEG-based BCIs has grown massively due to its potential to positively impact [32] the lives of several people by, for example, allowing more engaging post-stroke rehabilitation [33]. Changes in the state of the brain can be inferred from EEG recordings and used in BCIs via extracting a number of features, including power spectral density [34, 35], coherence [36, 37, 38], and more recently, amplitude-modulation measures [39]. Recent work [40] has shown an increase in the application of neural networks to BCIs involving diverse sets of tasks from motor imagery to affective state prediction. Despite the observed success of neural networks in such applications, EEG-based BCIs lack generalizability between different subjects, or even between different recording sessions acquired from the same subject [41].

Anatomic and environmental factors are attributed as the main causes of the typical difference of neural responses across individuals under the same stimulus [42, 43, 44]. Additionally, such shifts between training and testing conditions could be due to different data collection equipment, as well as changes in the electrodes positioning during an experimental session. A standard way of handling the high cross-subject variability seen with EEG-based applications is to *calibrate* the model prior to applying it to an unseen individual. This is achieved by collecting a number of labelled examples from this particular subject and retraining the model considering this new sample [45]. However, recent work [46, 47] highlighted that the calibration step might be too costly and slow. Improving the cross-subject generalization of current BCIs is therefore critical to make it possible to apply such models in real-world conditions and high-impact applications such as mental workload monitoring. An alternative to calibrating BCIs prior to using it on a new subject/conditions, lies on employing strategies to learn models which are less prone to relying on subject-specific information. To achieve this, recent work has considered techniques to handle shifts between training and testing distributions, such as domain adaptation approaches [48, 49, 50, 51].

In this dissertation, we consider two main applications of BCIs, namely, mental workload assessment and affective state prediction. Monitoring mental workload in a fast and accurate manner is critical in scenarios where the full attention of an individual is fundamental for the security of others. Firefighters, air traffic controllers, and first responders, for instance, are constantly exposed to such work conditions. In many cases, in addition to a demanding mental task, individuals are under varying levels of physical strain. Measuring mental workload under such scenarios is challenging, especially when relying on wearable sensors [52]. EEG-based mental workload monitoring has been developed in the past via the use of brain-computer interfaces [53, 54], but the high cross-subject variability often observed in the features employed by such methods hinder their application in real-world scenarios. As pointed out in [55], models are usually subject-specific and present poor generalization when training and testing conditions are distinct in terms of the represented individuals.

Affective state prediction, in turn, can be employed to favor the development of human-driven technologies with positive impact, such as online learning platforms, healthcare, and rehabilitation of psychological disorders [51]. As EEG signals offer a direct, objective, a high-resolution alternative for assessing neurophysiological responses to emotional stimuli [56], EEG-based BCIs are frequently considered for monitoring human's affective state. Despite their recent success in applications considering the same group of subjects on training and testing, EEG-based BCIs for emotion recognition also present a decay in performance when utilized for assessing the state of unseen subjects during training time [57, 58, 59, 60] due to factors such as individualized previous experience [61, 60]. Recent work has attempted to address this issue by applying domain adaptation and domain generalization approaches to learn subject-invariant models on top of features, as well as to learn subject-invariant representations [62].

1.1.2 Object recognition

Despite their success on computer vision tasks when large-scale datasets are available [63, 64, 65, 66], neural networks present a decrease in performance when faced with distribution shifts when employed to real-world tasks [67]. In the case of object recognition tasks, for example, changes in low-level features such as texture and illumination are sufficient to confuse a model to a point where it is not capable of yielding reliable predictions [67, 68]. Previous work [69, 70, 71] has shown that neural networks can be biased towards capturing features that are not necessarily *causing* changes in the object class. As an example, consider a scenario where the goal is to classify dogs and cats from photos. In the case of distractors such as bone-shaped toys are present in the majority of the pictures from dogs, it might be the case where a neural network will learn to distinguish dogs from cats by looking for such objects in an image, rather than more representative and generalizable features such as the shape of the ears.

In addition to object recognition tasks, neural networks have also been found to struggle in applications in medical imaging where data collection conditions might differ from training to testing time. Domain shifts can be observed, for instance, once data from new subjects or equipment are employed after a model is trained [72, 73]. Since a major requirement of deployed models consists of them being able to generalize even to previously unobserved conditions, i.e., a tumor segmentation model cannot be retrained once data from a new patient is observed, generalizing across domains has become a relevant research direction.

1.2 Thesis contributions

In this Section, we provide an overview of the contributions within this dissertation, separated by chapter.

- Chapter 2: We focus on providing resources to allow for the development of different strategies to assess mental workload on real-world conditions. For that, we introduced WAUC, a multimodal database of mental Workload Assessment Under physical aCtivity. The study involved 48 participants who performed the NASA Revised Multi-Attribute Task Battery II under three different activity level conditions. Physical activity was manipulated by changing the speed of a stationary bike or a treadmill. During data collection, six neural and physiological modalities were recorded, namely: electroencephalography, electrocardiography, breathing rate, skin temperature, galvanic skin response, and blood volume pulse, in addition to 3-axis accelerometry. In order to bring our experimental setup closer to real-world situations, signals were monitored using wearable, off-the-shelf devices.
- Chapter 3: We propose a strategy to estimate two types of discrepancies between multiple data distributions, namely marginal and conditional shifts, observed on data collected from different subjects. Besides shedding light on the assumptions that hold for a particular dataset, the estimates of statistical shifts obtained with the proposed approach can be used to investigate other aspects of a machine learning pipeline, such as quantitatively assessing the effectiveness of domain adaptation strategies. In particular, in this Chapter, we consider the EEG recordings from the WAUC dataset (introduced in Chapter 2) distinguishing between individuals who performed the experiment while running on a treadmill or pedaling on a stationary bike. We explored the effects of different normalization strategies commonly used to mitigate cross-subject variability and showed the effects that different normalization schemes have on statistical shifts and their relationship with the accuracy of mental workload prediction as assessed on unseen participants at train time.
- Chapter 4: We tackle the out-of-distribution problem by focusing on the domain generalization setting: a formalization where the data generating process at test time may yield samples

from never-before-seen distributions. Our work builds on the following lemma: by minimizing a notion of discrepancy between all pairs from a set of given domains, we also minimize the discrepancy between any pairs of mixtures of domains. Using this result, we derive a generalization bound for our setting. We then show that low risk over unseen domains can be achieved by representing the data in a space where (i) the training distributions are indistinguishable, and (ii) relevant information for the task at hand is preserved. Minimizing the terms in our bound yields an adversarial formulation which estimates and minimizes pairwise discrepancies. We validate our proposed strategy on standard domain generalization benchmarks involving image classification tasks, outperforming a number of recently introduced methods. Notably, we tackle a real-world application where the underlying data corresponds to multi-channel electroencephalography time series from different subjects, each considered as a distinct domain.

- Chapter 5: We revisit the multiple-discriminator setting introduced in Chapter 4 to train Generative Adversarial Networks by framing the simultaneous minimization of losses provided by different models as a multi-objective optimization problem. Specifically, we evaluate the performance of multiple gradient descent and the hypervolume maximization algorithm on a number of different datasets. Moreover, we argue that the previously proposed methods and hypervolume maximization can all be seen as variations of multiple gradient descent in which the update direction can be computed efficiently. Our results indicate that hypervolume maximization presents a better compromise between sample quality and computational cost than previous methods.
- Chapter 6: We propose a unified and versatile approach to mitigate both natural and artificial distribution shifts via the use of random projections. We show that such projections, when implemented as convolutional layers with random weights placed at the input of a model, are capable of increasing the overlap between the different distributions that may appear at training/testing time. We evaluate the proposed approach in settings where different types of distribution shifts occur, and show it provides gains in terms of improved out-of-distribution generalization in the domain generalization setting, as well as increased robustness to two types of adversarial perturbations on the CIFAR-10 dataset without requiring adversarial training.

1.3 Publications derived from the thesis

Publications included in the thesis

- "WAUC: a multi-modal database for mental workload assessment under physical activity", **Albuquerque, I.**, Tiwari, A., Parent, M., Cassani, R., Gagnon, J. F., Lafond, D., and Falk, T. H. (2020, December), *Frontiers in Neuroscience, 2020* [74] [Chapter 2].

- "Cross-subject statistical shift estimation for generalized electroencephalography-based mental workload assessment", **Albuquerque, I.**, Monteiro, J., Rosanne, O., Tiwari, A., Gagnon, J. F., and Falk, T. H. (2019, October), *International Conference on Systems, Man and Cybernetics (SMC), 2019* [72] [Chapter 3].

- "Estimating Distribution Shifts for Predicting Cross-Subject Generalization in Electroencephalography based Mental Workload Assessment", **Albuquerque, I.**, Monteiro, J., Rosanne, O., and Falk, T. H. (2021, September), *Under review at the IEEE Transactions on Human-Machine Systems, 2021* [75] [Chapter 3].

- "Generalizing to unseen domains via distribution matching", **Albuquerque, I.**, Monteiro, J., Darvishi, M., Falk, T. and Mitliagkas, I. (2020, July), in *Uncertainty and Robustness in Deep Learning Workshop held at the 2020 International Conference on Machine Learning, 2020* [29] and (2021, September) *Under review at the IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021* [76] [Chapter 4].

- "Multi-objective training of generative adversarial networks with multiple discriminators", **Albuquerque, I.**, Monteiro, J., Doan, T., Considine, B., Falk, T. and Mitliagkas, I. (2019, July), *In International Conference on Machine Learning (pp. 202-211) PMLR, 2019* [77] [Chapter 5].

- "Randomly projecting out distribution shifts for improved robustness", **Albuquerque, I.**, Monteiro, J., and Falk, T. (2021, October), *Workshop on Distribution Shifts: Connecting Methods and Applications at the Conference on Neural Information Processing Systems (NeurIPS)* [78] [Chapter 6].

Other publications

- "Single-shot real-time compressed ultrahigh-speed imaging enabled by a snapshot-to-video autoencoder", Liu, X., Monteiro, J., **Albuquerque, I.**, Lai, Y., Jiang, C., Zhang, S., Falk, T., and Liang, J., Photonics Research, in press, 2021 [79].

- "Adaptive filtering for improved EEG-based mental workload assessment of ambulant users", Rosanne, O., **Albuquerque, I.**, Cassani, R., Gagnon, J. F., Tremblay, S. and Falk, T. H., (2021), *Frontiers in Neuroscience, 15, p.341* [80].

- "PASS: A Multimodal Database of Physical Activity and Stress for Mobile Passive Body/Brain-Computer Interface Research", Parent, M., **Albuquerque, I.**, Tiwari, A., Cassani, R., Gagnon, J. F., Lafond, D., and Falk, T. H. (2020), *Frontiers in Neuroscience, 14, 1274* [81].

- "Deep learning-based electroencephalography analysis: a systematic review", Roy, Y., Banville, H., **Albuquerque, I.**, Gramfort, A., Falk, T. H. and Faubert, J., (2019) *Journal of Neural Engineering, 16(5), p.051001* [40]

- "Multi-scale heart beat entropy measures for mental workload assessment of ambulant users", Tiwari, A., **Albuquerque, I.**, Parent, M., Gagnon, J. F., Lafond, D., Tremblay, S., and Falk, T. H. (2019), *Entropy, 21(8), 783*, [82].

- "On the Analysis of EEG Features for Mental Workload Assessment During Physical Activity", **Albuquerque, I.**, Tiwari, A., Gagnon, J. F., Lafond, D., Parent, M., Tremblay, S., and Falk, T. H. (2018, October), *In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 538-543). IEEE* [52].

- "Multimodal Assessment of Human Innovation Perception Based on Eye Tracking, Electroencephalography and Electrocardiography", **Albuquerque, I.**, Monteiro, J., Falk, T. H., Pavlovic, V., Ephrem, F. and Lucaci, D., (2018, May), *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-4). IEEE* [83].

- "Learning to navigate image manifolds induced by generative adversarial networks for unsupervised video generation", **Albuquerque, I.**, Monteiro, J. and Falk, T. H., (2018), *1st*

LatinX in Artificial Intelligence Workshop held at the 2018 Conference on Neural Information Processing Systems (NeurIPS) [84].

- “Improving out-of-distribution generalization via multi-task self-supervised pretraining”, **Albuquerque, I.**, Naik, N., Li, J., Keskar, N. and Socher, R., (2020), *Uncertainty and Robustness in Deep Learning Workshop held at the 2020 In International Conference on Machine Learning (ICML)* [85].

- “Fusion of spectral and spectro-temporal EEG features for mental workload assessment under different levels of physical activity”, **Albuquerque, I.**, Rosanne, O., Gagnon, J.F., Tremblay, S. and Falk, T.H., (2019, March), *In 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 311-314). IEEE* [86].

- “A Comparison of Two ECG Inter-beat Interval Measurement Methods for HRV-Based Mental Workload Prediction of Ambulant Users”, Tiwari, A., **Albuquerque, I.**, Parent, M., Gagnon, J. F., Lafond, D., Tremblay, S., and Falk, T. H. (2019), *CMBES Proceedings, 42* [87].

- “Mental Workload Assessment During Physical Activity Using Non-linear Movement Artefact Robust Electroencephalography Features”, Tiwari, A., **Albuquerque, I.**, Gagnon, J. F., Lafond, D., Parent, M., Tremblay, S., and Falk, T. H. (2019, October), *In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 4149-4154). IEEE* [88].

- “A Multimodal Approach to Improve the Robustness of Physiological Stress Prediction During Physical Activity”, Parent, M., Tiwari, A., **Albuquerque, I.**, Gagnon, J. F., Lafond, D., Tremblay, S., and Falk, T. H. (2019, October), *In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 4131-4136). IEEE* [89].

- “On-line adaptative curriculum learning for GANs”, Doan, T., Monteiro, J., **Albuquerque, I.**, Mazoure, B., Durand, A., Pineau, J. and Hjelm, R. D., (2019, July), *In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 3470-3477)* [90].

- “Generalizable adversarial examples detection based on bi-model decision mismatch”, Monteiro, J., **Albuquerque, I.**, Akhtar, Z. and Falk, T. H., (2019, October), *In 2019 IEEE*

International Conference on Systems, Man and Cybernetics (SMC) (pp. 2839-2844). IEEE [91].

- "Self-supervised representation learning from electroencephalography signals", Banville, H., **Albuquerque, I.**, Hyvärinen, A., Moffat, G., Engemann, D. A. and Gramfort, A., (2019, October), *In 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE* [92].

- "An end-to-end approach for the verification problem: learning the right distance", Monteiro, J., **Albuquerque, I.**, Alam, J., Hjelm, R. D. and Falk, T., (2020, November), *In International Conference on Machine Learning (pp. 7022-7033). PMLR* [93].

- "Performance comparison of automated EEG enhancement algorithms for mental workload assessment of ambulant users", Rosanne, O., **Albuquerque, I.**, Gagnon, J. F., Tremblay, S. and Falk, T. H., (2019, March), *In 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 61-64). IEEE* [94].

- "EEG coupling features: Towards mental workload measurement based on wearables", Drouin-Picaro, A., **Albuquerque, I.**, Gagnon, J.F., Lafond, D. and Falk, T. H., (2017, October), *In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 28-33). IEEE* [39].

- "AMA: An Open-source Amplitude Modulation Analysis Toolkit for Signal Processing Applications", Cassani, R., **Albuquerque, I.**, Monteiro, J. and Falk, T. H., (2019, November), *In 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 1-4). IEEE* [95].

1.4 Thesis organization

In this thesis, there are contributions to different aspects of the Machine Learning pipeline. We start by introducing a new dataset for EEG-based mental workload assessment in Chapter 2, along with the remaining background required for the following chapters. We then proceed to introduce our first algorithmic contribution on the realm of out-of-distribution generalization in Chapter 3 where we propose an approach to estimate distribution shifts given samples from different domains and validate it on the WAUC dataset introduced in Chapter 2. We proceed to our next contributions to improve out-of-distribution generalization on neural networks by presenting, in Chapter 4, new theoretical results for the domain generalization setting and introducing a new algorithm based on the these results that relies on the use of multiple discriminators and random projections implemented as convolutions with random weights. In Chapter 5, we show that the proposed approach for domain generalization can also be employed in other applications with similar learning objectives. We consider generative modeling of probability distributions and show that the algorithm can be used to train generative adversarial networks. We also take an in-depth exploration of the different strategies to aggregate the different loss functions involved in the approach and propose to consider gradient-based multi-objective optimization strategies for learning. We then present our final algorithmic contribution in Chapter 6, where we propose to use random convolutions, as in Chapters 4 and 5, to increase the robustness of neural networks to both natural and adversarial distribution shifts by proposing an approach that does not rely on domain labels nor requires adversarial training. Lastly, we conclude the thesis in Chapter 7 by summarizing our main findings and contributions and introducing the future directions of investigation derived from the work presented throughout this thesis.

Within the development of this thesis, contributions to different parts of the machine learning pipeline were achieved. We introduced a new dataset, tackled fundamental problems such as generative modeling and out-of-distribution generalization by proposing new algorithms, and for that, utilized techniques such as multiple discriminators and random projections. Moreover, we showed that the proposed techniques can be applied to real-world use cases, such as estimating the cross-subject generalization capability of a predictor, as well as devising calibration-free brain-computer interfaces.

Chapter	Out-of-distribution generalization	Multiple discriminators	Random projections	BCI application
2				✔
3	✔			✔
4	✔	✔	✔	✔
5		✔	✔	
6	✔	✔	✔	

Table 1.1 – Overview of the main topics considered in each chapter of the thesis.

Given that our contributions span different topics and aspects of the machine learning pipeline, we found relevant to highlight our contributions in a concise and brief manner according to the respective type of algorithmic innovation, application, and goal. For that, we present in Table 1.1 the main aspects related to each chapter considering the following categories: out-of-distribution generalization, use of multiple discriminators, use of random projections, and applications to brain-computer interfaces.

Chapter 2

Background

2.1 Preamble

This chapter is compiled from material extracted from the manuscript published in the journal *Frontiers in Neuroscience* [74].

2.2 Empirical Risk Minimization

2.2.1 Notation

Let the data be represented by $\mathcal{X} \subset \mathbb{R}^d$, where d corresponds to the dimension of the input (or feature) space, and $\mathcal{Y}$ denotes the label space. In this case, examples correspond to pairs $(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}$, such that $y = f(x)$, and $f : \mathcal{X} \to \mathcal{Y}$ is a deterministic labeling function. A domain is defined as a tuple $\langle \mathcal{D}, f \rangle$ where $\mathcal{D}$ corresponds to a probability distribution over $\mathcal{X}$[1] and f is the respective labeling function.

Let h be a hypothesis defined as a mapping $h : \mathcal{X} \to \mathcal{Y}$, such that $h \in \mathcal{H}$, where $\mathcal{H}$ is a set of candidate hypothesis, and finally define the risk R associated with a given hypothesis h on domain

[1] Notice that in case we consider a stochastic model where the labeling function is not deterministic, the distribution $\mathcal{D}$ would be defined over $\mathcal{X} \times \mathcal{Y}$.

$\langle \mathcal{D}, f \rangle$ as:

$$R[h] = \mathbb{E}_{x \sim \mathcal{D}} \mathcal{L}[h(x), f(x)], \tag{2.1}$$

where the loss $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to R_+$ quantifies how different a hypothesis h is from the true labeling function f for a given instance (x, y).

2.2.2 Supervised learning as Empirical Risk Minimization

The problem of supervised learning can be defined as finding the minimum risk hypothesis h^* within the class $\mathcal{H}$:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} R[h]. \tag{2.2}$$

However, computing $R[h]$ is generally intractable since one does not have access to $\mathcal{D}$. In practice, we only have available to compute h^* a set of observed samples from $\mathcal{D}$.

Given the intractability of the risk minimization setting described above, *empirical risk minimization* (ERM) is a common practical alternative framework for supervised learning. In such case, a sample X of size N is observed from $\mathcal{D}$, i.e. $X = \{x_1, x_2, \ldots, x_N\}$, **where all x_n, $n = \{1, \cdots, N\}$ are assumed to be independently sampled from the same domain $\mathcal{D}$.** This requirement is commonly referred to in the Machine Learning literature as the "independent and identically distributed" (i.i.d.) assumption.

The empirical risk of a hypothesis h computed on a sample X, $\hat{R}_X[h_X]$, is thus defined as:

$$\hat{R}_X[h_X] = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}[h_X(x_i), f(x_i)]. \tag{2.3}$$

$$\hat{R}_X[h] = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), f(x_i)). \tag{2.4}$$

The generalization error of h_X (or generalization gap) on the domain $\mathcal{D}$ is defined as the difference between the true and empirical risks:

$$\varepsilon = |R[h_X] - \hat{R}_X[h_X]|. \tag{2.5}$$

Ideally, $\hat{R}_X[h_X] \approx 0$ and $\varepsilon \approx 0$, in which case h_X is able to attain a low risk across samples of $\mathcal{D}$ that were not observed at training time.

2.3 Domain Adaptation

We now consider scenarios where the i.i.d. assumption does not hold and examples are not expected to be identically distributed. We introduce the domain adaptation setting where samples from the *source domain* $\langle \mathcal{D}_S, f_S \rangle$ are considered at training time, but at test time one expects samples to be drawn from a *target domain* $\langle \mathcal{D}_T, f_T \rangle$. The discussion in [96] established the theoretical foundations for studying cross-domain generalization properties for domain adaptation problems, and we now state results from the domain adaptation literature which are relevant for this thesis.

A bound for the risk of a hypothesis h on the target domain $R_T[h]$ was introduced [17]. This result shows that $R_T[h]$ depends on $R_S[h]$, the risk of h on the source domain, a notion of divergence between both domains, as well as the minimum risk that can be achieved by some $h \in \mathcal{H}$ on both $\mathcal{D}_S$ and $\mathcal{D}_T$. We restate this result in the following Theorem.

Theorem 1 (Theorem 2 from Ben-David et al. [17][2]): *Let $\mathcal{D}_S$ and $\mathcal{D}_T$ be the distributions of two domains over a shared feature space. The risk of any hypothesis $h \in \mathcal{H}$ on the target domain will be thus bounded by:*

$$R_T[h] \leq R_S[h] + d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] + \lambda, \tag{2.6}$$

where λ accounts for how "adaptable" the class $\mathcal{H}$ is and it is defined as the minimal total risk over both domains that can be achieved by some $h \in \mathcal{H}$:

$$\lambda = \min_{h \in \mathcal{H}} [R_S[h] + R_T[h]]. \tag{2.7}$$

The term $d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ corresponds to the $\mathcal{H}\Delta\mathcal{H}$-divergence introduced in [97] for a hypothesis class $\mathcal{H}\Delta\mathcal{H} = \{h(x) \oplus h'(x) | h, h' \in \mathcal{H}\}$, where $\oplus$ is the XOR operation. The $\mathcal{H}$-divergence between two distributions $\mathcal{D}_S$ and $\mathcal{D}_T$ is defined as:

$$d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] = 2 \sup_{\eta \in \mathcal{H}} |\mathrm{Pr}_{x \sim \mathcal{D}_S}[\eta(x) = 1] - \mathrm{Pr}_{x \sim \mathcal{D}_T}[\eta(x) = 1]|. \tag{2.8}$$

[2] More precisely, in [17] the authors introduced a finite sample bound for a hypothesis class of finite VC-dimension.

As discussed in [17], an estimate of $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ can be directly computed from the error of a binary classifier trained to distinguish samples from $\mathcal{D}_S$ and $\mathcal{D}_T$. Therefore, an estimate $\hat{d}_{\mathcal{H}}[\hat{\mathcal{D}}_S, \hat{\mathcal{D}}_T]$ of the $\mathcal{H}$-divergence can be obtained with unlabeled samples of size m, $\hat{\mathcal{D}}_S$ and $\hat{\mathcal{D}}_T$, from the source and target domains, respectively, via the following Equation:

$$\hat{d}_{\mathcal{H}}[\hat{\mathcal{D}}_S, \hat{\mathcal{D}}_T] = 2\left(1 - \min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{x:h(x)=0} I[x \in \hat{\mathcal{D}}_S] + \frac{1}{m} \sum_{x:h(x)=1} I[x \in \hat{\mathcal{D}}_T]\right]\right), \tag{2.9}$$

where $I[x \in \hat{\mathcal{D}}_S]$ (similarly for $\hat{\mathcal{D}}_T$) denotes the binary indicator function, i.e., $I(x) = 1$, if $x \in \hat{\mathcal{D}}_S$.

An extension of that result presented in Theorem 1 was introduced in [98] in order to replace λ by a term that explicitly accounts for a possible mismatch between the labeling functions of source and target domains. In order to achieve that, the divergence between source and target is computed over a hypothesis class $\tilde{\mathcal{H}}$ defined as $\tilde{\mathcal{H}} = \{sign(|h(x) - h'(x)| - t) | h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. We state this result in the following Theorem:

Theorem 2 (Theorem 4.1 from Zhao et al. [98]): *Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains, respectively. For any hypothesis class $\mathcal{H}$ such that $h \in \mathcal{H}$, $h : \mathcal{X} \to [0, 1]$, the following inequality holds:*

$$R_T[h] \leq R_S[h] + d_{\tilde{\mathcal{H}}}[\mathcal{D}_S, \mathcal{D}_T] + \min\{\mathbb{E}_{x \sim \mathcal{D}_S} \mathbb{1}[f_S(x) \neq f_T(x)], \mathbb{E}_{x \sim \mathcal{D}_T} \mathbb{1}[f_S(x) \neq f_T(x)]\}, \tag{2.10}$$

where $\min\{\mathbb{E}_{x \sim \mathcal{D}_S} \mathbb{1}[f_S(x) \neq f_T(x)], \mathbb{E}_{x \sim \mathcal{D}_T} \mathbb{1}[f_S(x) \neq f_T(x)]\}$ *accounts for the mismatch between the labeling functions.*

Zhao et al. [99] extended Theorem 1 for the multi-source domain adaptation setting where multiple source domains are available at training time:

Theorem 3 (Zhao et al. [99]): *Let $\mathcal{D}_S^i$, $i \in \{1, \ldots, N_S\}$, be the N_S source domains and $\mathcal{D}_T$ be the target domain. Then, $\forall \alpha \in \Delta_{N_S}$, where Δ_{N_S} is the N_S-th dimensional simplex* [3]*, we have for $h \in \mathcal{H}$ that:*

$$R_T[h] \leq \sum_{i=1}^{N_S} \alpha_i \left(R_{S_i}[h] + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left[\mathcal{D}_T, \mathcal{D}_{S_i}\right]\right) + \lambda_\alpha, \tag{2.11}$$

[3] i.e. $\alpha \in \mathbb{R}^{N_S}$ and $\sum_{i=1}^{N_S} \alpha_i = 1$ and $\alpha_i \geq 0, i \in \{1, \ldots N_S\}$

where λ_α is the risk of the optimal hypothesis on the mixture source domain $S_\alpha = \sum_{i=1}^{N_S} \alpha_i S_i$ and the target domain T:

$$\lambda_\alpha = \min_{h \in \mathcal{H}} [R_{S_\alpha}[h] + R_T[h]]. \tag{2.12}$$

2.4 Minimax formulations

In this thesis, we considered distinct applications (i.e., generative modeling and domain generalization) which can be all seen as instances of a setting where different parameters of a model are estimated by optimizing different objectives. In these cases, finding optimal values for the such groups of parameters θ and ϕ can be formulated as the following optimization problem:

$$\min_\theta \max_\phi \mathcal{L}(\theta, \phi), \tag{2.13}$$

where $\mathcal{L}$ corresponds to an objective function. This kind of optimization problem can be solved by finding a saddle point of $\mathcal{L}$ where it is not possible to simultaneously improve its value with respect to both θ and ϕ. This kind of problem frequently appears in machine learning applications such as fairness [100], actor-critic methods for reinforcement learning [101], robust optimization [102], generative modeling [27], and learning domain-invariant representations [28].

In this thesis, we make contributions to improve the optimization of a variation of the problem defined in Eq. 2.13 and consider its application to both generative modeling and learning domain-invariant representations for domain generalization. In the following Sections, we introduce the settings and approaches we built our contributions on top of.

2.4.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [27] offer a new approach to generative modeling, using game-theoretic training schemes to implicitly learn a probability density. GANs are generally composed of a discriminator model D with parameters θ, such that $D(x) : \mathbb{R}^d \to [0, 1]$, where d is the dimensionality of the input space, and a generator G with parameters ϕ, such that $G(z) : \mathbb{R}^m \to \mathbb{R}^n$, where m is the size of an input noise vector z. $D(x)$ receives a sample from the data distribution p_{data} or a sample from the generator $G(z)$, $z \sim p_z$. During training, its goal is to learn how to tell apart these two different types of inputs. The generator, on the other hand, aims at *fooling* the discriminator by learning how to produce samples as close to the data distribution as possible.

The training of a GAN was originally defined a zero-sum game such that the objective of the discriminator $\mathcal{L}_D$ is equivalent to minimize an objective $\mathcal{L}$ and the objective of the generator $\mathcal{L}_G$ is to maximize $\mathcal{L}$:

$$\min_{\theta} \max_{\phi} \mathcal{L}(\phi, \theta), \tag{2.14}$$

where

$$\mathcal{L} = -\mathbb{E}_{x \sim p_{\text{data}}} \log D(x) - \mathbb{E}_{z \sim p_z} \log(1 - D(G(z))), \tag{2.15}$$

and $\mathcal{L} = \mathcal{L}_D = -\mathcal{L}_G$.

In practice, however, an alternative formulation of the GAN game referred to as a non-saturating game [103] is more frequently used. According to this training scheme, the discriminator loss $\mathcal{L}_D$ and the generator loss $\mathcal{L}_G$ are respectively defined as:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{\text{data}}} \log D(x) - \mathbb{E}_{z \sim p_z} \log(1 - D(G(z))), \tag{2.16}$$

$$\mathcal{L}_G = -\mathbb{E}_{z \sim p_z} \log D(G(z)). \tag{2.17}$$

2.4.1.1 Evaluating GANs - Fréechet Inception Distance

In [104], authors proposed to use as a quality metric the squared Fréchet distance [105] between Gaussians defined by estimates of the first and second order moments of the outputs obtained through a forward pass in a pretrained classifier of both real and generated data. They proposed the use of Inception V3 [106] for computation of the data representation and called the metric Fréchet Inception Distance (FID), which is defined as:

$$\text{FID} = ||m_d - m_g||^2 + \text{Tr}(\Sigma_d + \Sigma_g - 2(\Sigma_d \Sigma_g)^{\frac{1}{2}}), \tag{2.18}$$

where m_d, Σ_d and m_g, Σ_g are estimates of the first and second order moments from the representations of real data distributions and generated data, respectively.

We employ FID throughout our experiments for comparison of different approaches. However, for each dataset in which FID was computed, the output layer of a pretrained classifier on that particular dataset was used instead of Inception. m_d and Σ_d were estimated on the complete test partitions, which are not used during training.

2.4.2 Domain Adversarial Neural Networks

Domain Adversarial Neural Networks (DANNs) were proposed in [28, 18] as an adversarial approach for unsupervised domain adaptation, a setting where an unlabeled sample of the target domain is available at training time, via learning domain-invariant representations.

A DANN consists of a model with parameters ϕ composed by an encoder E with parameters ϕ_E and a classifier C with parameters ϕ_C. The rationale behind DANN relies on the hypothesis that such a model can have its performance on the target domain improved if the encoder E maps examples from the input space to a lower-dimensional feature space in such a way that domain-relevant cues are discarded by the encoding process. In order to achieve that, the training objective of the encoder is augmented with term that accounts for the $\mathcal{H}$-divergence between the source and the target domain, and ϕ_E is updated in such a way that the estimated divergence between the domains is decreased. The H-divergence is estimated by a domain discriminator D with parameters θ which is trained to be a good predictor of domain labels.

Similarly to the GAN framework, in the case of DANNs, the encoder model (or "generator") aims at fooling the discriminator in order to learn representations which are *invariant* to domain information, and, in practice, this is achieved by updating ϕ in a direction that increases the domain discrimination loss $\mathcal{L}_D$. On the other hand, the domain discriminator's parameters θ are updated in a direction that decreases $\mathcal{L}_D$ so that this model can be a good estimator of the $\mathcal{H}$-divergence between source and target domains. The training of a DANN, therefore, consists in optimizing the following objective:

$$\min_{\phi_E,\phi_C} \max_{\theta} \mathcal{L}(\phi_E, \phi_C, \theta), \tag{2.19}$$

where $\mathcal{L} = \mathcal{L}_C - \mathcal{L}_D$ and $\mathcal{L}_C$ consists on a loss that accounts how well the model composed by the encoder and the classifier predicts the labels of the inputs from the source domain for a particular task of interest. In Figure 2.1 we illustrate the main components of a DANN.

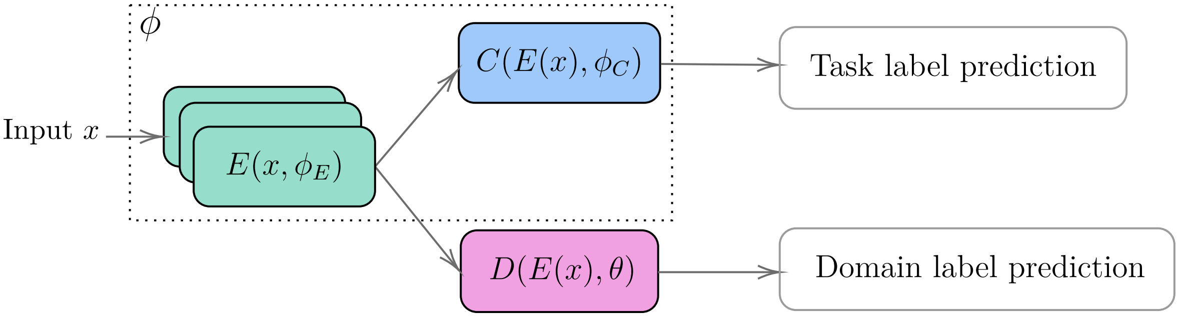


Figure 2.1 – Illustration of Domain Adversarial Neural Networks.

It is important to highlight that in the case of DANN, as well as any other approach for tackling problems under the unsupervised domain adaption setting, an unlabeled sample from the target domain is required at training time. Also, notice that DANN is trained in such a way that the encoder is adapted for a specific target domain.

2.5 Multi-objective optimization

In this section we provide some definitions regarding multi-objective optimization from prior literature which will be useful in the Chapter 5. Boldface notation is used to denote vector-valued functions. A multi-objective optimization problem is defined as [107]:

$$\min \mathbf{F}(x) = [f_1(x), f_2(x), ..., f_K(x)]^T,\ x \in \Omega \tag{2.20}$$

where K is the number of objectives, Ω is the variables space and $x = [x_1, x_2, ..., x_n]^T \in \Omega$ is a decision vector or possible solution to the problem. $\mathbf{F} : \Omega \to \mathbb{R}^K$ is a set of K-objective functions that maps the n-dimensional variables space to the K-dimensional objective space.

2.5.1 Pareto-dominance

Let x_1 and x_2 be two decision vectors. x_1 is said to dominate x_2 (denoted by $x_1 \prec x_2$) if and only if $f_i(x_1) \leq f_i(x_2)$ for all $i \in \{1, 2, \ldots, K\}$ and $f_j(x_1) < f_j(x_2)$ for some $j \in \{1, 2, \ldots, K\}$. In case a decision vector x is not dominated by any other vector in Ω, x is *a non-dominated solution.*

2.5.2 Pareto-optimality

A decision vector $x^* \in \Omega$ is said to be Pareto-optimal if and only if there is no $x \in \Omega$ such that $x \prec x^*$, i.e. x^* is a non-dominated solution. The Pareto-optimal Set (PS) is defined as the set of all Pareto-optimal solutions $x \in \Omega$, i.e., $\text{PS} = \{x \in \Omega \mid \text{x is Pareto optimal}\}$. The set of all objective vectors $\mathbf{F}(x)$ such that x is Pareto-optimal is called Pareto front (PF), that is $PF = \{\mathbf{F}(x) \in \mathbb{R}^K \mid x \in PS\}$.

2.5.3 Pareto-stationarity

Pareto-stationarity is a necessary condition for Pareto-optimality. For f_k differentiable everywhere for all k, $\mathbf{F}$ is Pareto-stationary at x if there exists a set of scalars $\alpha_k, k \in \{1, \ldots, K\}$, such that:

$$\sum_{k=1}^{K} \alpha_k \nabla f_k = \mathbf{0}, \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0 \quad \forall k. \tag{2.21}$$

2.5.4 Multiple Gradient Descent

Multiple gradient descent (MGD) [108, 109, 110, 111] was proposed for the unconstrained case of multi-objective optimization of $\mathbf{F}(x)$ assuming convex, continuously differentiable and smooth $f_k(x)$, $\forall k = \{1, \ldots, K\}$. For every iteration, MGD aims at finding a direction which will simultaneously maximally decrease all objectives. In [108], it was shown this can be achieved by finding a direction w^* such that the directional derivative towards w^* is negative for all objectives, i.e., $w^* \cdot \nabla f_k(x) \leq 0$, $\forall k = \{1, \ldots, K\}$. In the case of MGD, the problem of finding such common descent direction is solved by defining the convex hull of all $\nabla f_k(x)$, and finding the minimum norm element within it. Consider w^* given by:

$$\begin{aligned} w^* = \text{argmin}||w||^2, &\quad w = \sum_{k=1}^{K} \alpha_k \nabla f_k(x), \\ \text{s.t.} \quad \sum_{k=1}^{K} \alpha_k = 1, &\quad \alpha_k \geq 0 \quad \forall k. \end{aligned} \tag{2.22}$$

w^* will be either $\mathbf{0}$ in which case x is a Pareto-stationary point, or $w^* \neq \mathbf{0}$ and then w^* is a descent direction for all $f_i(x)$.

Similarly to gradient descent, MGD consists in finding the *common* steepest descent direction w_t^* at each iteration t, and then updating parameters with a learning rate λ according to:

$$x_{t+1} = x_t - \lambda \frac{w_t^*}{||w_t^*||}. \tag{2.23}$$

2.5.5 Hypervolume maximization (HV)

Let S be the set of solutions for a multi-objective optimization problem. The hypervolume H of S is defined as [112]:

$$H(S) = \mu(\cup_{x \in S}[\mathbf{F}(x), \boldsymbol{\eta}^*]), \tag{2.24}$$

where μ is the Lebesgue measure and $\boldsymbol{\eta}^*$ is a point dominated by all $x \in S$ (i.e., $f_i(x)$ is upper-bounded by η), referred to as the *nadir point*. $H(S)$ can be understood as the size of the space covered by $\{\mathbf{F}(x) \mid x \in S\}$ [113].

The hypervolume was originally introduced as a quantitative metric for coverage and convergence of Pareto-optimal fronts obtained through population-based algorithms [114]. Methods based on direct maximization of H exhibit favorable convergence even in challenging scenarios, such as simultaneous minimization of 50 objectives [113]. In the context of Machine Learning, single-solution HV has been applied to neural networks as a surrogate loss for mean squared error [115], i.e. the loss provided by each example in a training batch is treated as a single cost and the multi-objective approach aims to minimize costs over all examples. Authors show that such method provides an inexpensive boosting-like training.

2.6 Datasets

In this Section we introduced the datasets used throughout this thesis. We start by describing the image datasets we used in object recognition tasks involving distribution shifts in Chapters 4 and 6. Next, we introduce the EEG datasets utilized in Chapters 3 and 4. In addition to the datasets described in this Section, we also employed throughout our experiments standard datasets

Table 2.1 – Number of examples per domain of the PACS dataset.

	Number of examples
Photos	1670
Art painting	2048
Cartoon	2344
Sketch	3929

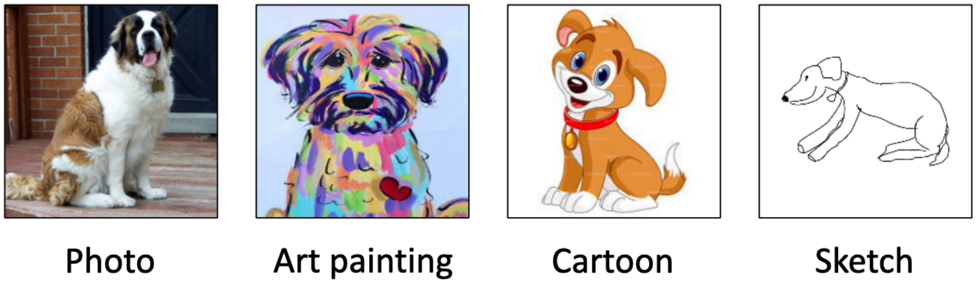


Figure 2.2 – Examples of images labeled as "dog" from each domain of the PACS dataset.

in the Machine Learning literature, such as MNIST [116], CIFAR-10 [117], ImageNet [118], CelebA [119], and Cats[4].

It is important to highlight that, besides using datasets that were already available in the literature, during this work we also collected a multi-modal database of psychophysiological recordings from 48 subjects. In this thesis we utilized the EEG recordings from this database in Chapter 3 and describe it in Section 2.6.3. For a complete description and analysis of all the collected modalities, the interested reader is referred to [74].

2.6.1 PACS

The PACS dataset was introduced with the aim of evaluating machine learning algorithms under distribution shifts [120]. The dataset is composed by 224×224 images corresponding to four different data sources, namely, (P)hotos, (A)rt Paintings, (C)artoon, and (S)ketches. In total, this dataset contains 9991 labeled examples divided into seven classes: dog, elephant, giraffe, guitar, horse, house, and person. In Table 2.1 we show the number of examples per domain and in Figure 2.2 we show examples of images labeled as "dog" from each one of the four domains. We use the original train/validation partitions provided by the dataset authors [120] in our experiments.

[4] https://www.kaggle.com/crawford/cat-dataset

Table 2.2 – Number of examples per domain of the VLCS dataset.

	Number of examples
Pascal VOC2007	3376
LabelMe	2656
Caltech-101	1415
SUN09	3282

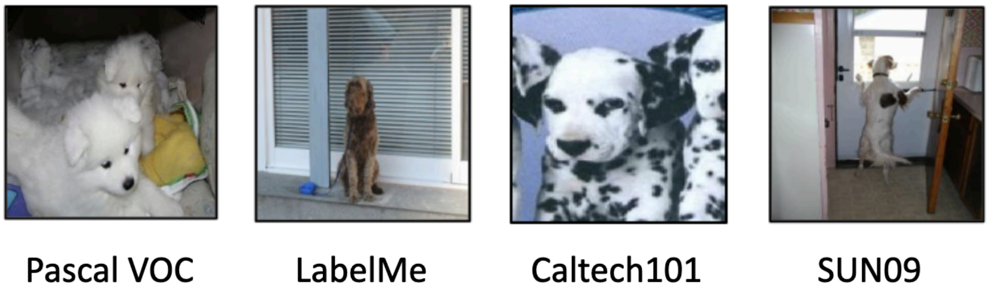


Figure 2.3 – Examples of images labeled as "dog" from each domain of the VLCS dataset.

2.6.2 VLCS

Similarly to PACS, the VLCS dataset [121] was also proposed to evaluate the ability of learning machines to generalize out-of-distribution. This dataset is composed by natural images collected from the following datasets proposed by the computer vision community: PASCAL VOC [122], LabelMe [123], Caltech101 [124], and SUN09 [125]. In total, VLSC contains 10729 images of size 224×224 which are divided into five classes, namely, bird, car, chair, dog, and person. In Table 2.2 we show the number of examples per domain and in Figure 2.3 we show examples of images labeled as "dog" from each one of the four domains. Following the common practice in the literature, we split each dataset into training and validation sets containing 80% and 20% of the data points from each domain, respectively.

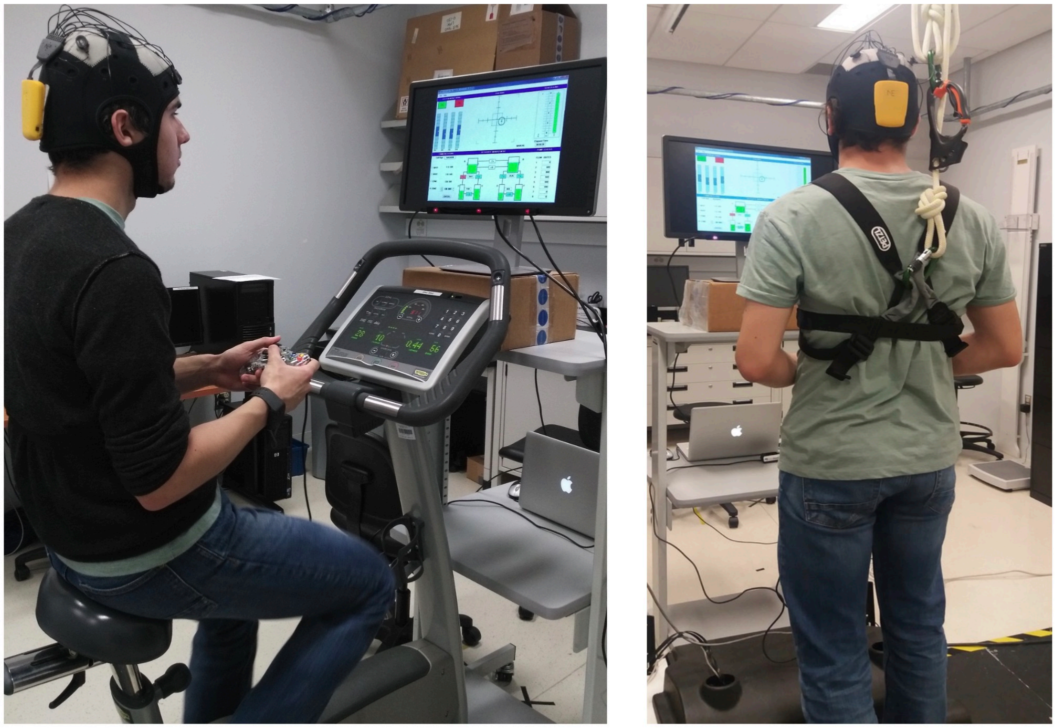
2.6.3 WAUC

Assessment of mental workload is crucial for applications that require sustained attention and where conditions such as mental fatigue and drowsiness must be avoided. Previous work that attempted to devise objective methods to model mental workload were mainly based on neurological or physiological data collected when the participants performed tasks that did not involve physical activity. While such models may be useful for scenarios that involve static operators, they may not apply in real-world situations where operators are performing tasks under varying levels of physical activity, such as those faced by first responders, firefighters, and police officers. In this Section, we introduce the multimodal database of mental Workload Assessment Under physical aCtivity (WAUC) collected with the aim of decreasing the gap between current research on mental workload assessment based on psychophysiological signals and real-world applications, as well as providing resources for allowing the development of new strategies for mental workload monitoring.

2.6.3.1 Participants

As the experimental protocol involved sustained physical and mental strain for a considerable period of time, recruited subjects were submitted to a pre-screening process in order to prevent any potential risk during the data collection. Hence, candidates with cardiovascular diseases, neurological disorders, history of feeling dizzy or fainting were not considered for the experiment. After the screening process, 4 participants were discarded and 48 were selected. Based on self-identified gender and the assigned physical activity modality (i.e., bike or treadmill) used during the experiment, a total of 22 participants used the treadmill (9 male, 13 female) and 26 performed the experiment using the bike (16 male, 10 female). The average age among the participants was 27.4 ± 6.6 years old. In order to avoid gender bias in our dataset, we intended to have a close number of male and female subjects, however, no candidate was rejected or accepted to participate in our experiment due to gender-related reasons. All participants consented to participating in the study and were remunerated (10 CAD/hour) for the time they spent at the experiment facility. The experimental protocol was approved by the Ethics Review Boards of INRS, Université Laval and the PERFORM Centre (Concordia University), the latter being the location in which data was collected.

Prior to arriving at the experiment facility, participants were advised to wear comfortable sportswear, and to not drink caffeinated beverages for at least 2 hours prior to the beginning of the data collection. Before starting the task tutorial, participants were asked to read and sign (in case of agreement) a consent form containing a brief description of the goals of our project and allowing the use and sharing of the collected data for research purposes.



(a) (b)

Figure 2.4 – Experimental set-up illustration for (a) bike and (b) treadmill sessions.

2.6.3.2 Experimental protocol

The experimental protocol aimed at simultaneously modulating mental and physical workload. Participants executed mental tasks while performing physical activity. A full factorial (2 mental workload × 3 Physical strain) design was employed to capture main effects and interactions. The data collection protocol was preceded by a tutorial to make the participants familiar with the tasks. The tutorial consisted in slides presentation to explain the experimental procedure and the tasks to be executed. Subjects were allowed to take as much time as necessary to go through the tutorial and to ask the experimenters as many questions as needed. After ensuring the participant understood the tasks to be performed, the next step involved donning the devices. To guarantee participants' safety during the experiment, a safety harness was placed at the participant's chest following the devices placement step mentioned above. This was only the case for the participants assigned to the treadmill task. For those assigned to the stationary bike, they were asked to adjust the seat according to their preference. In all cases, the height of the screen was adjusted lastly according to participants preferences. Figures 2.4-a and 2.4-b illustrate the experimental layout for the bike and treadmill, respectively, once all devices and safety features are in place. Before starting the data collection, each subject performed a practice session that corresponded to playing the Multi-task Attribute Battery II (MATB-II) (*c.f.* Section 2.6.3.3) for 10 minutes. While subjects were practicing, the experimenter observed whether they were capable of correctly performing each task.

Three levels of physical activity were considered: no movement, medium (treadmill: 3 km/h, bike: 50 rpm), and high movement (treadmill: 5 km/h, bike: 70 rpm). Since in the case of the stationary bike it was not possible to set the physical activity level for a fixed value during the experiment, we leveraged the training phase prior to each experimental section to let each participant get used to the speeds required during the data collection. Moreover, during each trial, the experimenter monitored whether the participant was deviating more than 5 rpm from the required speed and alerted the participant in case it did.

With respect to the mental workload levels elicited by MATB-II, two levels were considered, namely, low and high mental workload according to the task difficulty. In total, six possible combinations of joint mental workload and physical activity levels were tested. The experiment was then split into 6 sessions, each one corresponding to one of the six combinations previously described. The order in which each session was executed was counterbalanced among all the participants in order to avoid any ordering biases.

Before each session, data corresponding to two baseline periods was collected. During the first baseline, there were neither physical or mental activity. Participants were asked to stand still and relax during 60 seconds. Following this relaxation period, the second baseline was recorded where the subject was asked to start moving according to the corresponding physical activity level assigned to the current session, but without at mental workload manipulation. Recordings of the second baseline period only began once the activity level reached a stable period and the recording then lasted for 2 minutes. Lastly, the experimenter gave the joystick to the participant and the 10-minute session of combined mental physical effort started. After each task, a 5-minute break was given. During this resting period, participants were asked to perform a subjective evaluation corresponding to the past task by filling the NASA-TLX questionnaire [126]. They also reported their perceived fatigue level based on the Borg scale [127]. Overall, the duration of each experimental session comprising the baselines, task, and subjective evaluation was 18 minutes, and the complete experimental protocol lasted roughly two hours. Figure 2.5 summarizes the entire experiment and shows the duration in minutes corresponding to each part of a complete session.

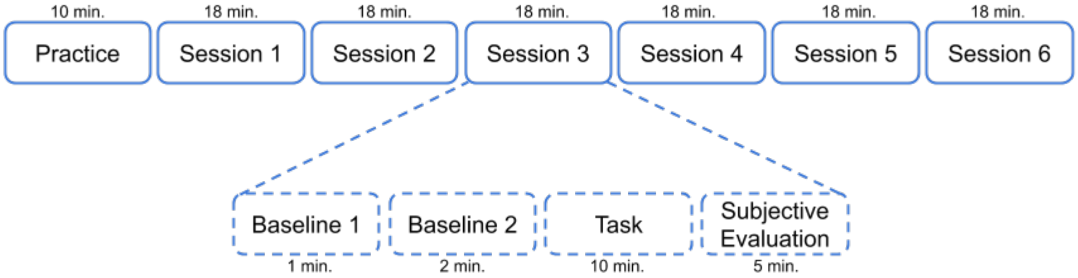
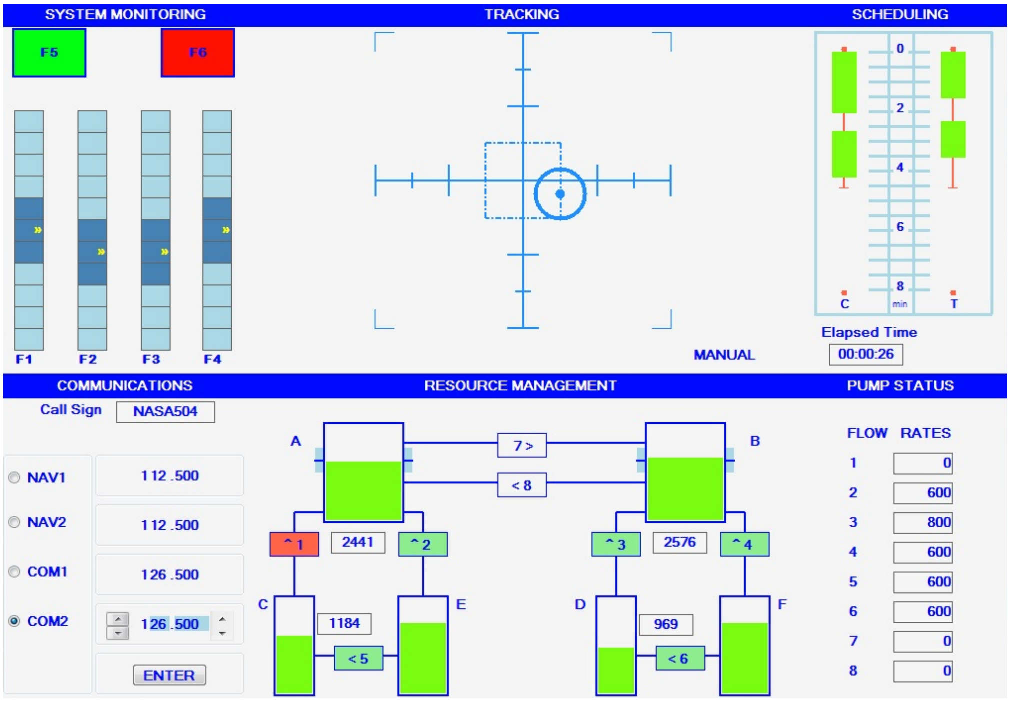


Figure 2.5 – Schematic of the steps executed by a participant during the experiment.



Elapsed Time 00:00:26

Call Sign NASA504

○ NAV1 112.500

○ NAV2 112.500

○ COM1 126.500

● COM2 126.500

FLOW	RATES
1	0
2	600
3	800
4	600
5	600
6	600
7	0
8	0

Figure 2.6 – Illustration of the MATB-II interface.

2.6.3.3 Stimuli

The Multi-task Attribute Battery II (MATB-II) [128] was employed to modulate the mental workload level on the participants. This set of tasks was originally devised to simulate different activities that need to be performed by an aircraft pilot. In our experiment, different mental strain levels are elicited by requiring the subjects to simultaneously perform three of the (four available) tasks involved in MATB-II, namely system monitoring, tracking, and resource management. Figure 2.6 shows a screenshot of the MATB-II interface, as seen by the participant. Note the top-right part of the screen was not used for the purposes of this study. An Xbox 360 controller was used to perform the three concurrent activities.

The system monitoring task (see top-left part of Fig. 2.6) requires the participant to monitor four sliders and report deviations from their normal state. The two warning lights (seen as F5 and F6 in the figure) were not used in this study. In their normal states, sliders oscillate around the center position. In their deviation state, sliders start oscillating around the top or the bottom of the panel. Participants had to use the directional pad of the controller to report deviations (one direction was assigned to each slider). When reported, the concerned slider reverted to its normal state. In case the deviated sliders were not reported within 10 seconds, they were reverted to their normal state and a false alarm was recorded.

The tracking task (top-middle part of Fig. 2.6), in turn, requires the participant to keep a target (a circular aim) within a square bounding box. As the trials progressed, the target started to move randomly. Participants had to use the joystick part of their controller to bring the target back near the center of the square. Lastly, the resource management task (bottom-centre part of Fig. 2.6) simulates the control of fuel reservoirs. Participants are asked to control pumps (which are subject to failure during the task) to transfer fuel across six reservoirs in order to keep the content levels of two main tanks (A and B) below a certain threshold. In particular, they were instructed to keep the level of the main tanks as close as possible to 2500 units (this level is indicated by ticks on the sides of tanks A and B). However, fuel gradually depleted from tanks A and B. To keep the tanks at the aimed level, participants could use eight pumps (labeled 1 to 8) to transfer fuel between the reservoirs. To activate pumps, participants had to use the second joystick of the controller to move the cursor and “click” on the pumps. When turned on, the pump would turn green. Pumps were configured to fail from time to time. When a pump failed, it turned red and was disabled. Pumps were automatically enabled for use after a while and the participant could resume using it if needed.

Modulation of the mental workload level relied on changing parameters in MATB-II. For example, for low mental workload cases, sliding bars speed, aim speed, volume of fuel in the reservoirs, and failure rate of the pumps were set to lower values. In the case of high mental workload, on the other hand, those parameters were set to larger values.

EEG data was collected using the 8-channel Neurolectrics Enobio portable headset [129]. The acquisition sampling rate was set to 500Hz. Electrode positions according to the 10-20 system were P3, T9, AF7, FP1, FP2, AF8, T10, and P4. References were placed at Fpz and Nz. Since our study involved physical activity, we decided to use wet electrodes on the regions that would be likely affected by sweat during the experiments to avoid signal quality issues [130]. Therefore, frontal and temporal regions were monitored using wet electrodes, while dry electrodes were used in the parietal region. Figures 2.4-a and 2.4-b illustrate Enobio's placement on the participant's head during the experiment.

Although we only considered EEG signals in our investigations, the WAUC dataset comprises further modalities such as skin temperature, acceleration, and electrocardiography. In addition to Enobio, two other wireless wearable devices were employed to acquire data and the open-source software MuLES [131] was utilized in order to allow simultaneous and synchronized acquisition of data streams from multiple devices. MuLES was also used to generate the synchronized markers indicating the beginning and the end of each phase of the experimental protocol.

2.6.4 SEED

The SJTU Emotion EEG Dataset (SEED) [132] was proposed with the aim of enabling cross-subject and cross-session assessment of human's affective state via EEG-based brain-computer interfaces. The dataset contains EEG recordings from 15 subjects, collected across three different experimental sessions. Data was acquired with a 62-channel (according to the international 10-20 system) ESI NeuroScan System at a sampling rate of 1000 Hz.

Each experimental session was composed of 15 parts, in which subjects watched a 4-minute long clip extracted from a movie with the aim of eliciting one out of three emotions (positive / negative / neutral). Subjects were asked to fill a questionnaire to self-evaluated the emotions elicited after watching each a movie clip. In total, each subject performed 3 experimental sessions.

2.7 Conclusion

In this Chapter, we presented the main definitions, settings, datasets, and previous results employed in the following Chapters of this thesis. We introduced the empirical risk minimization setting, along with the domain adaptation and multi-source domain adaptation frameworks that will be referred to especially in Chapters 3, 4, 6. This Chapter also contains in Section 2.4 a brief introduction to a standard optimization problem in Machine Learning that appears in Chapters 4 and 5 and connects two of the major contributions this thesis for the domain generalization setting and generative modeling. We also introduced basic definitions and algorithms in the realm of multi-objective optimization which are core components of the content presented in Chapters 4 and 5. Finally, we introduced the main datasets used throughout our experiments, giving a particular emphasis to the WAUC dataset which is employed in Chapter 3 and was collected by authors of the thesis and inspired the subsequent contributions to out-of-distribution generalization.

Chapter 3

Characterizing distribution shifts on EEG data

3.1 Preamble

This Chapter is compiled from material extracted from the manuscript presented at the System, Man, and Cybernetics Conference [72] and the manuscript under review at the IEEE Transactions on Human-Machine Systems [75].

3.2 Introduction

An alternative strategy to calibrating BCIs to unseen subjects/conditions is to develop methods that reduce the variability between training and testing conditions. To this end, approaches such as domain adaptation have been proposed [48, 49]. A standard domain adaptation strategy corresponds to augmenting the learning objective of an algorithm with a term that accounts for how *invariant* the current model is with respect to data from different distributions [18, 133]. The goal of this regularization term is to enforce the learned model to ignore domain-specific cues.

Previous work on domain adaptation has shown that different techniques rely on distinct assumptions over the training and testing distributions [134, 98]. For example, a common requirement is the *covariate shift* assumption, which considers that the distributions of labels y conditioned on data x, $p(y|x)$, do not shift across training and testing conditions and only the marginal distributions $p(x)$ shift [134]. In the case of EEG-based passive BCI applications, however, previous work has argued that $p(y|x)$ is likely to shift between different subjects [135, 44, 43, 42]. Therefore, the covariate shift assumption cannot be taken for granted since, given feature vectors x_1 and x_2 respectively acquired from two distinct subjects and represented in a shared feature space, $p_1(y|x_1) \neq p_2(y|x_2)$ even in the case where $x_1 = x_2$. As discussed in [98, 136], when the covariate shift assumption does not hold, there is a trade-off between learning domain-invariant representations and obtaining a small prediction error across different domains that needs to be optimized.

Verifying whether the underlying assumptions of a particular approach hold in practice is a frequently overlooked step by domain adaptation approaches [137]. In this Chapter, we claim that it is necessary to evaluate the underlying structure of a particular dataset in order to verify which types of distribution shifts exist and which assumptions could be safely considered (or not), when utilizing domain adaptation strategies. To this end, our main contributions are: (i) We introduce a method to estimate the cross-subject mismatch between the conditional label distributions; (ii) We apply a notion of divergence introduced in [97] to estimate the mismatch between marginal distributions of pairs of subjects; (iii) We investigate whether common practices in the EEG literature to handle cross-subject variability, such as normalizing spectral features, are able to mitigate both conditional and marginal distributional shifts. Given the relevance of decreasing cross-subject variability on EEG-based mental workload assessment, we empirically validate our proposed method on the WAUC dataset, previously introduced in Chapter 2 [74]. We considered a subset of this dataset comprised of 18 subjects and investigated how different ways to modulate physical activity affect the cross-subject statistical shifts on EEG correlates of mental workload.

The remainder of this Chapter is organized as follows: in Section 3.3 we formalize the problem of generalizing across subjects under the domain adaptation. In Section 3.4, the proposed strategies to estimate conditional and marginal shifts are presented. In Sections 3.5 and 3.6, we describe the experimental setup and present the results, respectively. Finally, we outline the main conclusions in Section 3.7.

3.3 Cross-subject generalization and Domain Adaptation

In light of Theorems 1 and 2 from [17] and [98], respectively, and presented in Chapter 2, it is possible to identify the the two main aspects that determine how well a hypothesis h generalizes from the source to the target domain. The first quantity that affects such generalization capability is the divergence between domains measured in the feature space, i.e., in order to allow cross domain generalization the input space $\mathcal{X}$ must be such that the $\mathcal{H}$-divergence $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ measured between the marginal distributions is low.

In addition to a divergence between the marginal distributions of source and target domains, the mismatch between labeling functions should be taken into account. Theorem 2 provides a way to account for this discrepancy in case it exists via the following quantity:

$$\min\{\mathbb{E}_{x\sim\mathcal{D}_S}\mathbb{1}[f_S(x) \neq f_T(x)], \mathbb{E}_{x\sim\mathcal{D}_T}\mathbb{1}[f_S(x) \neq f_T(x)]\}. \tag{3.1}$$

A feature space $\mathcal{X}$ that enables good transferability from the source to the target domain must be one such that the above quantity is as low as possible.

Previous work on domain adaptation (e.g., [18]) has mostly focused on mitigating the mismatch between marginal distribution and assumed that labeling functions were the same across domains. However, when this is not case, decreasing the discrepancy between marginal distributions [98] or adding more data [138] might actually hurt the performance of a model on the target domain.

3.3.1 Formalizing cross-subject generalization

In this work, we formalize the problem of learning passive BCIs (i.e., BCIs that monitor implicit user mental states) that generalize across subjects under the domain adaptation setting. For that, consider a dataset with a total of M subjects and that each subject is associated with domain $\mathcal{D}_i$ and labeling function f_i, $\forall\, i = \{1, \cdots, M\}$. Without loss of generality, assume that recordings from the first $M-1$ subjects are available at training time and we are interested in predicting how well a hypothesis $h \in \mathcal{H}$ would perform in the M-th subject, which was not considered at training time. Let $\mathcal{D}_S = \bigcup_{i=1}^{M-1} \mathcal{D}_i$ be the source domain defined as the union of the domains corresponding to the training subjects. Taking into consideration Theorem 2, (Equation 2.10), we can bound the risk on the M-th unseen subject, $R_M[h]$ as

$$R_M[h] \leq R_S[h] + d_{\tilde{\mathcal{H}}}[\mathcal{D}_S, \mathcal{D}_M] + \min\{\mathbb{E}_{x\sim\mathcal{D}_S}\mathbb{1}[f_S(x) \neq f_M(x)], \mathbb{E}_{x\sim\mathcal{D}_M}\mathbb{1}[f_S(x) \neq f_M(x)]\}. \quad (3.2)$$

In practice, we aggregate the available test samples from all the training subjects to estimate the risk of h in the source domain $R_S[h] = \sum_{i=1}^{M-1} R_i[h]$. However, there is no such straightforward way of estimating the two remaining terms of the bound. In the next Section, a strategy to compute these two terms is proposed.

3.4 Practical approaches to estimate distribution shifts

In this Section, we provide practical strategies to estimate both conditional and marginal shifts for the case where multiple domains (subjects) are available. Quantifying such mismatch will enable us to:

- Shed light on which domain adaptation strategies should be used for a given scenario by verifying whether, for example, the covariate shift assumption holds.
- As these quantities are related to how well a particular hypothesis will perform on unseen subjects, we can use their estimates computed considering different feature spaces and infer which one would achieve better performance on unseen subjects.

3.4.1 Conditional shift

A conditional shift is observed across subjects when the labeling function (or, in the stochastic case, the conditional distribution of the labels given the input features) differ among the subjects, i.e., for M subjects, we have $f_i(x) \neq f_j(x)$, $\forall i, j = \{1, \cdots, M\}$. In order to characterize the cross-subject conditional shift of a dataset of M subjects, we consider the following quantity on the generalization bound presented in Theorem 2 for all pairs of subjects:

$$\min\{\mathbb{E}_{\mathcal{D}_i}[|f_i - f_j|], \mathbb{E}_{\mathcal{D}_j}[|f_j - f_i|]\}, \quad (3.3)$$

where $i, j = \{1, \cdots, M\}$. In practice, it is not possible to compute such quantity as one does not have access to the true labeling functions and computing the expectations in Eq. 3.3 is intractable.

We thus propose to estimate such values by learning a labeling rule for each one of the domains, and account for how well it classifies examples from the other domain. Assuming that we are able to learn a good predictor for the labels of each domain, such approach is capable of accounting for how "close" the true labeling functions of different domains are. In practice, we consider that two labeled samples of size N from domains i and j are available and compute the following estimator $\mu_{i,j}$ for the quantity $\mathbb{E}_{\mathcal{D}_i}[||f_i - f_j||]$:

$$\mu_{i,j} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}[f_i(x_n^i) \neq \tilde{f}_j(x_n^i)], \tag{3.4}$$

where $(x_n^i, y_n^i) \sim \mathcal{D}_i$, and $\tilde{f}_j$ is an approximated labeling function for the j-th subject. We decided to have as $\tilde{f}_j$ a non-parametric decision procedure based on the Euclidean distance between data points in a fixed feature space. For that, we use a k-nearest neighbor (k-NN) labeling function, i.e., a k-NN binary classifier trained on $\mathcal{D}_j$ to classify as low or high mental workload condition data sampled from $\mathcal{D}_i$. Based on $\mu_{i,j}$ and $\mu_{j,i}$ we estimate the value $d_{i,j} = d_{j,i} = \min\{\mu_{i,j}, \mu_{j,i}\}$ and compose a Hermitian (elements symmetric with respect to the main diagonal are equal) disparity matrix D defined as:

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,M} \\ d_{2,1} & d_{2,2} & \dots & d_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \dots & d_{M,M} \end{bmatrix}. \tag{3.5}$$

Notice that in the case we obtain optimal approximate labeling functions, i.e., $f_i(x_n^i) = \tilde{f}_j(x_n^i)$, $\forall i = j$, the trace of D is equal to 0. Finally, in order to obtain a single value representing the conditional shift of all subjects in a dataset, we aggregate the values of pairwise conditional shifts. For that, we compute the Frobenius norm $||.||_F$ of the disparity matrix D:

$$||D||_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} |d_{i,j}|^2}. \tag{3.6}$$

The resulting $||D||_F$ is then rescaled to the $[0, 1]$ interval to allow for easier comparison across feature spaces.

3.4.2 Marginal shift

The $\mathcal{H}$-divergence between two distributions $\mathcal{D}_S$ and $\mathcal{D}_T$ is defined as:

$$d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] = 2 \sup_{\eta \in \mathcal{H}} |\mathrm{Pr}_{x \sim \mathcal{D}_S}[\eta(x) = 1] - \mathrm{Pr}_{x \sim \mathcal{D}_T}[\eta(x) = 1]|. \tag{3.7}$$

As discussed in [17], $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ can be estimated from the error ϵ of a binary classifier trained to distinguish samples from $\mathcal{D}_S$ and $\mathcal{D}_T$. The lower ϵ is, the highest the estimate of $d_{\mathcal{H}}$ will be, since in this case, there is a hypothesis η capable of distinguishing between $\mathcal{D}_S$ and $\mathcal{D}_T$ with high accuracy. Notice that the $\mathcal{H}$-divergence only accounts for discrepancies between the marginal distributions of the domains, not accounting for how each data point is labeled. Therefore, it is not necessary to have access to labeled samples from the considered domains to estimate its value.

Our proposed approach to estimate the cross-subject marginal shift from a group of M domains (subjects) relies on estimating pair-wise domain divergences, i.e., we compute $d_{\mathcal{H}}[\mathcal{D}_i, \mathcal{D}_j]$ $\forall i, j = \{1, \cdots, M\}$. In the case of scenarios where EEG datasets are taken into account, estimating cross-domain marginal shifts consists in obtaining models capable of performing pair-wise discrimination of features extracted from recordings of different subjects. Similarly to the proposed strategy to estimating cross-subject conditional shift values, we introduce a Hermitian matrix H that accounts for marginal shifts between all subjects. Each entry of H corresponds to the average error rate of pair-wise subject classification. In practice, we use 5-fold cross validation to estimate the error rates. An aggregate value of marginal shift can also be obtained via the rescaled Frobenius norm of H.

3.5 Experimental setup

In this Section we provide a recap of the WAUC dataset, as well as introduce the features, normalization approaches, and the mental workload classification scheme utilized in the experiments. Moreover, we describe the implementation details in order to allow reproducibility of our experiments.

3.5.1 WAUC dataset

We consider for our empirical evaluation the EEG recordings of the Workload Assessment Under physical aCtivity (WAUC) dataset [74] previously introduced in Chapter 2. It is important to highlight that there are two different types of baseline recordings in the WAUC dataset (c.f. Figure 2.5): 1) EEG was recorded when no mental or physical effort was demanded from the participant (eyes-closed, no movement), and 2) data was acquired when only physical effort was taken into account, i.e., subjects were running on the treadmill or pedalling at the specified speed while executing no mental task. In this Chapter, we considered a total of 18 subjects from the dataset, whom half performed physical activity with the treadmill and the other half with the bike, and only take into account experimental sessions recorded under high physical activity levels.

3.5.2 Feature extraction

Our preprocessing and feature extraction pipeline consisted of downsampling the EEG recording to 250Hz, filtering it with a band-pass filter from 0.5-45 Hz, and computing features over 4-second epochs with 3 seconds of overlap between consecutive windows. Considering a 10-minute experimental session, after downsampling and epoching the data, we obtained an approximate total of 600 points per subject×session. As the literature has shown that increases in mental workload induce changes in alpha, beta, and theta bands in the frontal cortex [139, 140], we considered power spectral density (PSD) features in standard EEG frequency bands, namely: delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), and beta (2-30 Hz). The power of each electrode in each of these bands was calculated by decomposing the EEG signal using band-pass filters and then calculating the normalized squared-magnitude for the corresponding time-series.

3.5.3 Normalization

Feature normalization is a common practice used to minimize the effects of cross-subject variability for EEG-based classification tasks. Features extracted during tasks (mental or physical) are typically normalized with respect to the statistics of features from recordings corresponding to baseline periods [141, 142, 143, 144]. The main goal of this strategy is to emphasize changes in the features that correspond to factors that were modulated during the experimental task. In the case of the WAUC dataset, normalizing the features with respect to the first baseline period (baseline-1) highlights changes on the PSD due to both mental and physical stimuli. In turn, normalization with respect to the statistics of recording collected during the second baseline highlights modifications stemming only from mental workload changes, as only physical strain was modulated during this step.

While commonly believed to improve classification accuracy, it is not clear from a statistical learning perspective whether and why these different normalization strategies work. Here, we quantitatively assess the impact that normalization has on mental workload performance under the lens of conditional and marginal shifts, as well as of cross-subject classification performance. As such, we perform a subject-wise normalization of each feature according to:

$$x'_n = \frac{x_n - \beta}{\gamma^2}, \tag{3.8}$$

where β corresponds to the average feature vector and γ the standard deviation considering the data recorded for the respective subject during the baseline periods.

In addition to the aforementioned normalization strategies, we also perform experiments with features obtained after per-subject whitening of the data i.e., β is the sample average and γ the standard deviation for a given subject. This procedure is commonly referred to as z-score normalization. Lastly, we considered features without any normalization. As such, a total of four feature spaces are considered across our experiments: no normalization, whitening, and baseline-1/-2 normalization.

3.5.4 Cross-subject mental workload classification

In addition to analyzing the estimated cross-subject conditional and marginal shift for a mental workload assessment task, we also evaluate the cross-subject classification performance in this scenario. For that, we consider a leave-one-subject-out (LOSO) cross-validation scheme and train a different classifier per subject not included in the training set. Using this approach, we set our problem as a single-source single-target domain adaptation, where the source domain corresponds to the data of the all subjects pooled together, and the target domain corresponding to the subject left out as the test set. Although this is the setting considered in the experiments, we did not apply any domain adaptation scheme when learning classifiers since our objective in this Chapter is to investigate distributional shifts and their relationship with out-of-distribution generalization.

3.5.5 Implementation details

We implemented all classifiers, normalization, and cross-validation schemes using Scikit-learn [145]. For all experiments, we performed 30 independent repetitions considering slightly different partitions of the available data examples by randomly selecting 300 data points out of the 600 total available per subject/session. To enforce reproducibility, the random seed for all experiments was set to 10 and the code corresponding to the following experiments will be is available on GitHub[1].

A Random Forest with 20 estimators is used as the subject classifier to estimate $d_{\mathcal{H}}$ for computing the marginal shift. For predicting mental workload using LOSO cross-validation, we also use a Random Forest classifier, but in this case with 30 estimators.

3.6 Results and Discussion

In this Section, we aim at answering the following main questions: i) Do different feature normalization schemes yield different values of distributional shifts? ii) Can the estimation of distributional shifts indicate how difficult it is to learn BCIs that generalize well on unseen subjects? iii) For a fixed feature space, are our findings consistent across two partitions of the WAUC containing subjects that had physical activity levels modulated by either bike or treadmill?

3.6.1 Statistical shifts estimation

Figures 3.1 and 3.2 show the boxplots with 30 independent estimates of the conditional shift for subjects corresponding to treadmill and bike, respectively. Considering the results obtained with the non-normalized version of the features as reference, it is possible to observe that whitening the features significantly improved the estimated aggregate conditional shift values (Equation 3.6) for both treadmill and bike cases. As expected, this type of normalization is widely used in machine

[1] `https://github.com/belaalb/EEG-DA`

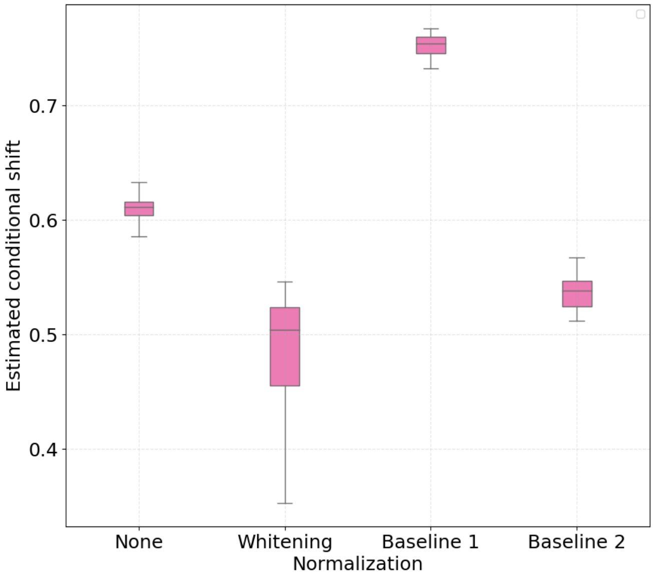


Figure 3.1 – Boxplots with 30 independent estimates of the aggregate cross-subject conditional shift across different normalization strategies for participants which performed physical activity using a *treadmill*. Lower values represent smaller estimated conditional shift.

learning and known to improve overall classification performance in different applications of EEG data [146, 147, 148].

In the case of normalizing the features with respect to the baseline periods, our findings show large differences when comparing the treadmill and bike conditions. For the bike case, normalizing the features yielded only a slight decrease in the observed conditional shift for both baseline-1 and baseline-2 periods. For the treadmill condition, on the other hand, normalizing relative to baseline-1 (no physical activity) resulted in an increase of the aggregated conditional shift, thus potentially negatively affecting the performance of the mental workload assessment model to unseen subjects. Baseline-2 normalization, in turn, reduced the estimated conditional shift to levels closer to that achieved with per-subject whitening.

In addition to investigating the aggregated conditional shift values, an in-depth analysis is also performed for the conditional shift values across all pairs of subjects in order to better understand the effects of feature normalization and the dependency on activity type. For that, Figures 3.3 and 3.4 display the disparity matrices D computed considering features without normalization and whitening for both activity types, respectively. Notice that the entries at the main diagonal (i.e.,

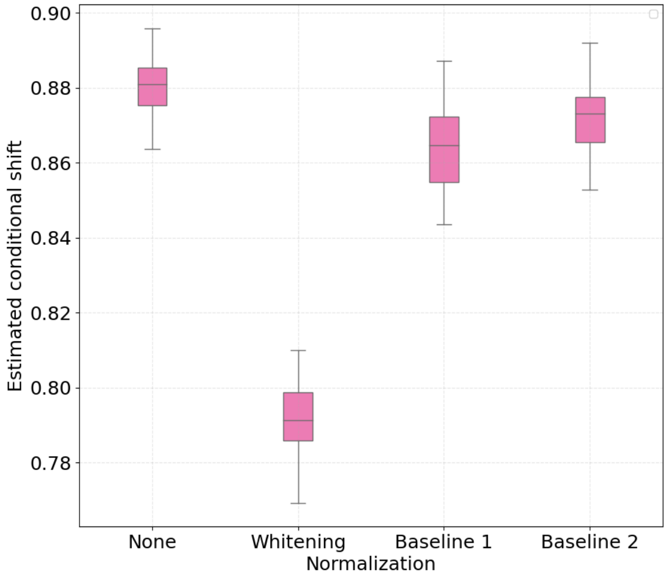
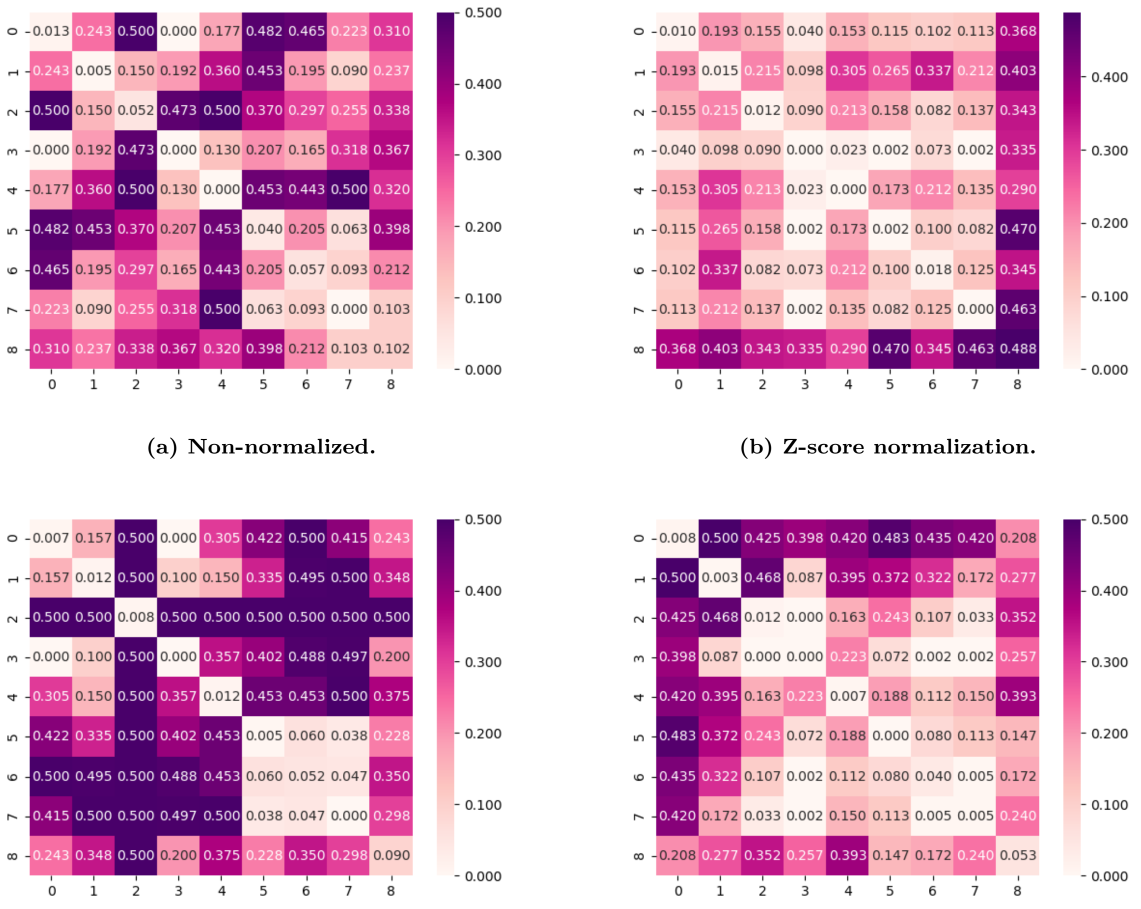


Figure 3.2 – Boxplots with 30 independent estimates of the aggregate cross-subject conditional shift across different normalization strategies for participants which performed physical activity using a *bike*. Lower values represent smaller estimated conditional shift.

within-subject disparity) were computed by having disjoint training and test sets, thus these values provide information about how good the employed labeling function approximation was. Also, these results correspond to a single estimate, thus do not show the variability of the reported quantities as it is the case in Figures 3.1 and 3.2.

It can be observed that the cross-subject conditional shift for the bike condition is much higher in comparison to the treadmill condition. This observation agrees with the findings of [74] and [149], which observed that different methods for inducing physical activity generate different EEG responses. Our results indicate that in the case of PSD features, this difference can be observed in practice by EEG responses which are more subject-specific, resulting in lower classification performance for the case of performing activity with a stationary bike, as reported in [74].

Similarly to the conditional shift analysis, we show in Figures 3.5 and 3.6 boxplots for the estimated aggregated marginal shift computed 30 times for all the considered normalization procedures, for treadmill and bike conditions, respectively. It is important to highlight that higher values of marginal shift (i.e., high $d_{\mathcal{H}}$) indicate a higher accuracy on pair-wise cross-subject classification. As such, discriminating data from two subjects in the PSD feature space consists in an easier task, and



(a) Non-normalized.

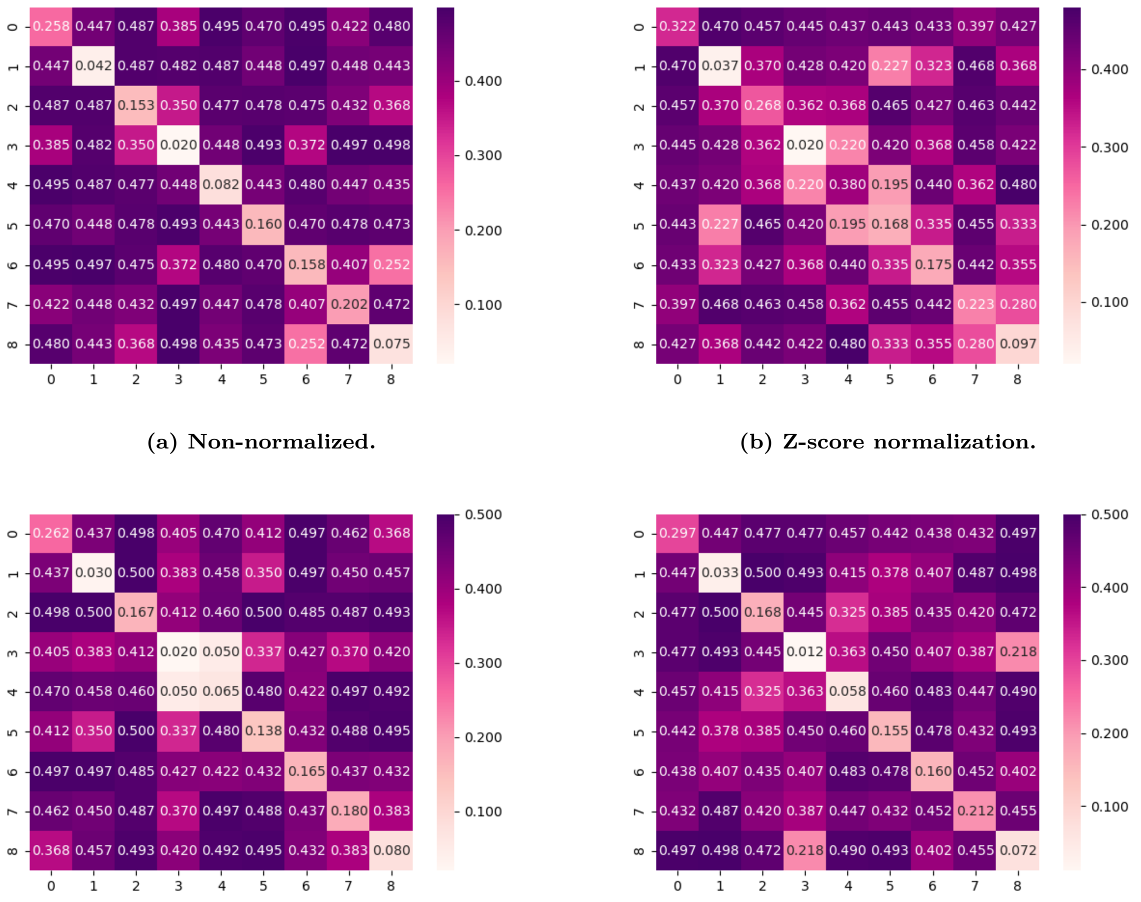
(b) Z-score normalization.

(c) Baseline 1 normalization.

(d) Baseline 2 normalization.

Figure 3.3 – Pair-wise cross-subject conditional shift computed from subjects that performed physical activity on the *treadmill*.

this contributes to higher cross-subject variability. We observe that for both treadmill and bike cases, subject-wise feature whitening decreased the estimated marginal shift, while baseline-1 and baseline-2 normalization increased it. Intuitively, we expected z-score normalization to decrease the marginal shift, as the normalized features for all subjects have equal first and second order statistics. On the other hand, according to previous results on baseline normalization for EEG features, we expected that both baseline-1 and baseline-2 methods would make it more difficult for the classifier to discriminate subjects in the PSD feature space.



(a) Non-normalized.

(b) Z-score normalization.

(c) Baseline 1 normalization.

(d) Baseline 2 normalization.

Figure 3.4 – Pair-wise cross-subject conditional shift with non-normalized features computed from subjects that performed physical activity on the *bike*.

3.6.2 Generalization gap

Lastly, target domain accuracy (i.e., test set or left-out subject) is reported for low/high mental workload classification using a LOSO cross-validation scheme. In addition to the test accuracy calculated on data from the subject left out, we also compute the classifier performance on the source domain by taking out from the training data 200 data points per subject. Based on the bound shown in Equation 2.10, our goal is to verify whether the estimated conditional and marginal shift values provide a way to assess the generalization gap between source and target domains. We use the training accuracy to compute the empirical risk, as it is equal to $1 - \hat{R}_X[h_X]$ calculated with a 0-1 loss. Likewise, the true risk $R_X[h_X]$ was estimated as the accuracy on the test set. We calculated training and test average accuracy and the corresponding standard deviation across 30

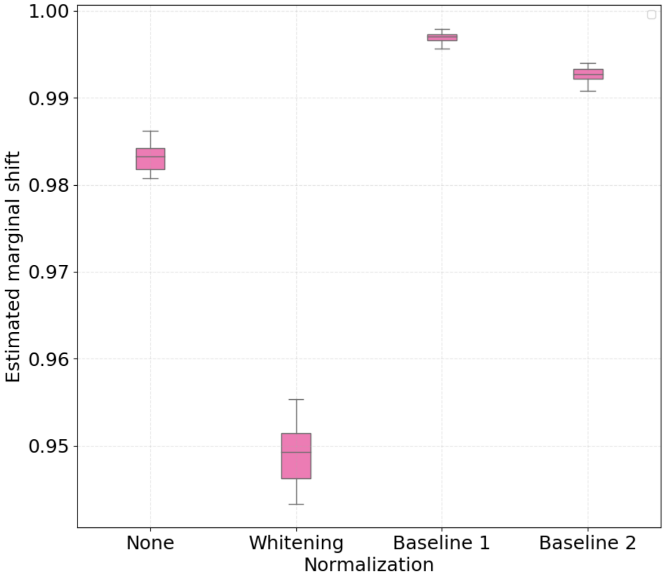


Figure 3.5 – Boxplots with 30 independent estimates of the aggregate cross-subject marginal shift across different normalization strategies for participants which performed physical activity using a *treadmill*. Lower values represent smaller estimated marginal shifts.

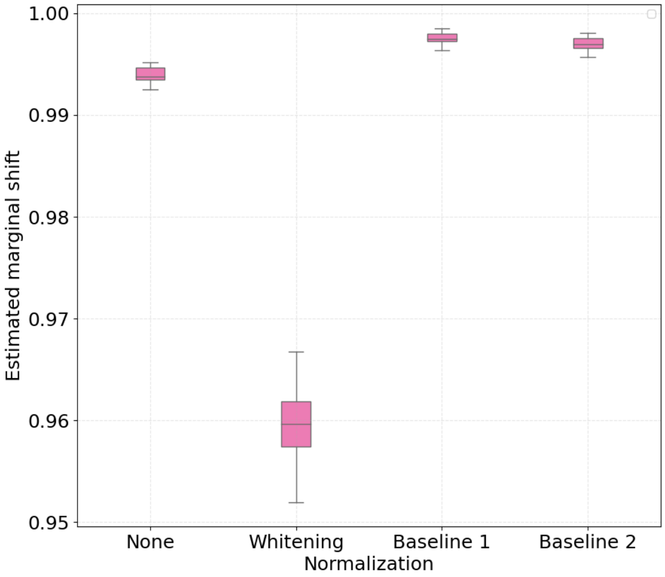


Figure 3.6 – Boxplots with 30 independent estimates of the aggregate cross-subject marginal shift across different normalization strategies for participants which performed physical activity using a *bike*. Lower values represent smaller estimated marginal shifts.

independent runs. These values are shown per subject left out during training and averaged across all subjects. We also report average and standard deviation values of the generalization gap for each subject, calculated as the absolute difference between training and test accuracy. Tables 3.1 and 3.2 present these quantities for the treadmill and the bike conditions, respectively.

Table 3.1 – Results of binary mental workload classification with leave-one-subject-out cross validation for subjects that performed physical activity on the *treadmill*. For each subject, top and middle rows represent training and test accuracy, respectively. The estimated generalization gap is shown below the dotted line. Average and standard deviation across 30 independent runs are reported.

Subject	None	Whitening	Baseline 1	Baseline 2
S0	0.974±0.004	0.936±0.007	0.985±0.003	0.982±0.004
	0.764±0.055	0.588±0.018	0.889±0.044	0.704±0.028
	0.210±0.055	0.348±0.018	0.096±0.044	0.279±0.029
S1	0.974±0.005	0.939±0.010	0.985±0.003	0.976±0.003
	0.543±0.043	0.628±0.042	0.550±0.037	0.560±0.051
	0.431±0.045	0.311±0.044	0.435±0.037	0.416±0.050
S2	0.974±0.004	0.941±0.007	0.985±0.003	0.979±0.005
	0.575±0.046	0.602±0.058	0.524±0.015	0.630±0.052
	0.399±0.046	0.340±0.060	0.461±0.016	0.349±0.051
S3	0.974±0.005	0.934±0.008	0.984±0.003	0.978±0.004
	0.700±0.079	0.968±0.060	0.643±0.082	0.603±0.055
	0.249±0.063	0.054±0.065	0.292±0.104	0.281±0.093
S4	0.977±0.003	0.939±0.008	0.985±0.003	0.983±0.004
	0.662±0.032	0.771±0.056	0.540±0.022	0.541±0.024
	0.315±0.032	0.168±0.056	0.445±0.022	0.441±0.024
S5	0.973±0.003	0.942±0.009	0.989±0.004	0.979±0.004
	0.601±0.044	0.851±0.067	0.530±0.030	0.554±0.030
	0.372±0.044	0.092±0.067	0.454±0.030	0.425±0.028
S6	0.980±0.005	0.945±0.009	0.987±0.003	0.978±0.005
	0.751±0.042	0.595±0.037	0.588±0.074	0.567±0.049
	0.229±0.043	0.350±0.039	0.399±0.074	0.411±0.049
S7	0.973±0.004	0.935±0.007	0.985±0.003	0.975±0.004
	0.613±0.088	0.862±0.044	0.821±0.074	0.565±0.047
	0.360±0.089	0.073±0.047	0.164±0.075	0.410±0.047
S8	0.984±0.003	0.959±0.006	0.989±0.003	0.982±0.003
	0.608±0.058	0.508±0.004	0.597±0.054	0.584±0.061
	0.375±0.059	0.451±0.006	0.392±0.055	0.398±0.060
All	0.976±0.005	0.941±0.011	0.985±0.004	0.979±0.005
	0.649±0.093	0.708±0.155	0.637±0.150	0.600±0.078
	0.327±0.093	0.242±0.147	0.348±0.140	0.379±0.078
Cond. shift	0.608±0.013	0.482±0.060	0.753±0.009	0.537±0.014
Marg. shift.	0.981±0.002	0.949±0.003	0.997±0.001	0.993±0.001

According to the results presented in Table 3.1, we observe that, as predicted by the bound presented in Theorem 2 (Equation 2.10), z-score normalization, i.e., the features with lower conditional and marginal shifts, presented the smallest approximated generalization gap between source and

Table 3.2 – Results of binary mental workload classification with leave-one-subject-out cross validation for subjects that performed physical activity on the *bike*. For each subject, top and middle rows represent training and test accuracy, respectively. The estimated generalization gap is shown below the dotted line. Average and standard deviation across 30 independent runs are reported.

Subject	None	Whitening	Baseline 1	Baseline 2
S0	0.921 ± 0.008	0.845 ± 0.009	0.923 ± 0.007	0.920 ± 0.006
	0.534 ± 0.026	0.536 ± 0.023	0.525 ± 0.016	0.558 ± 0.022
	0.388 ± 0.028	0.309 ± 0.026	0.398 ± 0.016	0.363 ± 0.024
S1	0.893 ± 0.009	0.826 ± 0.015	0.899 ± 0.008	0.892 ± 0.007
	0.545 ± 0.041	0.579 ± 0.027	0.550 ± 0.046	0.568 ± 0.055
	0.348 ± 0.043	0.246 ± 0.030	0.348 ± 0.047	0.324 ± 0.053
S2	0.906 ± 0.007	0.829 ± 0.011	0.909 ± 0.008	0.904 ± 0.009
	0.545 ± 0.037	0.550 ± 0.026	0.507 ± 0.007	0.519 ± 0.014
	0.361 ± 0.038	0.279 ± 0.025	0.402 ± 0.012	0.385 ± 0.018
S3	0.892 ± 0.009	0.814 ± 0.012	0.896 ± 0.009	0.895 ± 0.009
	0.541 ± 0.039	0.681 ± 0.067	0.613 ± 0.078	0.578 ± 0.063
	0.351 ± 0.038	0.133 ± 0.067	0.284 ± 0.079	0.317 ± 0.063
S4	0.903 ± 0.008	0.836 ± 0.013	0.907 ± 0.008	0.900 ± 0.007
	0.549 ± 0.027	0.541 ± 0.024	0.575 ± 0.054	0.542 ± 0.034
	0.354 ± 0.029	0.295 ± 0.028	0.331 ± 0.051	0.358 ± 0.036
S5	0.910 ± 0.009	0.837 ± 0.012	0.914 ± 0.007	0.909 ± 0.008
	0.529 ± 0.020	0.555 ± 0.037	0.531 ± 0.019	0.522 ± 0.019
	0.380 ± 0.023	0.283 ± 0.040	0.383 ± 0.021	0.387 ± 0.020
S6	0.914 ± 0.008	0.847 ± 0.013	0.918 ± 0.008	0.914 ± 0.007
	0.529 ± 0.022	0.520 ± 0.015	0.535 ± 0.026	0.536 ± 0.025
	0.385 ± 0.022	0.327 ± 0.020	0.383 ± 0.027	0.378 ± 0.027
S7	0.900 ± 0.007	0.841 ± 0.009	0.905 ± 0.007	0.898 ± 0.010
	0.549 ± 0.033	0.547 ± 0.026	0.553 ± 0.033	0.557 ± 0.043
	0.350 ± 0.032	0.294 ± 0.027	0.352 ± 0.033	0.341 ± 0.043
S8	0.896 ± 0.009	0.841 ± 0.012	0.904 ± 0.008	0.900 ± 0.008
	0.551 ± 0.039	0.599 ± 0.030	0.546 ± 0.033	0.542 ± 0.028
	0.345 ± 0.040	0.242 ± 0.034	0.358 ± 0.033	0.358 ± 0.031
All	0.904 ± 0.012	0.835 ± 0.016	0.908 ± 0.011	0.904 ± 0.012
	0.541 ± 0.033	0.567 ± 0.057	0.548 ± 0.050	0.547 ± 0.042
	0.363 ± 0.037	0.268 ± 0.065	0.360 ± 0.054	0.357 ± 0.045
Cond. shift	0.880 ± 0.008	0.792 ± 0.010	0.864 ± 0.011	0.872 ± 0.010
Marg. shift.	0.994 ± 0.001	0.959 ± 0.004	0.998 ± 0.001	0.997 ± 0.001

target domains. This finding is similarly observed in the case of the group of subjects that performed the experiment with the stationary bike, as shown by the results reported in Table 3.2. An overall comparison between treadmill and bike subjects also reveals that inter-subject generalization, as measured by the estimate of the risk on the source domain (training subjects), is considerably lower for the bike condition. This aspect could also have been predicted by the diagonal values of the disparity matrix (Figures 3.3b and 3.4b) which show that for the majority of the subjects the approximated labeling function seems to be easier to approximate for the treadmill condition.

Moreover, in the case of the treadmill group, we observe that baseline-1 normalization yielded a slightly smaller average generalization gap in comparison to baseline-2, even though it presented a considerably higher conditional shift. As both normalization strategies obtained close values of average marginal shift, we believe this indicates that the two analyzed statistical shifts might differ in their contribution to the generalization bound. Furthermore, considering the average results across all subjects, z-score normalization presented the best performance in terms of accuracy, being able to correctly classify roughly 70% of points from subjects not considered during training. It is important to highlight that as opposed to normalizing with respect to baseline recordings, which requires a calibration step to collect data prior to the actual task, z-score normalization does not need any extra information other than the features extracted from data corresponding to the task. On the other hand, despite better mitigating cross-subject variability and being more efficient in terms of data collection time, the intra-subject classification performance of models trained on z-score normalized features is worse in comparison with other strategies, indicating there might be a trade-off between improving cross-subject performance and maintaining good accuracy on the source domains.

To provide further empirical evidence that the analysis of the statistical shifts as employed in this Chapter can be used to select a feature normalization that yields better cross-domain (i.e., cross-subject) generalization, we show in Fig. 3.7 boxplots of 30 independent generalization gap estimates for each subject within the treadmill group. In addition, we provide in Fig. 3.8 a bar plot with average values of cross-subject disparity for all subjects that had physical workload modulated by the treadmill. These values were computed using the columns of the average disparity matrix resulting from the 30 repetitions executed to generate Fig. 3.1. Notice that within this analysis we are not taking into account the marginal shift. By comparing Figs. 3.7 and 3.8 we observe that for subjects 2, 3, 4, 5, 7, and 8 the normalization method with lower average conditional shift, yielded a smaller median estimated generalization gap. Importantly, we observe that subject 8 did not benefit from z-score normalization, as the conditional shift increased, along with an increase in the generalization and a decrease in the accuracy as shown in Table 3.1.

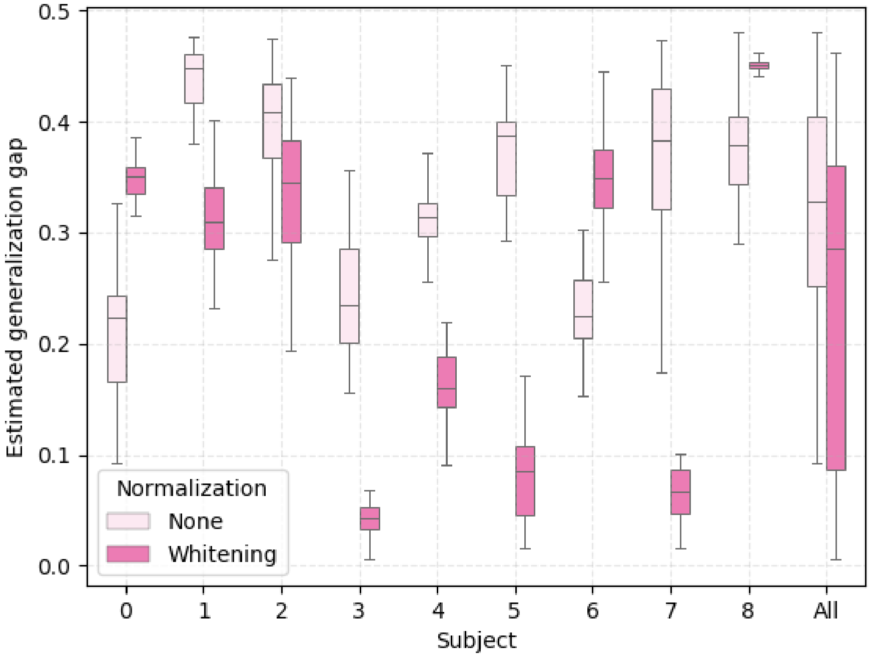


Figure 3.7 – Boxplot with 30 independent estimates of the generalization gap for the subjects that performed the experiment using a *treadmill*. The generalization gap is computed as the difference between training and test accuracy using a leave-one-subject-out cross-validation setting.

3.6.3 Main takeaways

In light of our results and discussion, we highlight the observations we found most relevant to be considered by future research. In case the goal is to improve out-of-distribution performance, normalization procedures that decrease the overall cross-subject conditional shift should be prioritized since they yield smaller generalization gaps. To devise passive BCIs with the aim of monitoring mental workload under physical activity, our analysis showed that z-score normalization provided the best strategy for normalizing EEG power spectral density features. Moreover, such normalized feature spaces should be considered in case representation learning based on domain adaptation is used to learn domain-invariant classifiers. Notice there is a caveat that should also be taken into account: the results shown in Tables 3.1 and 3.2 consistently indicate (i.e., across equipment for modulating physical activity and normalization procedures) that improving out-of-distribution performance via normalizing the features leads to a decrease on the model accuracy computed on unseen data from the training subjects.

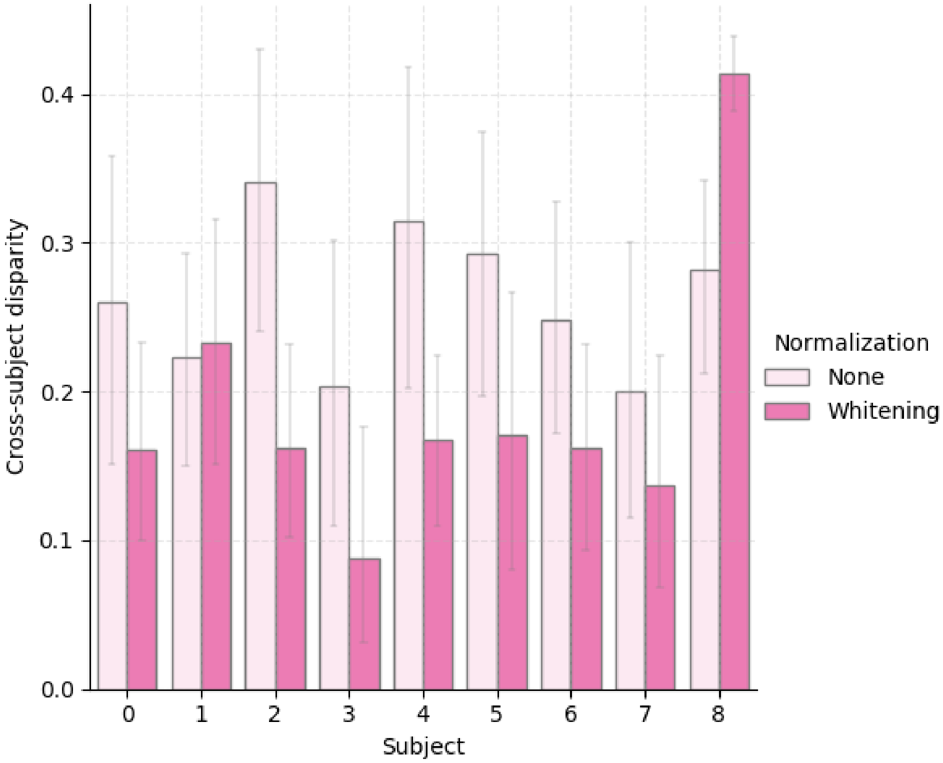


Figure 3.8 – Bar plot with the average cross-subject disparity for 30 independent estimates of the disparity matrix for the subjects that performed the experiment using a *treadmill*.

3.7 Conclusion

In this Chapter, we present the first steps towards better understanding the cross-subject variability phenomena seen with passive EEG-based BCIs from a statistical learning perspective. We looked at this problem through the lens of domain adaptation and proposed strategies to estimate distributional shifts between conditional and marginal distributions corresponding to the data generating process of features and labels from different subjects. To evaluate the proposed approach, the WAUC dataset was used and binary mental workload assessment from EEG power spectral features was performed. Our analysis showed that feature normalization, as well as data collection conditions such as the equipment used to induce physical workload, had a relevant impact in the estimated values of conditional shift. Importantly, our results indicate that whitening the features (i.e., performing z-score normalization) mitigated both conditional and marginal shifts and improved mental workload assessment on unseen subjects at training time.

Chapter 4

Generalizing to unseen domains via distribution matching

4.1 Preamble

This chapter is compiled from material extracted from the manuscript presented at the Uncertainty and Robustness and Deep Learning Workshop at the International Conference on Machine Learning [29] and the manuscript under review at the IEEE Transactions on Systems, Man, and Cybernetics: Systems [76].

4.2 Introduction

In this Chapter, we make contributions in the direction of devising more general learning machines. We take a step further from the domain adaption and consider a more general framework which is often referred to in the literature as *domain generalization* [22]. In this case, it is assumed that a set of distributions over the input space is available at training time. At test time, however, both observed distributions, as well as unseen novel domains, might appear, and a low risk should be obtained regardless of the underlying domain. More importantly, unlike domain adaptation in which the goal is to find a representation that aligns training data distributions with a specific target domain, *domain generalization strategies aim at finding a representation space that yields* *good performance on novel distributions, unknown at training time.* Recent work on domain generalization has included the use of data augmentation [150, 151] at training time, meta-learning to simulate domain shift [152], adding a self-supervised task to encourage an encoder to learn robust representations [153, 85], and learning domain-invariant representations [154], among other approaches.

Our contributions in this chapter are within the domain generalization setting and can be divided into three main categories: theoretical, algorithm, and applications. We first argue and prove that, given a set of distributions over data, if the distances measured between any pair of such distributions are small, so is the distance between mixtures obtained from the same set of distributions. This leads to the development of a bound on the risk measured against any distribution, and further shows that generalization can be expected if one considers distributions on the neighborhood of the "convex hull"[1] defined by the set of domains accessible during training. Inspired by these findings, we introduce an approach so that an encoder is enforced to map the data from the input space to a space where domain-dependent cues are filtered away while relevant information to the task of interest is preserved. While doing so, unlike standard domain adaptation approaches, no data from test distributions is considered to be observed. We evaluate the proposed algorithm in problems where different sets of assumptions are likely to hold, showing its versatility as well as its improved performance in comparison to multiple baselines.

We summarize our contributions in the following:

1. We introduce assumptions on the data generating process tailored to the domain generalization setting, which we argue are more general than standard i.i.d. requirements and more likely to hold in practice. In other words, given a data sample, it is more likely that our assumptions will hold compared to the more restrictive i.i.d. property;
2. We prove a generalization bound for the risk over unseen domains and show that generalization can be expected for domains on the neighborhood of a notion of convex hull of distributions observed at training time;
3. Aiming to minimize the terms of the introduced bound, we devise an adversarial approach so that pairwise domain divergences are estimated and minimized. In order to do so, several

[1] i.e., the set of all mixtures obtained from given distributions.

practical improvements are proposed on top of previous approaches for domain adaption including the use of random projection layers prior to domain discriminators.

4. We provide evidence through empirical evaluation showing that the proposed approach yields improvements relative to alternative methods across scenarios where different assumptions over the observed domains hold, including realistic cases where the labeling functions might shift.

The remainder of this Chapter is organized as follows: In Section 4.3 we review the related work. In Section 4.4, we define the domain generalization setting and present our main results, as well as the resulting algorithm. Section 4.5 provides the experiment descriptions and their respective results. Section 4.6 concludes the Chapter.

4.3 Related work

A number of contributions under the domain generalization setting borrowed tools from causal inference to enforce the learned representations to be invariant across the different domains presented to the model at training time [155, 156, 157]. Other contributions such as [153] and [85] proposed different strategies to leverage self-supervised tasks to improve the out-of-distribution performance of a given model. Inspired by the domain adaptation literature [97, 17, 18], previous work on domain generalization also proposed to add a regularization term based on the minimization of an estimate of a divergence between the source domains to the empirical loss computed on the training data. This is the case of the Conditional Invariant Deep Domain Generalization (CIDDG) [154], where class-specific domain classifiers are employed to induce the encoder to learn representations where the mismatch between the labels conditional distributions is minimized. Moreover, [158] proposed the Maximum Mean Discrepancy Adversarial Autoencoder (MMD-AAE), an approach that relies on an adversarial autoencoder used along with a maximum mean discrepancy penalty [159] to remove domain-specific information.

Recent work has proposed settings where domain-shifts are simulated at training time by splitting the source domains into meta-train and meta-test sets [152, 160, 73, 161]. Strategies based on learning domain-invariant representations [22], data augmentation [151, 150], and on decomposing the model's parameters into domain-agnostic and domain-specific components [120] have also been introduced. Work on other settings with more restrictive assumptions than domain generalization are also related to our contribution. For example, recent work on multi-domain learning [162], a setting where multiple domains are available at training time and test data is drawn from the same distributions seen during training [163], also leveraged an adversarial approach to perform $\mathcal{H}$-divergence minimization.

A straightforward approach to extend the empirical risk minimization (ERM) setting for domain generalization would be to learn h minimizing the empirical risk $\hat{R}[h]$ measured over all N_S source domains and *hope* generalization would be achieved to the target data, i.e.:

$$h = \arg\min \hat{R} = \frac{1}{N_S} \sum_{j=1}^{N_S} \frac{1}{M_j} \sum_{i=1}^{M_j} \ell[h(x_i), f(x_i)]. \tag{4.1}$$

In fact, as will be discussed in more detail in next sections, such a rather simplistic approach often yields strong baselines.

4.4 Learning domain agnostic representations for domain generalization

4.4.1 Formalizing domain generalization

We start by defining a set of assumptions over the data generating process considering the domain generalization case as well as the notion of risk we are concerned with. We then define $\mathfrak{D}$, referred to as *meta-distribution*, corresponding to a probability distribution over a countable set of possible domains. Under this view, a query for a data example consists of: i) sampling a domain from the meta-distribution, and ii) sampling a data point according to that particular domain. Such process is repeated m times so as to yield a training sample $(x^m \sim \mathfrak{D}^m, y^m)$. We remark the described model of data generating processes is sufficiently general so as to include the i.i.d. case (the meta-distribution yields a single domain) as well as the domain adaptation setting (if two domains are allowed), but further supports several other cases where multiple domains exist. Figure 4.1 illustrates the data generating process of the domain generalization setting by representing the meta-distribution along with possible domains. Once a finite train sample is collected, a set of

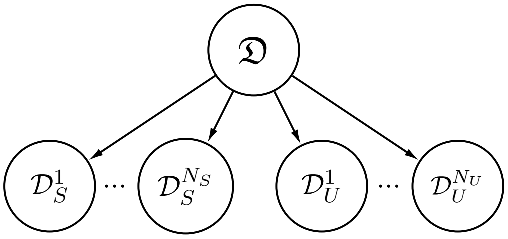


Figure 4.1 – Illustration of the meta-distribution $\mathfrak{D}$ composed by the source and unseen domains.

N_S domains is observed. Each distribution $\mathcal{D}_S^i$, $i \in [N_S]$, in such set will be referred to as source domain. At test time, however, drawing samples from $\mathfrak{D}$ might yield data distributed according to new unseen domains. We then introduce extra notation and represent the set of possible domains unobserved while train data is acquired by $\mathcal{D}_U^j$, $j \in [N_U]$. The labeling rules corresponding to each domain are denoted as f_{S_i} and f_{U_j}, for the source and unseen domains, respectively. For the sake of clarity, we hereinafter omit the index from the notation corresponding to unseen domains whenever it can be inferred from the context.

We proceed and define a risk minimization framework similar to that corresponding to the i.i.d. setting: find the predictor $h^* \in \mathcal{H}$ that minimizes the meta-risk $R_{\mathfrak{D}}[h]$ defined as follows:

$$\begin{aligned} &h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} R_{\mathfrak{D}}[h], \\ &R_{\mathfrak{D}}[h] = \mathbb{E}_{\mathcal{D} \sim \mathfrak{D}}[\mathbb{E}_{x \sim \mathcal{D}}[\ell(h(x), f_{\mathcal{D}}(x))]]. \end{aligned} \tag{4.2}$$

However, within the domain generalization setting, no information regarding possible test distributions is available at training time, which renders estimating $R_{\mathfrak{D}}[h]$ uninformative for a practical number of source domains. Moreover, we argue that no-free-lunch type of impossibility results may be used to conclude that it is impossible to generalize to any possible unknown distribution[2], so that one must assume something about the test domains in order to enable generalization. In the following results, we tackle this issue and introduce generalization guarantees for a particular set of domains lying close to the set of mixtures of *source distributions*, i.e., those observed once train data is collected.

[2] For a fixed h, one can always define a distribution yielding high risk.

4.4.2 Matching distributions in the convex hull

Let a set S of source domains such that $|S| = N_S$ be denoted by $\mathcal{D}_S^i$, $i \in \{1, \dots, N_S\}$. The convex hull Λ_S of S is defined as the set of mixture distributions given by: $\Lambda_S = \{\bar{\mathcal{D}} : \bar{\mathcal{D}}(\cdot) = \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i(\cdot), \pi \in \Delta_{N_S}\}$, where Δ_{N_S} is the N_S-th dimensional simplex. The following lemma shows that for any pair of domains such that $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^2$, the $\mathcal{H}$-divergence between $\mathcal{D}'$ and $\mathcal{D}''$ is upper-bounded by the largest $\mathcal{H}$-divergence measured between elements of S.

Lemma 1 *(Bounding the $\mathcal{H}$-divergence between domains in the convex hull of the sources). Let $d_{\mathcal{H}}[\mathcal{D}_S^i, \mathcal{D}_S^k] \leq \epsilon$, $\forall\, i, k \in [N_S]$. The following inequality holds for the $\mathcal{H}$-divergence between any pair of domains $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^2$:*

$$d_{\mathcal{H}}[\mathcal{D}', \mathcal{D}''] \leq \epsilon. \tag{4.3}$$

Proof. Consider two unseen domains, $\mathcal{D}_U'$ and $\mathcal{D}_U''$ on the convex-hull Λ_S of N_S source domains with support Ω. Consider also $\mathcal{D}_U'(\cdot) = \sum_{k=1}^{N_S} \pi_k \mathcal{D}_S^k(\cdot)$ and $\mathcal{D}_U''(\cdot) = \sum_{l=1}^{N_S} \pi_l \mathcal{D}_S^l(\cdot)$ The $\mathcal{H}$-divergence between $\mathcal{D}_U'$ and $\mathcal{D}_U''$ can be written as:

$$\begin{aligned}
d_{\mathcal{H}}[\mathcal{D}_U', \mathcal{D}_U''] =& 2 \sup_{h \in \mathcal{H}} |\mathrm{Pr}_{x \sim \mathcal{D}_U'}[h(x) = 1] - \mathrm{Pr}_{x \sim \mathcal{D}_U''}[h(x) = 1]|, \\
=& 2 \sup_{h \in \mathcal{H}} |\mathbb{E}_{x \sim \mathcal{D}_U'}[\mathbf{I}(h(x))] - \mathbb{E}_{x \sim \mathcal{D}_U''}[\mathbf{I}(h(x))]|, \\
=& 2 \sup_{h \in \mathcal{H}} \left| \int_\Omega \mathcal{D}_U'(x)\mathbf{I}(h(x))dx - \int_\Omega \mathcal{D}_U''(x)\mathbf{I}(h(x))dx \right|, \\
=& 2 \sup_{h \in \mathcal{H}} \left| \int_\Omega \sum_{k=1}^{N_S} \pi_k \mathcal{D}_S^k(x)\mathbf{I}(h(x))dx - \int_\Omega \sum_{l=1}^{N_S} \pi_l \mathcal{D}_S^l(x)\mathbf{I}(h(x))dx \right|, \\
=& 2 \sup_{h \in \mathcal{H}} \left| \int_\Omega \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \mathcal{D}_S^k(x)\mathbf{I}(h(x))dx - \int_\Omega \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \mathcal{D}_S^l(x)\mathbf{I}(h(x))dx \right|, \\
=& 2 \sup_{h \in \mathcal{H}} \left| \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \left(\int_\Omega \mathcal{D}_S^k(x)\mathbf{I}(h(x))dx - \int_\Omega \mathcal{D}_S^l(x)\mathbf{I}(h(x))dx \right) \right|.
\end{aligned} \tag{4.4}$$

Using the triangle inequality, we can write:

$$d_{\mathcal{H}}[\mathcal{D}_U', \mathcal{D}_U''] \leq 2 \sup_{h \in \mathcal{H}} \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \left| \int_\Omega \mathcal{D}_S^k(x)\mathbf{I}(h(x))dx - \int_\Omega \mathcal{D}_S^l(x)\mathbf{I}(h(x))dx \right|. \tag{4.5}$$

Finally, using the sub-additivity of the sup:

$$\begin{aligned} d_{\mathcal{H}}[\mathcal{D}'_U, \mathcal{D}''_U] \leq & \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k 2 \sup_{h \in \mathcal{H}} \left| \int_{\Omega} \mathcal{D}^k_S(x) \mathbf{I}(h(x)) dx - \int_{\Omega} \mathcal{D}^l_S(x) \mathbf{I}(h(x)) dx \right|, \\ = & \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k d_{\mathcal{H}}[\mathcal{D}^k_S, \mathcal{D}^l_S]. \end{aligned} \tag{4.6}$$

Given $d_{\mathcal{H}}[\mathcal{D}^k_S, \mathcal{D}^l_S] \leq \epsilon \; \forall \; k, l \in [N_S]$:

$$d_{\mathcal{H}}[\mathcal{D}'_U, \mathcal{D}''_U] \leq \epsilon. \qquad \square$$

We thus argue that if one minimizes the maximum pairwise $\mathcal{H}$-divergence between source domains, which can be achieved by an encoding process that filters away domain discriminative cues, the $\mathcal{H}$-divergence between any two domains in Λ_S also decreases.

4.4.3 Generalizing to unseen domains

Now we turn our attention to the set of unseen distributions $\mathcal{D}^j_U$, $j \in \{1, \ldots, N_U\}$, i.e., those in the support of the meta-distribution but not observed within the training sample. Given an unseen domain $\mathcal{D}_U$, we further introduce $\bar{\mathcal{D}}_U$, the element within Λ_S which is closest to $\mathcal{D}_U$, i.e., $\bar{\mathcal{D}}_U$ is given by $\text{argmin}_{\pi_1, \ldots, \pi_{N_S}} d_{\tilde{\mathcal{H}}} \left[\mathcal{D}_U, \sum_{i=1}^{N_S} \pi_i \mathcal{D}^i_S \right]$. We now use Lemma 1 and previously proposed generalization bounds for the domain adaptation setting [99, 98] to derive a generalization bound for the risk $R_U[h]$.

Theorem 4 *(Upper-bounding the risk on unseen domains). Given the previous setup, let S be the set of source domains and $\mathcal{Y} = [0, 1]$. The risk $R_U[h]$, $\forall h \in \mathcal{H}$, for* ***any*** *unseen domain $\mathcal{D}_U$ such that $d_{\tilde{\mathcal{H}}}[\bar{\mathcal{D}}_U, \mathcal{D}_U] = \gamma$ is bounded as:*

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R^i_S[h] + \gamma + \epsilon + \min\{\mathbb{E}_{\bar{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\}, \tag{4.7}$$

where ϵ is the highest pairwise $\tilde{\mathcal{H}}$-divergence measured between pairs within S, $\tilde{\mathcal{H}} = \{sign(|h(x) - h'(x)| - t) | h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$ and $f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$ is the labeling function for any $x \in Supp(\bar{\mathcal{D}}_U)$ resulting from combining all f_{S_i} with weights π_i, $i \in [N_S]$, determined by $\bar{\mathcal{D}}_U$.

Proof. Let the source and target domains be $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$, respectively. For the single-source, single-target domain adaptation case, it was previously shown that the risk of any $h \in \mathcal{H}$, $h : \mathcal{X} \rightarrow [0, 1]$ is bounded by [98]:

$$R_T[h] \leq R_S[h] + d_{\tilde{\mathcal{H}}}[\mathcal{D}_S, \mathcal{D}_T] + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_T - f_S|]\}, \tag{4.8}$$

where $\tilde{\mathcal{H}} = \{sign(|h(x) - h'(x)| - t)|h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$.

In order to devise a generalization bound for the risk on any unseen domain in terms of quantities related to the distributions seen at training time, we start by writing (4.8) considering $\mathcal{D}_U$ and its "projection" onto the convex-hull of the sources $\bar{\mathcal{D}}_U = \text{argmin}_{\pi_1,\dots,\pi_{N_S}} d_{\mathcal{H}} \left[\mathcal{D}_U, \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\right]$. For that, we introduce the labeling function $f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$, which is an ensemble of the respective labeling functions from each source domain weighted by the mixture coefficients that determine $\bar{\mathcal{D}}_U$. $R_U[h]$ can thus be bounded as:

$$R_U[h] \leq R_{\bar{U}}[h] + d_{\tilde{\mathcal{H}}}[\bar{\mathcal{D}}_U, \mathcal{D}_U] + \min\{\mathbb{E}_{\bar{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\}.$$

Similarly to the proof of our Lemma 1 for the case where $\mathcal{D}' = \mathcal{D}_U$ and $\mathcal{D}'' = \bar{\mathcal{D}}_U$ (and to [99]), it follows that:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \sum_{i=1}^{N_S} \pi_i d_{\tilde{\mathcal{H}}}[\mathcal{D}_S^i, \mathcal{D}_U] + \min\{\mathbb{E}_{\bar{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\}. \tag{4.9}$$

Using the triangle inequality for the $\mathcal{H}$-divergence along with Lemma 1, we can bound the $\tilde{\mathcal{H}}$-divergence between $\mathcal{D}_U$ and any source domain $\mathcal{D}_S^i$, $d_{\tilde{\mathcal{H}}}[\mathcal{D}_U, \mathcal{D}_S^i]$, according to:

$$\begin{aligned} d_{\tilde{\mathcal{H}}}[\mathcal{D}_U, \mathcal{D}_S^i] &\leq d_{\tilde{\mathcal{H}}}[\mathcal{D}_U, \bar{\mathcal{D}}_U] + d_{\tilde{\mathcal{H}}}[\bar{\mathcal{D}}_U, \mathcal{D}_S^i], \\ &\leq \gamma + \epsilon, \end{aligned}$$

where $\gamma = d_{\tilde{\mathcal{H}}}[\mathcal{D}_U, \bar{\mathcal{D}}_U]$. Using this result, we can now upper-bound $\sum_{i=1}^{N_S} \pi_i d_{\tilde{\mathcal{H}}}[\mathcal{D}_U, \mathcal{D}_S^i]$ by $\gamma + \epsilon$ and finally re-write (4.9) as:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \gamma + \epsilon + \min\{\mathbb{E}_{\bar{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\}. \qquad \square$$

Remark 1: Notice that the right-most term in Theorem 4 accounts for the mismatch between the labeling functions f_{S_π} and f_U, which reduces to 0 in most adopted scenarios within domain adaption/generalization applications, since it is often considered that the *covariate shift assumption* holds [134]. Under such setting, the labeling functions are the same across all domains in the support of $\mathfrak{D}$, i.e. $f_{S_i} = f_{U_j} = f$ for all $i \in \{1, \dots, N_S\}$ and $j \in \{1, \dots, N_U\}$. Besides the covariate shift assumption, previous work on multi-source domain adaptation [164] considered the case where the unseen domain $\mathcal{D}_U$ can be represented as a mixture of the sources with weights π_i, $i \in \{1, \dots, N_S\}$, i.e. $\mathcal{D}_U = \bar{\mathcal{D}}_U$. When such assumption holds, the term indicated by γ in Theorem 4 will vanish. We thus re-state in Corollary 1 the previous result under such simplifying assumptions.

Corollary 1 *(Generalization to unseen domains within Λ_S under the covariate shift assumption). Let all domains within the the support of the meta-distribution $\mathfrak{D}$ have labeling function f. Let S be set of source domains and its convex-hull be denoted as Λ_S. The risk $R_U[h]$ of a hypothesis h on an unseen domain $\mathcal{D}_U \in \Lambda_S$, is upper-bounded by:*

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \epsilon. \tag{4.10}$$

Proof. The right-most term of (4.7) accounts for the mismatch between the labeling functions of $\mathcal{D}_U$ and $\bar{\mathcal{D}}_U$. Since all domains within $\mathfrak{D}$ have the same labeling function, this term is equal to 0. As $\mathcal{D}_U \in \Lambda_S$, $\mathcal{D}_U = \bar{\mathcal{D}}_U$, which results in $d_{\tilde{\mathcal{H}}}[\mathcal{D}_U, \bar{\mathcal{D}}_U] = \gamma = 0$. $\square$

Remark 2: Based on the introduced results we define in the following an algorithm relying solely on source data, *unlike domain adaptation approaches*. While the total source risk can be minimized as usual, ϵ can be minimized by encoding source data to a space where source domains are hard to distinguish. We remark that we empirically found (c.f. Sections 4.5.1.2 and 4.5.3) the proposed algorithm was able to succeed even in scenarios where the considered assumptions are not likely to hold.

Remark 3: We further highlight that the introduced results also provide insights regarding *the importance of acquiring diverse datasets* in practice when targeting domain generalization (and hint as to why data augmentation is often helpful). The more diverse a dataset is regarding the number of domains present at training time, more likely it is that an unseen distribution lies within the convex hull of the source domains (i.e. $\gamma \to 0$). Therefore, not only the amount of data is important to achieve better generalization on unseen domains, but also the diversity of the training data is crucial.

Remark 4: Another practical aspect worth remarking is that, even though our domain generalization setting is more general than ERM, Theorem 4 suggests that source domain labels should also be available, since they are required to estimate ϵ, which is not the case for ERM. However, collecting domain labels is inherent to the data acquisition procedure for several tasks and commonly available as meta-data in cases such as, speech recognition, where different speakers or recording devices can be viewed as different domains.

4.4.4 Practical contributions

Motivated by the previous results, we propose to design algorithms that minimize the terms in the bound in (4.7) that can be estimated even if only source data is observed, i.e., ϵ as well as the risks over the train sample. We thus aim at learning an encoder $E : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z} \subset \mathbb{R}^d$ preserves information relevant for separating classes while removing domain-specific cues in such a way that it is harder to distinguish examples from different domains in comparison to the original space $\mathcal{X}$.

4.4.4.1 Efficiently estimating ϵ

Previous work on domain adaptation introduced strategies based on minimizing the empirical $\mathcal{H}$-divergence between sources and a given target domain [18, 99]. Instead, as per the discussion following Theorem 4, the domain generalization setting requires estimating pairwise $\mathcal{H}$-divergences across all available sources, not considering target data of any sort. Naively extending previous methods to our case would require $\mathcal{O}(N_S^2)$ estimators, which is unpractical given real-world cases where several source domains are available. We thus propose to use *one-vs-all* classifiers. In this case, there is one domain discriminator per source domain and the k-th discriminator estimates $\sum_{l \neq k} d_{\mathcal{H}}[\mathcal{D}_S^k, \mathcal{D}_S^l]$, and improves the method to a number of $\mathcal{H}$-divergence estimators linear on N_S.

We illustrate the estimation of $\mathcal{H}$-divergences using one-vs-all discriminators by considering an example in which three source domains are available. Consider samples of size M from $N_S = 3$ source domains which are available at training time. The loss $\mathcal{L}_1$ for the domain discriminator D_1 accounting for estimating $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_2]$ and $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_3]$ can be written as:

$$\begin{aligned}\mathcal{L}_1 &= \frac{1}{3M} \sum_{i=1}^{3M} \ell(D_1(x_i), y_1), \\ &= \frac{1}{M} \sum_{i=1}^{M} \ell(D_1(x_i), y_1) + \frac{1}{M} \sum_{i=M+1}^{2M} \ell(D_1(x_i), y_1) + \frac{1}{M} \sum_{i=2M+1}^{3M} \ell(D_1(x_i), y_1),\end{aligned} \tag{4.11}$$

where ℓ represents a loss function (e.g., 0-1 loss) and each term accounts for the loss provided by examples from one domain and y_1 is the corresponding domain label (i.e., in this case, $y_1 = 1$ for examples from $\mathcal{D}_1$ and $y_1 = 0$ otherwise.). Splitting the first term in two parts and replacing the domain labels y_1 by their corresponding values, we obtain:

$$\mathcal{L}_1 = \frac{1}{M} \sum_{i=1}^{M/2} \ell(D_1(x_i), 1) + \frac{1}{M} \sum_{i=M+1}^{2M} \ell(D_1(x_i), 0) + \frac{1}{M} \sum_{i=\frac{M}{2}+1}^{M} \ell(D_1(x_i), 1) + \frac{1}{M} \sum_{i=2M+1}^{3M} \ell(D_1(x_i), 0). \tag{4.12}$$

The first two terms from Eq. 4.12 account for $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_2]$ and the last two terms account for $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_3]$.

4.4.4.2 Training

The proposed approach contains three main modules, all parameterized by neural networks: an encoder E with parameters ϕ_E, a task classifier C with parameters ϕ_C, and a set of $\mathcal{H}$-divergence estimators D_k with parameters θ_k, $k \in [N_S]$. Intuitively, E attempts to minimize a classification loss $\mathcal{L}_C(\cdot; \phi_C)$ (standard cross-entropy in our case) and empirical $\mathcal{H}$-divergences, which is achieved through the maximization of domain discrimination losses, denominated $\mathcal{L}_k$. Each domain discriminator, on the other hand, aims at minimizing $\mathcal{L}_k$. The procedure for estimating ϕ_E, ϕ_C, and all

θ_k's can be thus formulated as the following multiplayer minimax game:

$$\min_{\phi_E,\phi_C} \max_{\theta_1,\ldots,\theta_{N_S}} \mathcal{L}_C(C(E(x;\phi_E);\phi_C),y_C) - \sum_{k=1}^{N_S} \mathcal{L}_k(D_k(E(x;\phi_E);\theta_k),y_k), \tag{4.13}$$

where y_C corresponds to the task label for the example x, and y_k is equal to 1 in case $x \sim \mathcal{D}_S^k$, or 0 otherwise. Training is carried out with alternate updates. A pseudocode describing the training procedure is presented in Algorithm 1. To further illustrate our proposed approach, we provide in Figure 4.2 a diagram showing the main components of the model in a case where three domains are available at training time. We refer to the proposed approach as G2DM (Generalizing to unseen Domains via Distribution Matching).

Algorithm 1 Generalizing to unseen Domains via Distribution Matching

```
Requires: classifier and encoder learning rate ($\beta_C$), domain discriminators learning rate ($\beta_D$),
scaling ($\alpha$), mini-batch size ($m$).
Initialize $\phi_E$, $\phi_C$, $\theta_1, \ldots, \theta_{N_S}$ as $\phi_E^0$, $\phi_C^0$, $\theta_1^0, \ldots, \theta_{N_S}^0$.
for $t = 1, \ldots,$ number of iterations do
  Sample one mini-batch from each source domain $\{(x_1^i, y_C^i, y_1^i, \ldots, y_{N_S}^i)\}_{i=1}^m$
  # Update domain discriminators
  for $k = 1, \ldots, N_S$ do
    $\theta_k^t \leftarrow \theta_k^{t-1} + \frac{\beta_D}{N_S \cdot m} \sum_{i=1}^{N_S \cdot m} \nabla_{\theta_k} \mathcal{L}_k(D_k(E(x^i;\phi_E^{t-1});\theta_k^{t-1}), y_k^i)$
  end for
  # Update task classifier
  $\phi_C^t \leftarrow \phi_C^{t-1} + \frac{\beta_C}{N_S \cdot m} \sum_{i=1}^{N_S \cdot m} \nabla_{\phi_C} \mathcal{L}_C(C(E(x^i;\phi_E^{t-1});\phi_C^{t-1}), y_C^i)$
  # Update encoder
  $\phi_E^t \leftarrow \phi_E^{t-1} + \frac{\beta_C}{N_S \cdot m}(\sum_{i=1}^{N_S \cdot m} \alpha \nabla_{\phi_E} \mathcal{L}_C(C(E(x^i;\phi_E^{t-1});\phi_C^{t-1}), y_C^i)$
      $-(1-\alpha)\nabla_{\theta_k} \mathcal{L}_k(D_k(E(x^i;\phi_E^{t-1});\theta_k^t), y_k^i))$
end for
```

4.4.4.3 Improving training stability

Previous work on domain adaptation/generalization [18, 154] proposed solving the problem stated in 4.13 using a gradient reversal layer [28]. We empirically observed such approach to be heavily dependent on the choice of hyperparameters in order for training to converge. We propose to augment the described adversarial approach using strategies originally utilized for stabilizing the training of generative adversarial networks with multiple discriminators [165, 77]. Namely, we include a random projection layer in the input of each domain discriminator with the goal of making examples from different distributions harder to be distinguished. In addition, we use the negative

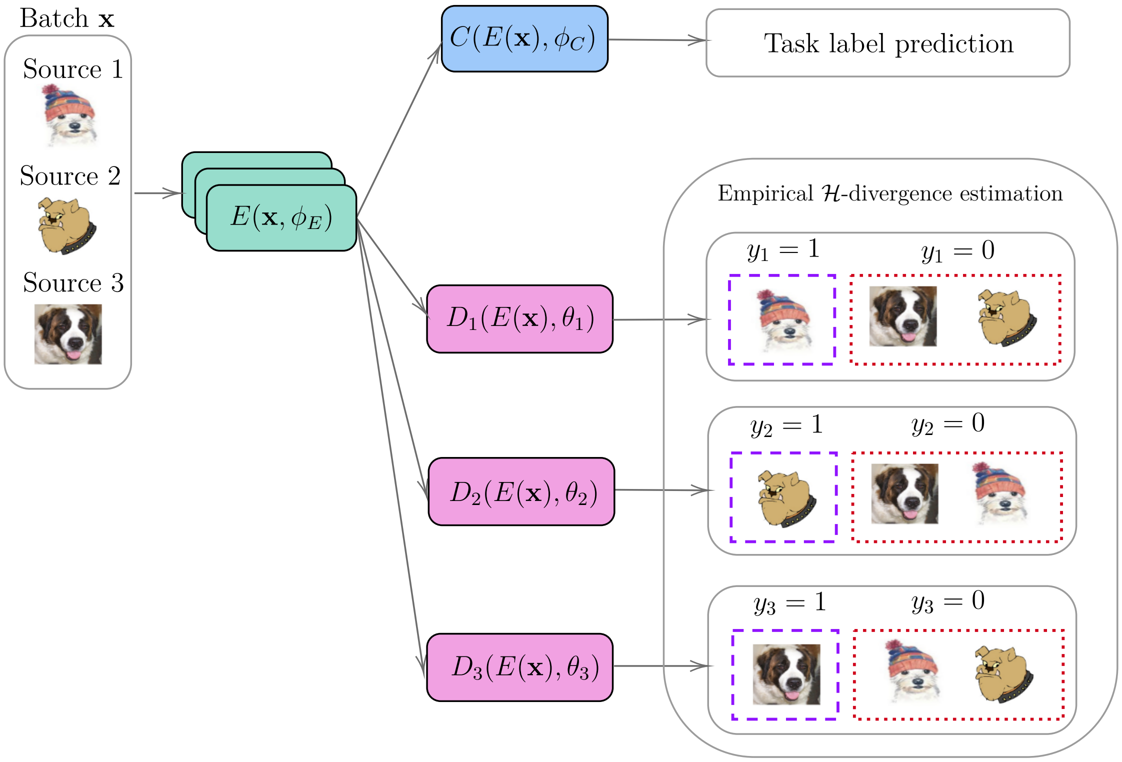

Batch $\mathbf{x}$
Source 1
Source 2
Source 3
$E(\mathbf{x}, \phi_E)$
$C(E(\mathbf{x}), \phi_C)$
Task label prediction
$D_1(E(\mathbf{x}), \theta_1)$
$D_2(E(\mathbf{x}), \theta_2)$
$D_3(E(\mathbf{x}), \theta_3)$
Empirical $\mathcal{H}$-divergence estimation
$y_1 = 1$
$y_1 = 0$
$y_2 = 1$
$y_2 = 0$
$y_3 = 1$
$y_3 = 0$

Figure 4.2 – Illustration of G2DM main components.

log hypervolume instead of the summation in the game represented in 4.13 in order to assign more preference to solutions which decrease all pairwise divergences uniformly even in cases where there is a trade-off in their minimization.

4.4.4.4 Differences to multi-source domain adaptation

We further remark the differences between G2DM and previous adversarial approaches which are often employed in domain adaptation. Essentially, G2DM compares examples *only from source domains* to learn domain-agnostic representations, i.e., there is no notion of target distribution. Other settings such as [166, 167] are more restricted in that a particular distribution is targeted and data from that distribution is required, besides the source data we use in our case. Moreover, those approaches do not aim at matching source distributions and only consider $\mathcal{H}$-divergences computed between each source domain and the given target. In the case of G2DM, on the other hand, the goal is to match source domain distributions to decrease ϵ, and thus only pairwise discrepancies between training domains are considered.

4.5 Experimental Setup and Results

We design our empirical evaluation to validate G2DM in conditions where different assumptions are satisfied. In the first scenario, we chose experimental conditions such that the covariate shift assumption holds. For that, we employ G2DM on object recognition tasks. In this case, we aim to answer the following research questions: i) Can G2DM perform better than standard ERM under i.i.d. assumptions by using information of source domains only? ii) Where does G2DM performance stand in comparison to previously proposed domain generalization strategies? iii) Is G2DM indeed enforcing distribution matching across source and unseen domains? And iv) What is the effect on the resulting performance given by different access models to test distributions during training?

We then evaluate whether G2DM is able to attain good out-of-domain performance even in the challenging scenario where the covariate shift assumption is likely to be violated. For that, we consider a real-world task that involves classifying EEG time series for affective state prediction, a burgeoning area within the human-machine systems field. In applications involving EEG data, subjects are often considered as distinct domains with different labeling functions, as also discussed and shown in Chapter 3 [72]. Implementation details can be found in Section 4.5.4.

4.5.1 Evaluation under covariate shift

As previously discussed in Chapter 2, the VLCS benchmark [121] is composed by examples from five overlapping classes from the VOC2007 [122], LabelMe [123], Caltech-101 [168], and SUN [125] datasets. PACS [120], in turn, consists of images distributed into seven classes from four different datasets: photo (P), art painting (A), cartoon (C), and sketch (S). We compare the performance of our proposed approach with a model trained with no mechanism to enforce domain generalization (referred to as ERM throughout this Section). Moreover, we consider for comparison the recently introduced invariant risk minimization (IRM) strategy [155] and include results reported in the literature achieved by Epi-FCR [161], JiGen [153] along with the ERM results they provided (referred to as ERM-JiGen), and MMD-AAE [158]. Finally, the adaptation of DANN for domain generalization reported in [161] was also considered. All such methods have as encoder the convolutional stack of AlexNet [63] and the weights are initialized from the pre-trained model on ImageNet [169].

Table 4.1 – Classification accuracy (%) on VLCS datasets for models trained with leave-one-domain-out validation.

Unseen domain (→)	V	L	C	S	Average
DANN	66.40	64.00	92.60	63.60	71.70
MMD-AAE	67.70	62.60	94.40	64.40	72.28
Epi-FCR	67.10	64.30	94.10	65.90	72.90
JiGen	70.62	60.90	96.93	64.30	73.19
ERM - JiGen	71.96	59.18	96.93	62.57	72.66
IRM	72.16	62.36	**98.35**	67.82	75.17
ERM	**73.44**	60.44	97.88	67.92	74.92
G2DM	71.14	**67.63**	95.52	**69.37**	**75.92**

Table 4.2 – Classification accuracy (%) on PACS datasets for models trained with leave-one-domain-out validation.

Unseen domain (→)	P	A	C	S	Average
DANN	88.10	63.20	67.50	57.00	69.00
Epi-FCR	86.10	64.70	72.30	65.00	72.00
JiGen	89.00	**67.63**	71.71	65.18	73.38
ERM - JiGen	89.98	66.68	69.41	60.02	71.52
IRM	89.97	64.84	71.16	63.63	72.39
ERM	**90.02**	64.86	70.18	61.40	71.61
G2DM	88.12	66.60	**73.36**	**66.19**	**73.55**

In Tables 4.1 and 4.2, we report the average best accuracy across three runs with different random seeds on the test partition of the unseen domain under a leave-one-domain-out validation scheme. Results show that G2DM outperforms ERM in terms of average performance across the unseen domains for both benchmarks, and support the claim that leveraging source domain information as done by G2DM provides an improvement on generalization to unseen distributions in comparison to simply considering the i.i.d. requirement is satisfied. G2DM further presented better average performance when compared to our implementation of IRM, as well as results from other methods previously reported in the literature. We finally highlight that G2DM showed an improvement in performance in more challenging domains [120], such as LabelMe and Sketch.

4.5.1.1 Impact of source domains diversity on unseen domain accuracy

In this experiment, we verify whether removing examples from one source domain impacts the performance on the target domain. We evaluate each target domain on models trained using all possible combinations of the remaining domains as sources. The ERM baseline is also included for

Table 4.3 – Impact of decreasing the number of source domains on VLCS. Rows represent the two source domains used.

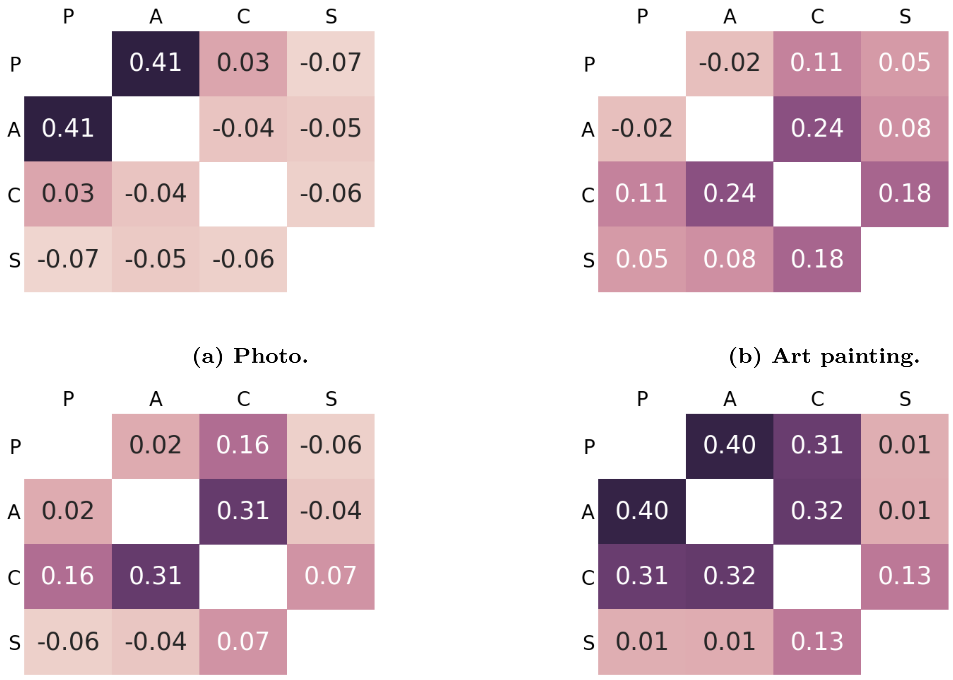
		Source					
Target	Method	VC	VL	VS	LC	LS	CS
V	ERM	-	-	-	66.14	72.16	69.89
	G2DM	-	-	-	62.39	69.89	67.23
L	ERM	58.32	-	62.11	-	-	59.85
	G2DM	65.37	-	65.87	-	-	64.37
C	ERM	-	98.82	98.58	-	84.67	-
	G2DM	-	95.75	96.70	-	81.84	-
S	ERM	69.04	66.29	-	59.80	-	-
	G2DM	69.54	68.43	-	57.06	-	-

reference. Results presented in Table 4.3 show that for all unseen domains, decreasing the number of source domains from 3 (see Table 4.1) to 2 decreased classification performance for almost all combinations of source domains. We notice that in some cases, excluding a particular source from training severely decreases the target loss. As an example, for the Caltech-101, excluding from training examples from the VOC dataset decreased the accuracy by more than 10% for the proposed approach, as well as for ERM.

4.5.1.2 Estimating H-divergences across sources and unseen domains

We now investigate whether cross-domain $\mathcal{H}$-divergences are being in fact reduced by G2DM. We use ERM as a baseline as it does not include any mechanism to enforce distribution matching. We estimate $\mathcal{H}$-divergences by computing the proxy pairwise $\mathcal{A}$-distance [17] for each pair of domains on the PACS benchmark. Classifiers are trained on top of the representations $\mathcal{Z}$ obtained with ERM and G2DM. We show in Figures 4.3a, 4.3b, 4.3c, and 4.3d the differences in estimated discrepancies between ERM and G2DM for each unseen domain. Each entry corresponds to a pair of domains indicated in the row and the column and positive values indicate that G2DM *decreased* the corresponding pairwise $\mathcal{A}$-distance in comparison to ERM. Notice that the diagonals are left blank as we do not compute the classification accuracy between the same domains.

We observe that, apart from the case where 'photo' is the test domain, G2DM was in fact able to better match most of the source distributions, thus yielding smaller ϵ which favours generalization as predicted by Theorem 4. Notably, we highlight that although our proposed approach has no access to data from the unseen domain at training time and, therefore, does not directly implement



(a) Photo.

(b) Art painting.

(c) Cartoon.

(d) Sketch.

Figure 4.3 – Differences between estimated pairwise $\mathcal{H}$-divergences under ERM and G2DM on PACS (captions denote unseen domains). Higher values indicate that G2DM better matched domains. Overall, G2DM is able to decrease pairwise discrepancies.

a strategy to decrease the divergence between the unseen domain and the convex hull of the sources (i.e. γ), the results presented in Figures 4.3a, 4.3b, 4.3c, and 4.3d show that the estimated pairwise $\mathcal{H}$-divergence between the unseen domain and sources also decreased in most of the considered cases.

In fact, the only mechanism the encoder has in order to reduce ϵ corresponds to learning how to filter domain information from the data, in the sense that once samples from two distinct distributions are encoded, one cannot distinguish from which distribution each sample came from. Observed results thus suggest that such encoder also removes domain information from the unseen distributions observed at test time, preventing the learning algorithm to yield a high γ. This effect is explained by the fact that the $\mathcal{H}$-divergence satisfies the triangle inequality $d_{\mathcal{H}}[\mathcal{D}_U, \mathcal{D}_S^i] \leq d_{\mathcal{H}}[\mathcal{D}_U, \mathcal{D}_S^j] + d_{\mathcal{H}}[\mathcal{D}_S^j, \mathcal{D}_S^i]$, which shows that an upper-bound for the discrepancy between the unseen domain and any source gets tighter once ϵ decreases.

4.5.1.3 The effect of different access methods to test data during training

Results of previous experiments correspond to an optimistic scenario where data from the unseen domain is made available for selecting the best performing model. This is not the case in practice since varying unseen distributions might appear. In Table 4.4, we compare results obtained further considering different access methods to the test data. Namely, we consider the case where no access to the unseen distribution is allowed and only source data can be used in order to define stopping criteria. In such cases, both validation accuracy and training loss computed on left-out in-domain data are employed (referred in Table 4.4 as source accuracy and source loss, respectively). Moreover, as a reference of performance, we further report the accuracy achieved assuming access to unseen domain data during training in order to select the best model (referred in Table 4.4 as unseen accuracy). For comparison, we further present the performance reported by [154] for CIDDG, since a stopping criterion using solely data from source domains was employed in that case. We observe that, when using the training loss as stopping criterion, our strategy outperforms CIDDG for almost all domains, while the baseline performance severely degrades when 'sketch' is the unseen domain.

As an alternative to AlexNet, we further evaluate the performance of the proposed approach using the convolutional stack of a ResNet-18 [170], since it has shown promising results in recent work [153]. We compare our approach with JiGen[3] adopting the same previously-discussed test data access methods for both approaches. We further report in Table 4.4 the performance obtained by JiGen as reported in [153] although it is unclear which stopping criteria were adopted for that case. We observe that replacing AlexNet by ResNet-18 yields a more stable average performance across stopping criteria. Based on the results obtained with AlexNet, we remark that results might be too optimistic/pessimistic depending on the assumed access method to unseen distributions, and as such, in order to allow fair comparison between different approaches, *the performance across different access methods should be reported.*

4.5.2 Effect of random projection size

We further investigate the effectiveness on providing a more stable training of the random projection layer in the input of each discriminator. For that, we run experiments with 7 different projection sizes, as well as directly using the output of the feature extractor model. Besides the

[3] Results are generated using JiGen authors' source code (`https://github.com/fmcarlucci/JigenDG`).

Table 4.4 – Accuracy (%) on PACS dataset with different stopping criteria and test data access methods (see text for a description of the different criterion methods listed).

Method	Criterion	P	A	C	S	Average
			AlexNet			
CIDDG	Results from [154]	78.65	62.70	69.73	64.45	68.88
G2DM	Source accuracy	85.33	57.76	69.71	49.45	65.56
	Source loss	87.37	66.70	70.26	50.98	68.82
	Unseen accuracy	88.80	66.70	73.29	65.03	73.45
			ResNet-18			
JiGen	Source accuracy	95.83	78.52	73.31	69.14	79.20
	Source loss	95.83	78.89	73.32	70.73	79.69
	Unseen accuracy	96.11	79.56	74.25	71.00	80.23
	Results from [153]	96.03	79.42	75.25	71.35	80.51
G2DM	Source accuracy	93.70	79.22	76.34	75.14	81.10
	Source loss	93.75	77.78	75.54	77.58	81.16
	Unseen accuracy	94.63	81.44	79.35	79.52	83.34

random projection size, we use the same hyperparameters values (the same used in the previous experiment) and initialization for all models. We report in Figure 4.4 the best target accuracy achieved with all random projection sizes on the PACS benchmark considering the Sketch dataset as unseen domain. Overall, we observed that the random projection layer has indeed an impact on the generalization of the learned representation and that the best result was achieved with a size equal to 1000. Moreover, we notice that, in this case, having a smaller (500) random projection layer is less hurtful for the performance than using a larger one. We also found that removing the random projection layer did not allow the training to converge with this experimental setting.

4.5.3 Evaluation beyond the covariate shift assumption

We proceed to evaluate G2DM on a unfavorable scenario where the covariate shift is unlikely to hold. The goal of the selected task is to perform affective state estimation with three classes (positive, neutral, or negative) based on EEG signals from the SEED dataset [132] collected from 15 subjects. We use the architecture described in [171] for both G2DM and ERM. For each subject left out for testing, we use 10 out of the remaining 14 domains for training and use the other 4 as validation data.

We report in Table 4.5 the classification accuracy (%) averaged across all unseen subjects and three independent training runs. Under **source data validation**, the reported performance was computed on the epoch of highest accuracy on the validation partition. The results under **semi-**

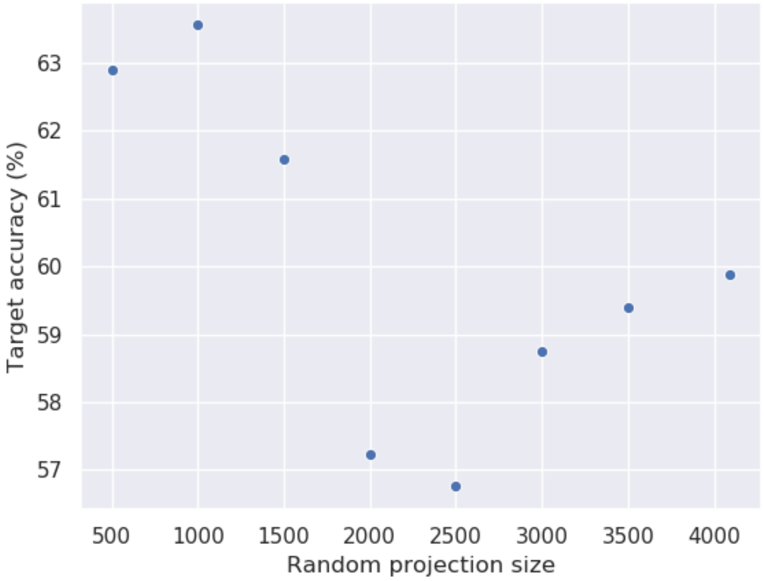


Figure 4.4 – Accuracy on the PACS dataset considering Sketch as unseen domain considering different sizes of random projection layer.

privileged were obtained on the epoch of highest accuracy on the unseen subject data. This is similar to considering the "Oracle" model selection criterion as defined in [172], with the difference that we report the highest accuracy on the full unseen domain rather than reporting the performance only on its corresponding validation partition. The comparison between G2DM and ERM shows that even in this challenging case where the mismatch between labeling functions is not negligible [72, 43, 135, 61], G2DM is able to successfully leverage the available domain information (which in this case comes with no additional effort at the data collection) and presents an improvement of more than 3.4% in accuracy in comparison to ERM in both considered scenarios for the DG setting.

4.5.3.1 Comparison with domain adaptation strategies

We further report in Table 4.5 results obtained by domain adaptation strategies. Such methods, reported in Table 4.5 under **privileged baselines**, are privileged in the sense that unlabeled data belonging to the unseen domain (unknown in our case) is used to adapt representations to yield subject-specific models. We consider for comparison the results of four different approaches as reported by [173] on the SEED dataset with the same architecture we used for G2DM. Namely, we compare the performance of G2DM with Deep Adaptation Networks (DANs) [174], Multi-source

Table 4.5 – Average accuracy (%) on the SEED dataset across 15 subjects. Semi-privileged approaches correspond to the best performing model under the domain generalization setting. Privileged baselines (domain adaptation setting) have access to unseen domain data at training time.

Setting	Method	Average accuracy. (%)
Domain Generalization	*Source data validation*	
	ERM	51.98
	G2DM	**55.77**
	Semi-privileged	
	ERM	56.82
	G2DM	**60.26**
	Privileged baselines	
Domain Adaptation	DAN [174, 173]	50.28
	DANN [18, 173]	55.87
	MDAN [99, 173]	56.65
	MDMN [173]	60.59

Domain Adversarial Network (MDAN) [99], Multiple Domain Matching Network (MDMN) [173], as well as DANN.

It is worth highlighting that, when comparing the domain adaptation strategies with domain generalization (DG) approaches, one should keep in mind that such strategies aim to obtain domain-agnostic models, as opposed to DA methods which target a specific distribution. As such, one would expect domain adaptation approaches to achieve better performance than DG when evaluated on samples drawn from such target distributions. However, we observe G2DM's performance to be on par with, or even better than, some of the considered DA strategies. We conjecture a larger number of source domains available at training time would decrease the gap between domain generalization and adaptation even further; i.e., it would be more likely that unseen domains are exactly represented in the convex hull of the sources yielding low γ (c.f. Theorem 4).

4.5.4 Implementation details

4.5.4.1 VLCS and PACS benchmarks

In order to obtain a consistent comparison with the aforementioned baseline models, we follow previous work and employ the weights of a pre-trained AlexNet [63] and ResNet-18 [170] as the initialization for the feature extractor model on the experiments. The last layer is discarded and the representation of size 4096 for AlexNet and 512 for ResNet-18 is used as input for the task classifier and the domain discriminators. The random projection layer is implemented as a linear layer with weights normalized to have unitary L2-norm. The task classifier is a one-layer fully-connected network of input size 4096 and 512 in the case of AlexNet and ResNet-18, respectively. The output size equal to the number of classes in both cases.

Following previous work on domain generalization [120, 161], we use models pre-trained on the ILSVRC dataset [175] as initialization. For fair comparison, all models we implemented were given a budget of 200 epochs. We use label smoothing [106] on the task classifier in order to prevent overfitting. Models were trained using Stochastic Gradient Descent (SGD) with Polyak's acceleration. One epoch corresponds to the length of the largest source domain training sample. The learning rate was "warmed-up" for a number of training iterations equal to nw. Hyperparameter tuning was performed through random search over a pre-defined grid so as to find the best values for the learning rate (lr), momentum, weight decay, label smoothing parameter ls, nw, random projection size[4], learning rate reduction factor, and weighting (α). Each model was run with three different initializations (random seeds 1, 10, and 100 selected *a priori*) and the average best accuracy on the test partition of the unseen domain is reported. For our ERM models, we used the same hyperparameters as in [153], while for IRM we employed the same hyperparameter values reported in the authors implementation of the colored MNIST experiments.

The grids used on the hyperparameter search for each hyperparameter are presented in the following. A budget of 200 runs was considered and for each combination of hyperparameters each model was trained for 200 and 30 epochs in the case of AlexNet and ResNet-18, respectively. The best hyperparamters values for AlexNet on PACS and VLCS benchmarks are respectively denoted by *, †. For the ResNet-18 experiments on PACS we indicate the hyperparameters by $^+$. Moreover, in the case of ResNet-18, we aggregated the discriminators losses by computing the corresponding hypervolume as in [77], with a nadir slack equal to 2.5. All experiments were run considering a minibatch size of 64 (training each iteration took into account 64 examples from each source domain) on single GPU hardware (either an NVIDIA V100 or NVIDIA GeForce GTX 1080Ti).

- Learning rate for the task classifier and feature extractor: $\{0.01^{*,+}, 0.001^{\dagger}, 0.0005\}$;
- Learning for the domain classifiers: $\{0.0005^{*}, 0.001, 0.005^{\dagger,+}\}$;
- Weight decay: $\{0.0005^{*}, 0.001, 0.005^{\dagger+}\}$;

[4] The option of not having the random projection layer was included in the search.

- Momentum: $\{0.5, 0.9^{*,\dagger,+}\}$
- Label smoothing: $\{0.0^{+}, 0.1, 0.2^{*,\dagger}\}$;
- Losses weighting (α): $\{0.35, 0.8^{*,\dagger,+}\}$;
- Random projection size: $\{1000^{*}, 3000, 3500^{\dagger}, \text{None}^{+}\}$;
- Task classifier and feature extractor learning rate warm-up iterations: $\{1, 300^{*,\dagger}, 500^{+}\}$;
- Warming-up threshold: $\{0.00001^{*}, 0.0001^{\dagger,+}, 0.001\}$;
- Learning rate schedule patience: $\{25^{+}, 60^{\dagger}, 80^{*}\}$;
- Learning rate schedule decay factor: $\{0.1^{+}, 0.3^{\dagger}, 0.5^{*}\}$.

4.5.4.2 Affective state prediction

We use SyncNet [171] as the encoder for the experiments with the SEED dataset. We follow previous work and apply a simple pre-processing that consists of clipping artifacts with amplitude 5 times higher than the mean of the channel signal and windowing data with chunks of 60 seconds. Each window was normalized to have zero mean and unit variance. For the encoder network, we adopt a one-layer parameterized convolutional filter with 2 filters (designed to extract synchrony coherence which interpretable features based on the previous neuroscience literature [171]). We train all models for 100 epochs using SGD with Polyak's acceleration. The learning rate was "warmed-up" for a number of training iterations equal to 500.

The output of the encoder with size 602 is used as input for the task classifier and the domain discriminators. The domain discriminator architecture consists of a four-layer fully-connected neural network of size 602 → random projection size → 256 → 128 → 2. The random projection layer is implemented as a linear layer with weights normalized to have unitary L2-norm. The task classifier is a two-layer fully-connected network of size 602 → 100 → number of classes.

The summary of parameters is presented in the following:

- Window size: 60 seconds
- Number of filters: 2
- Filters length: 40
- Pooling size: 40
- Input drop out rate: 0.2

- Initial learning rate task classifier: 9.963e-04
- Initial learning rate discriminator: 9.963e-05
- Random projection size: 602

4.5.4.3 Proxy A-distance estimation

We implement the domain discriminators using tree ensemble classifiers with 100 estimators. We report the average classification accuracy using 5-fold cross-validation independently run for each domain pair. Each domain is represented by a random sample of size 500.

4.6 Conclusion

In this chapter, we tackled the domain generalization problem and showed that generalization can be achieved in the neighborhood of the set of mixtures of distributions observed during training. Based on this result, we introduced G2DM, an efficient approach in yielding invariant representations across unseen distributions. Our method employs multiple one-vs-all domain discriminators, such that pairwise divergences between source distributions are estimated and minimized at training time. We provide empirical evidence supporting the claim that making use of domain information improves performance relative to standard settings relying on i.i.d. requirements. Moreover, the introduced approach outperformed recent methods which also leverage domain labels. We further showed that our proposed method resulted in strong results in a realistic setting, with performance comparable to privileged systems tailored to test distributions.

Chapter 5

Multi-objective training of Generative Adversarial Networks with multiple discriminators

5.1 Preamble

This chapter is compiled from material extracted from the manuscript presented at the International Conference on Machine Learning [77].

5.2 Introduction

Generative Adversarial Networks (GANs) [27] offer a new approach to generative modeling, using game-theoretic training schemes to implicitly learn a given probability density. Prior to the emergence of GAN architectures, realistic generative modeling remained elusive. While offering unprecedented realism, GAN training still remains fraught with stability issues. Commonly reported shortcomings involve the lack of useful gradient signal provided by the discriminator, and mode collapse, i.e., lack of diversity in the generator's samples. In this Chapter, we apply a similar architecture as that introduced in Chapter 4 for learning invariant representations for the problem of training GANs. As earlier shown in Chapter 2, both problems can be formulated as minimax optimization, and we leverage this fact to show that the same approach can be used for both generative modeling (in this Chapter) and representation learning (Chapter 4). Moreover, in this Chapter, we propose to use multi-objective optimization to solve such problems and introduce an algorithm that is a better alternative in comparison to previously proposed multi-objective approaches.

More specifically, we build upon the framework of [165] and propose reformulating their average loss minimization to further stabilize GAN training by treating the loss signal provided by each discriminator as an independent objective function. To achieve this, we simultaneously minimize the losses using multi-objective optimization techniques. Namely, we exploit well-known methods in optimization literature such as the multiple gradient descent (MGD) algorithm [109], previously introduced in Chapter 2. However, due to MGD's prohibitively high cost in the case of large neural networks, we propose to use more efficient alternatives such as maximizing the hypervolume (c.f. Chapter 2) within the region defined between a fixed, shared upper bound on the losses, which we refer to as the *nadir point* $\boldsymbol{\eta}^*$, and each of the component losses. In contrast to the approach described in [165], where the average loss is minimized when training the generator, hypervolume maximization (HV) optimizes a weighted loss, and the generator's training will assign greater importance to feedback from discriminators against which it performs poorly.

Experiments performed on MNIST show that HV presents a useful compromise between *computational cost* and *sample quality* when compared to similar approaches such as GMAN's average loss minimization (low quality and cost), and MGD (high quality and cost). Our results indicate that increasing the number of discriminators consequently increases the generator's robustness to hyperparameter settings. In addition, experiments performed on CIFAR-10 indicate the method described produces higher quality and more diverse generator samples as measured by several quantitative metrics. Moreover, image quality and sample diversity are once more shown to consistently improve as we increase the number of discriminators.

In summary, our main contributions are as follows:

1. A variation of the single-solution hypervolume maximization algorithm is introduced as an alternative to MGD for the case of training large neural networks.
2. We offer a new perspective on multiple-discriminator GAN training by framing it as a multi-objective optimization problem, and draw similarities between previous approaches under

this setting and MGD, commonly employed as a general solver for multi-objective optimization.

3. Our experiments across several datasets showed that hypervolume maximization offers better trade between computational cost and improvement in performance in comparison with both single- and multi-objective approaches for training GANs.

The remainder of this Chapter is organized as follows: In Section 5.3, we describe prior relevant literature. The HV algorithm from training GANs is detailed in Section 5.4, with experiments and results presented in Section 5.5. The main conclusions from this Chapter are outlined in Section 5.6.

5.3 Related work

Considerable research has been devoted in recent literature to overcome training instability[1] within the GAN framework. Some architectures such as BEGAN [177] have applied auto-encoders as discriminators and proposed a new loss function to help stabilize training. Methods such as TTUR [104], in turn, have attempted to define separate schedules for updating the generator and discriminator. The PacGAN algorithm [178] proposes to modify the discriminator's architecture to accept m concatenated samples as input. These samples are jointly classified as either real or generated, and the authors show that such an approach can help to enforce sample diversity. Furthermore, *spectral normalization* was applied to the discriminator's parameters in SNGAN [179] aiming to ensure Lipschitz continuity, which is empirically shown to yield high quality samples across several sets of hyperparameters.

5.3.1 Training GANs with multiple discriminators

While we would prefer to always have strong gradients from the discriminator during training, the vanilla GAN makes this difficult to ensure, as the discriminator quickly learns to distinguish real and generated samples [103], thus providing no meaningful error signal to improve the generator thereafter. The work in [180] proposed the Generative Multi-Adversarial Networks (GMAN) which

[1] *Instability* in the sense commonly used in GANs literature, i.e. when the discriminator is able to easily distinguish between real and fake samples during the training phase [165, 176, 177].

consists of training the generator against a *softmax* weighted arithmetic average of K different discriminators:

$$\mathcal{L}_G = \sum_{k=1}^{K} \alpha_k(-\mathcal{L}_{D_k}), \tag{5.1}$$

where $\alpha_k = \frac{e^{\beta(-\mathcal{L}_{D_k})}}{\sum_{j=1}^{K} e^{\beta(-\mathcal{L}_{D_j})}}$, $\beta \geq 0$, and $\mathcal{L}_{D_k}$ is the loss of discriminator k and is defined as:

$$\mathcal{L}_{D_k} = -\mathbb{E}_{x\sim p_{\text{data}}} \log(D_k(x)) - \mathbb{E}_{z\sim p_z} \log(1 - D_k(G(z))). \tag{5.2}$$

$D_k(\mathbf{x})$ and $G(\mathbf{z})$ correspond to the outputs of the k-th discriminator and the generator, respectively. The goal of using the proposed averaging scheme is to favor discriminators yielding higher losses to the generator (i.e., high $-\mathcal{L}_{D_k}$), thus providing more useful gradients during training. Experiments were performed with $\beta = 0$ (equal weights), $\beta \to \infty$ (only worst discriminator is taken into account), $\beta = 1$, and β learned by the generator. Models with $K = \{2, 5\}$ were tested and evaluated using a proposed metric and the Inception score [181]. Results showed that the simple average of discriminator's losses provided the best values for both metrics in most of the considered cases.

The work described in [165] proposed training a GAN with K discriminators using the same architecture. Each discriminator D_k sees a different randomly projected lower-dimensional version of the input image. Random projections are defined by a randomly initialized matrix W_k, which remains fixed during training. Theoretical results provided show the distribution induced by the generator G will converge to the real data distribution p_{data}, as long as there is a sufficient number of discriminators. Moreover, discriminative tasks in the projected space are harder, i.e., real and fake examples are more alike, thus avoiding early convergence of discriminators, which leads to common stability issues in GAN training such as mode-collapse [103]. Essentially, the authors trade one hard problem for K easier subproblems. The losses of each discriminator $\mathcal{L}_{D_k}$ are the same as shown in Eq. 5.2. However, the generator loss $\mathcal{L}_G$ is defined as the sum of the losses provided by each discriminator, as shown in Eq. 5.3. This choice of $\mathcal{L}_G$ does not exploit available information such as the performance of the generator with respect to each discriminator.

$$\mathcal{L}_G = -\sum_{k=1}^{K} \mathbb{E}_{z\sim p_z} \log D_k(G(z)). \tag{5.3}$$

5.4 Multi-objective training of GANs with multiple discriminators

We introduce a variation of the GAN game in which the generator solves the following multi-objective problem:

$$\min \boldsymbol{\mathcal{L}}_G(z) = [l_1(z), l_2(z), ..., l_K(z)]^T, \tag{5.4}$$

where each $l_k = -\mathbb{E}_{z \sim p_z} \log D_k(G(z))$, $k \in \{1, ..., K\}$, is the loss provided by the k-th discriminator. Training proceeds in the usual fashion [27], i.e. with alternate updates between the discriminators and the generator. Updates of each discriminator are performed to minimize the loss described in Eq. 5.2.

A natural choice for our generator's updates is the MGD algorithm, described in Chapter 2. However, computing the direction of steepest descent w^* before every parameter update step, as required in MGD, can be prohibitively expensive for large neural networks. Therefore, we propose an alternative scheme for multi-objective optimization and argue that both our proposal and previously published methods can all be viewed as performing a computationally more efficient version of the MGD update rule, without the burden of needing to solve a quadratic program, i.e. computing w^*, every iteration.

5.4.1 Hypervolume maximization for training GANs

Previous work [112] has shown that finding the set of candidate solutions for a multi-objective optimization problem that maximizes the hypervolume H (c.f Eq. 2.24) yields Pareto-optimal solutions. Since MGD converges to a set of Pareto-stationary points, i.e., a superset of the Pareto-optimal solutions, hypervolume maximization yields a subset of the solutions obtained using MGD. We exploit this property and define the generator loss as the negative log-hypervolume, as defined in Eq. 5.5:

$$\mathcal{L}_G = -\mathcal{V} = -\sum_{k=1}^{K} \log(\eta - l_k), \tag{5.5}$$

where the nadir point coordinate η is an upper bound for all l_k. In Fig. 5.1 we provide an illustrative example for the case where $K = 2$. The highlighted region corresponds to $e^{\mathcal{V}}$. Since the nadir point $\boldsymbol{\eta}^*$ is fixed, $\mathcal{V}$ will be maximized, and consequently $\mathcal{L}_G$ minimized, if and only if each l_k is minimized. Moreover, by adapting the results shown in [115], the gradient of $\mathcal{L}_G$ with respect to any generator's

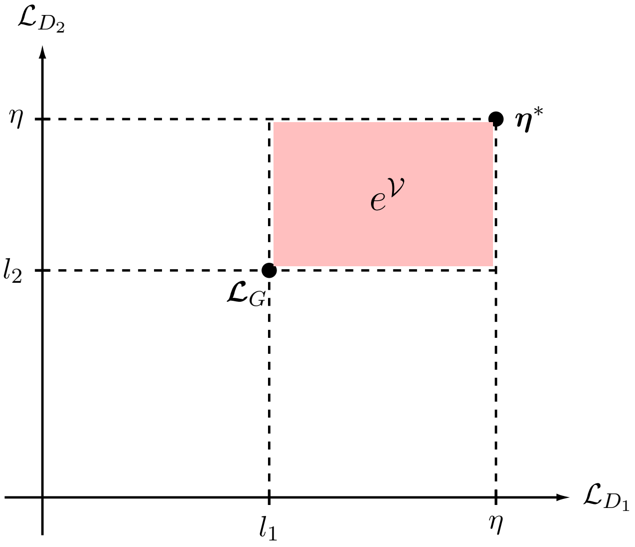


Figure 5.1 – 2D example of the objective space where the generator loss is being optimized.

parameter ϕ is given by:

$$\frac{\partial \mathcal{L}_G}{\partial \phi} = \sum_{k=1}^{K} \frac{1}{\eta - l_k} \frac{\partial l_k}{\partial \phi}. \tag{5.6}$$

In other words, the gradient can be obtained by computing a weighted sum of the gradients of the losses provided by each discriminator, whose weights are defined as the inverse distance to the nadir point components. This formulation will naturally assign more importance to higher losses in the final gradient, which is another useful property of hypervolume maximization.

5.4.1.1 Nadir point selection

It is evident from Eq. 5.6 that the selection of η directly affects the importance assignment of gradients provided by different discriminators. Particularly, as the quantity $\min_k\{\eta - l_k\}$ grows, the multi-objective GAN game approaches the one defined by the simple average of l_k. Previous literature has discussed in depth the effects of the selection of η in the case of population-based methods [182, 183]. However, those results are not readily applicable for the single-solution case. As will be shown in Section 5.5, our experiments indicate that the choice of η plays an important role in the final quality of samples. Nevertheless, we observed that this effect becomes less relevant as the number of discriminators increases.

5.4.1.2 Nadir point adaptation

Similarly to [115], we propose an adaptive scheme for η such that at iteration t: $\eta^t = \delta \max_k\{l_k^t\}$, where $\delta > 1$ is a user-defined parameter which will be referred to as *slack*. This enforces $\min_k\{\eta^t - l_k^t\}$ to be higher when $\max_k\{l_k^t\}$ is high and low otherwise, which induces a similar behavior as an average loss when training begins and automatically places more importance on the discriminators in which performance is worse as training progresses.

We further illustrate the proposed adaptation scheme in Fig. 5.2. Consider a two-objective problem with $l_1^t > 0$ and $l_2^t > 0$ corresponding to $\mathcal{L}_{D_1}$ and $\mathcal{L}_{D_2}$ at iteration t, respectively. If no adaptation is performed and η is left unchanged throughout training, as represented by the red dashed lines in Fig. 5.2, $\eta - l_1^t \approx \eta - l_2^t$ for a large enough t. This will assign similar weights to gradients provided by the different losses, which defeats the purpose of employing hypervolume maximization rather than average loss minimization. Assuming that losses decrease with time, after T updates, $\eta^T = \delta \max\{l_1^T, l_2^T\} < \eta$, since losses are now closer to 0. The employed adaptation scheme thus keeps the gradient weighting relevant even when losses become low. This effect will become more aggressive as training progresses, assigning more importance to the gradients of higher losses, as $\eta^T - \max\{l_1^T, l_2^T\} < \eta^0 - \max\{l_1^0, l_2^0\}$.

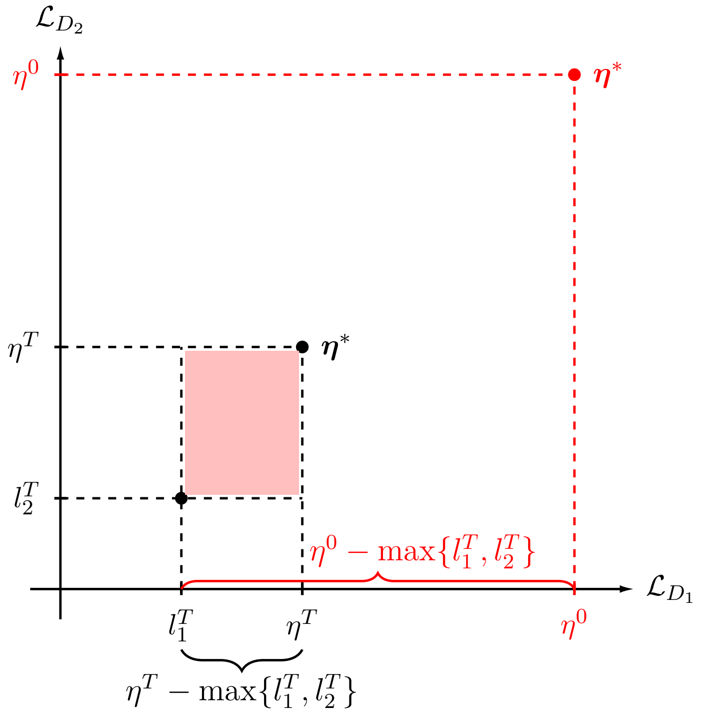


Figure 5.2 – Losses and nadir point at $t = T$, and nadir point at $t = 0$ (in red).

5.4.1.3 Comparison to average loss minimization

The upper bound proven by [165] assumes that the marginals of the real and generated distributions are identical along all random projections. However, average loss minimization does not ensure equally good approximation between the marginals in all directions. In the case of competing discriminators, i.e., when decreasing the loss on one projection increases the loss on another, the distribution of losses can be uneven. With HV on the other hand, especially when η is reduced during training, the overall loss will remain high as long as there are discriminators with high loss. This objective tends to prefer central regions, where all discriminators present roughly equally low losses.

5.4.2 Relationship between multiple discriminator GANs and MGD

All methods described previously for the solution of GANs with multiple discriminators, i.e., average loss minimization [165], GMAN's weighted average [180] and HV can be defined as MGD-like two-step algorithms consisting of: *Step 1* - consolidate all gradients into a single update direction (compute the set $\alpha_{1,...,K}$); *Step 2* - update parameters in the direction returned in Step 1. The definition of *Step 1* for the considered methods can be summarized as follows:

1. MGD: $\alpha_{1:K} = \text{argmin}_{\alpha}||w||, \quad \text{s.t.} \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0 \; \forall k \in \{1, ..., K\}$
2. Average loss minimization [165]: $\alpha_k = \frac{1}{K}$
3. GMAN [180]: $\alpha_k = \text{softmax}(l_{1:K})_k$
4. Hypervolume maximization: $\alpha_k = \frac{1}{T(\eta - l_k)}$, $T = \sum_{k=1}^{K} \frac{1}{\eta - l_k}$

5.5 Experiments

We performed four sets of experiments aiming to understand the following phenomena: (i) How alternative methods for training GANs with multiple discriminators perform in comparison to MGD; (ii) How alternative methods perform in comparison to each other in terms of sample quality and coverage; (iii) How the varying number of discriminators impacts performance given the studied methods; and (iv) Whether the multiple-discriminator setting is practical given the added cost involved in training a set of discriminators.

Firstly, we exploited the relatively low dimensionality of MNIST and used it as testbed for comparing MGD with the other approaches, i.e., average loss minimization (AVG), GMAN's weighted average loss, and HV, proposed in this work. Moreover, multiple initializations and *slack* combinations were evaluated in order to investigate how varying the number of discriminators affects robustness to those factors. Then, experiments were performed with an upscaled version of CIFAR-10 at the resolution of 64x64 pixels while increasing the number of discriminators. Upscaling was performed with the aim of running experiments utilizing the same architecture described in [165]. We evaluated HV's performance compared to baseline methods in terms of its resulting sample quality. Additional experiments were carried out with CIFAR-10 at its original resolution in order to provide a clear comparison with well known single-discriminator settings. We further analyzed HV's impact on the diversity of generated samples using the stacked MNIST dataset [184] and also compare the computational cost and performance for the single- vs. multiple-discriminator cases. Moreover, we provide generated samples to illustrate the impact of the number of random projections on the quality of the learned generative model. Lastly, we perform experiments with high resolution datasets, namely upscaled CelebA and Cats, and show that the proposed multi-objective approach for training GANs is also capable of learning high dimensional distributions.

In all experiments performed, the same architecture, set of hyperparameters and initialization were used for both AVG, GMAN and our proposed method, the only variation being the generator loss. Unless stated otherwise, Adam [185] was used to train all the models with learning rate, β_1 and β_2 set to 0.0002, 0.5 and 0.999, respectively. Mini-batch size was set to 64. The Fréchet Inception Distance (FID) [104], introduced in Chapter 2, was used for comparison.

5.5.1 MGD compared with alternative methods

We employed MGD in our experiments with MNIST and, in order to do so, a quadratic program has to be solved prior to every parameters update. For this, we used Scipy's implementation of the Serial Least Square Quadratic Program solver. Three and four fully connected layers with *LeakyReLU* activations were used for the generator and discriminator, respectively. Dropout was also employed in the discriminator and the random projection layer was implemented as a randomly initialized norm-1 fully connected layer, reducing the vectorized dimensionality of MNIST from 784 to 512. The output layer of a pretrained *LeNet* [186] was used for FID computation. Experiments

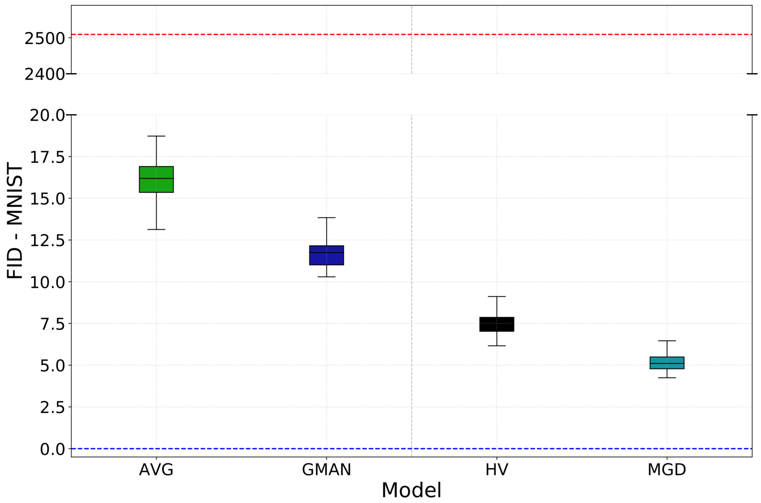


Figure 5.3 – Boxplots corresponding to 30 independent FID computations with 10000 images. MGD performs consistently better than other methods, followed by HV. Models that achieved minimum FID at training time were used. Red and blue dashed lines represent FID values for a random generator and real data, respectively.

over 100 epochs with 8 discriminators are reported in Fig. 5.3 and Fig. 5.4. In Fig. 5.3, boxplots refer to 30 independent computations of FID over 10000 images sampled from the generator which achieved the minimum FID at train time. FID results are measured at training time with 1000 images and the best values are reported in Fig. 5.4 along with the necessary time to achieve it.

MGD outperforms all tested methods. However, its cost per iteration does not allow its use in more relevant datasets outside MNIST. Hypervolume maximization, on the other hand, performs closer to MGD than the considered baselines, while introducing no relevant extra cost. In Fig. 5.5, we analyze convergence in the Pareto-stationarity sense by plotting the norm of the update direction for each method, given by $||\sum_{k=1}^{K} \alpha_k \nabla l_k||$. All methods converged to similar norms, leading to the conclusion that similar Pareto-stationary solutions will result in generators with distinct sample quality.

5.5.1.1 Sensitivity of hypervolume maximization to the initialization and the choice of the slack hyperparameter

Analysis of the performance sensitivity with the choice of the slack parameter δ and initialization was performed under the following setting: models were trained for 50 epochs on MNIST with HV

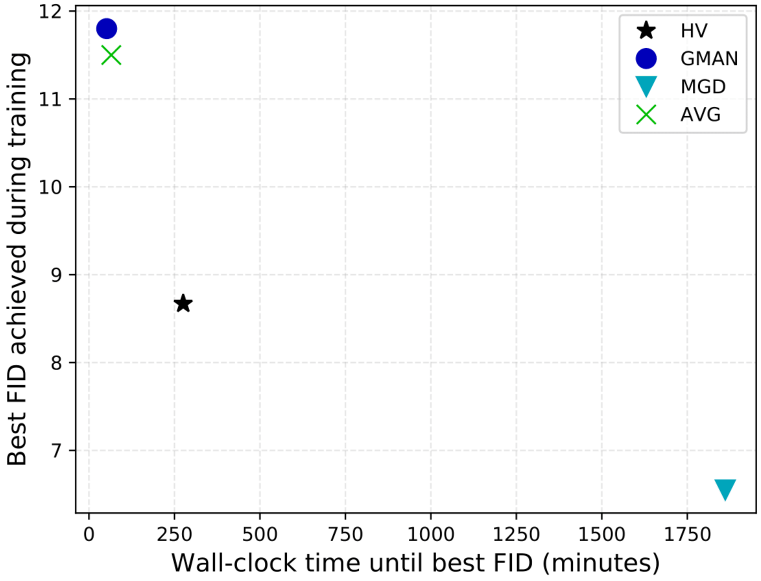


Figure 5.4 – Time vs. best FID achieved during training for each approach. FID values are computed over 1000 generated images after every epoch. MGD performs relevantly better than others in terms of FID, followed by HV. However, MGD is approximately 7 times slower than HV. HV is well-placed in the time-quality trade-off.

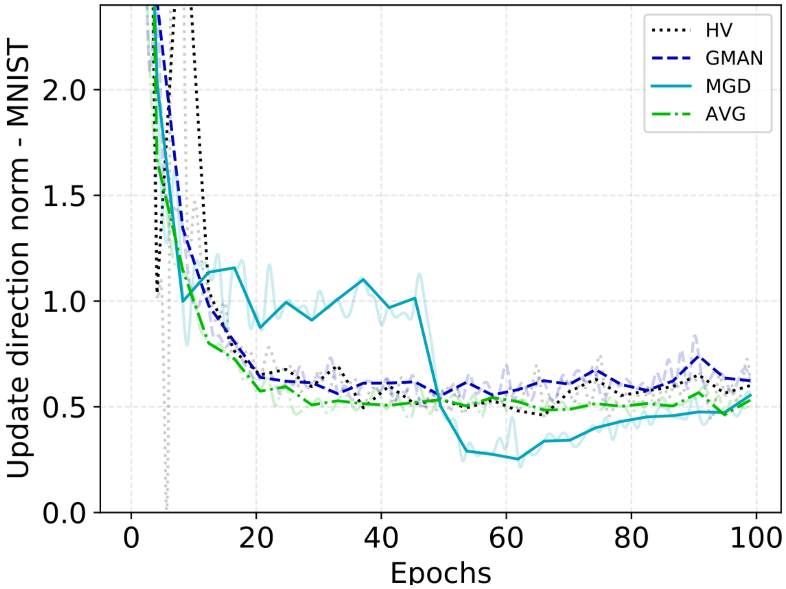


Figure 5.5 – Norm of the update direction over time for each method. While Pareto-stationarity is approximately achieved by all methods, performance varies relevantly in terms of FID.

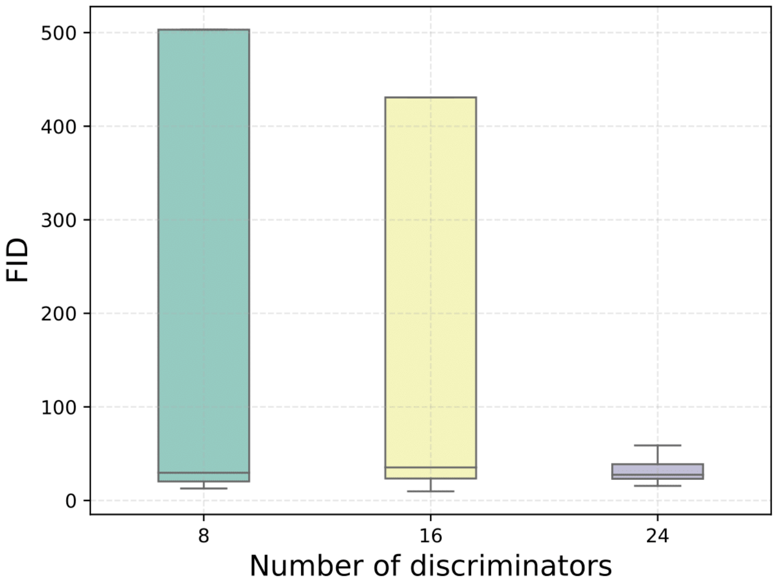


Figure 5.6 – Independent FID evaluations for models obtained with different initializations and slack parameter δ. Sensitivity reduces as the number of discriminators increases.

using 8, 16, 24 discriminators. Three independent runs (different initializations) were executed with each $\delta = \{1.05, 1.5, 1.75, 2\}$ and number of discriminators, totaling 36 final models. Figure 5.6 reports the boxplots obtained for 5 FID independent computations using 10000 images, for each of the 36 models obtained under the setting described. Results indicate that increasing the number of discriminators yields much smaller variation in the FID obtained by the final model.

5.5.2 Hypervolume Maximization as an alternative for MGD

5.5.2.1 Upscaled CIFAR-10

We evaluate the performance of HV compared to baseline methods using the upscaled CIFAR-10 dataset. FID was computed with a pretrained ResNet [170] that was trained on the 10-class classification task of CIFAR-10 up to approximately 95% test accuracy. DCGAN [187] and WGAN-GP [188] were included in the experiments for FID reference. Same architectures as in [165] were employed for all multi-discriminators settings. An increasing number of discriminators was used.

In Fig. 5.7, we report the boxplots of 15 independent evaluations of FID on 10000 images for the best model obtained with each method across 3 independent runs. Results once more indicate that

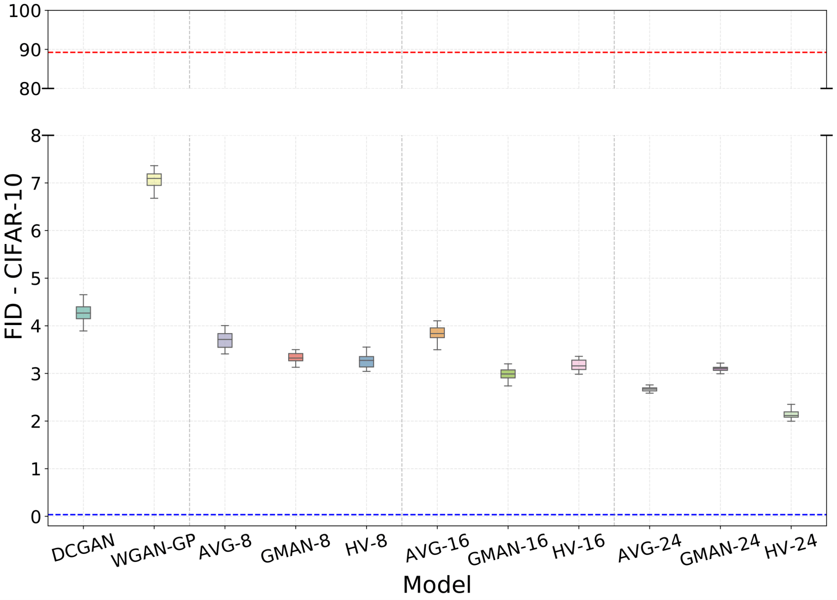


Figure 5.7 – Boxplots of 15 independent FID computations with 10000 images. Dashed lines represent the FID for real data (blue) and a random generator (red). FID was computed with a pretrained ResNet.

HV outperforms other methods in terms of quality of the generated samples. Moreover, performance clearly improves as the number of discriminators grows. Figure 5.8 shows the FID at train time, i.e., measured with 1000 generated images after each epoch, for the best models across runs. Models trained against more discriminators converge to smaller values.

We report the norm of the update direction $|| \sum_{k=1}^{K} \alpha_k \nabla l_k ||$ for each method in Figure 5.9. Interestingly, different methods present similar behavior in terms of convergence in the Pareto-stationary sense, i.e., the norm upon convergence is lower for models trained against more discriminators, regardless of the employed method.

5.5.2.2 CIFAR-10

We run experiments with CIFAR-10 in its original resolution, aiming to put our proposed approach in context with previous methods. A similar analysis can be found in the Table 2 of [179], for the model referred to as *Standard CNN*. We employ the same architecture except for the spectral normalization which was removed from the discriminators, while a random projection input layer was added. FID and IS are evaluated on 5000 generated images as in [179] as well as 10000 images,

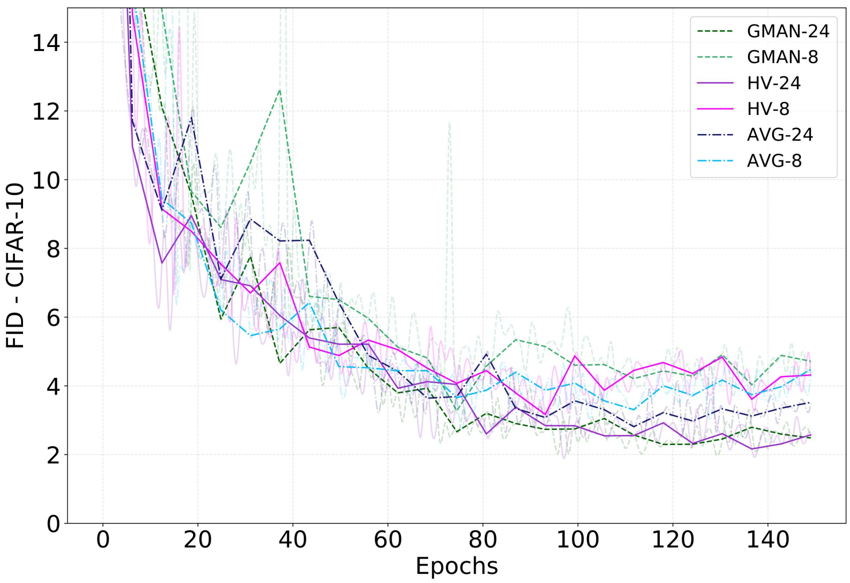


Figure 5.8 – FID estimated over 1000 generated images at train time. Models trained against more discriminators achieve lower FID. FID was computed with a pretrained ResNet.

as reported in Table 5.1. These results include our proposed approach and our implementation of [179], alongside the FID measured using a ResNet classifier trained in advance on the CIFAR-10 dataset. For completeness, we also report the Inception score (IS) [181] and compute both metrics considering samples of size 5000 and 10000.

As can be seen, the addition of the multiple discriminators setting along with HV yields a relevant shift in performance for the DCGAN-like generator, improving the evaluated metrics while the generator architecture was kept unchanged. Both IS and FID improved relative to WGAN-GP, while outperforming our own implementation of SNGAN. It is worth noting that for this experiment we selected the best performing set of hyperparameters for SNGAN, following the reported setting in prior literature [179].

	FID-ResNet	FID (5k)	IS (5k)	FID (10k)	IS (10k)
SNGAN [179]	-	25.5	7.58 ± 0.12	-	-
WGAN-GP [179]	-	40.2	6.68 ± 0.06	-	-
DCGAN [179]	-	-	6.64 ± 0.14	-	-
SNGAN (our implementation)	1.55	27.93	7.11 ± 0.30	25.29	7.26 ± 0.12
DCGAN + 24 Ds and HV	1.21	27.74	7.32 ± 0.26	24.90	7.45 ± 0.17

Table 5.1 – An evaluation of the effect of adding discriminators on a DCGAN-like model trained on CIFAR-10. Results reach the same level as the best-reported scores for the given architecture in the multiple-discriminator setting.

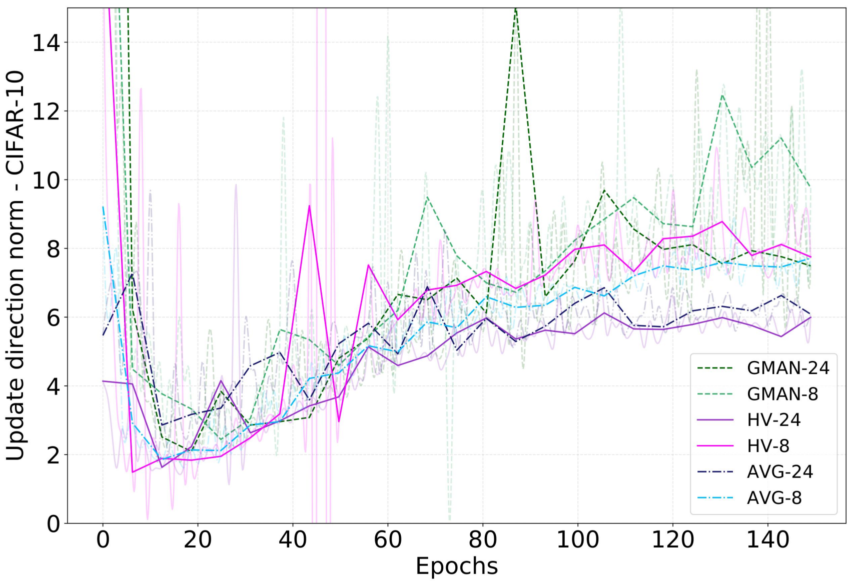


Figure 5.9 – Norm of the update direction over time for each method. Higher number of discriminators yield lower norm upon convergence.

5.5.3 Computational cost

In Table 5.2 we present a comparison of minimum FID (measured with a pretrained ResNet) obtained during training, along with computation cost in terms of time and space for different GANs, with both 1 and 24 discriminators. By design, the computational cost of training GANs under a multiple-discriminator setting is higher in terms of both FLOPS and memory, if compared with the single-discriminator GAN setting. However, the additional cost results in a corresponding improvement in performance. This effect was consistently observed using three different well-known approaches, namely DCGAN [187], Least-square GAN (LSGAN) [189], and HingeGAN [179]. The architectures of all single discriminator models follow that of DCGAN, described in [187]. For the 24-discriminator models, we used the setting described in Section 5.5.2.1. All models were trained with a minibatch of size 64 over 150 epochs.

We further emphasize that even though training with multiple discriminators may be more computationally expensive when compared to conventional approaches, such a framework supports fully parallel training of the discriminators, a feature which is not trivially possible in other GAN settings. For example for WGANs, the discriminator is serially updated multiple times for each generator update. In Figure 5.10, we provide a comparison in terms of wall-clock time per iteration

	# Disc.	FID-ResNet	FLOPS	Memory
DCGAN	1	4.22	8e10	1292
	24	1.89	5e11	5671
LSGAN	1	4.55	8e10	1303
	24	1.91	5e11	5682
HingeGAN	1	6.17	8e10	1303
	24	2.25	5e11	5682

Table 5.2 – Comparison between different GANs with 1 and 24 discriminators in terms of minimum FID-ResNet obtained during training, and FLOPs (MAC) and memory consumption (MB) for a complete training step.

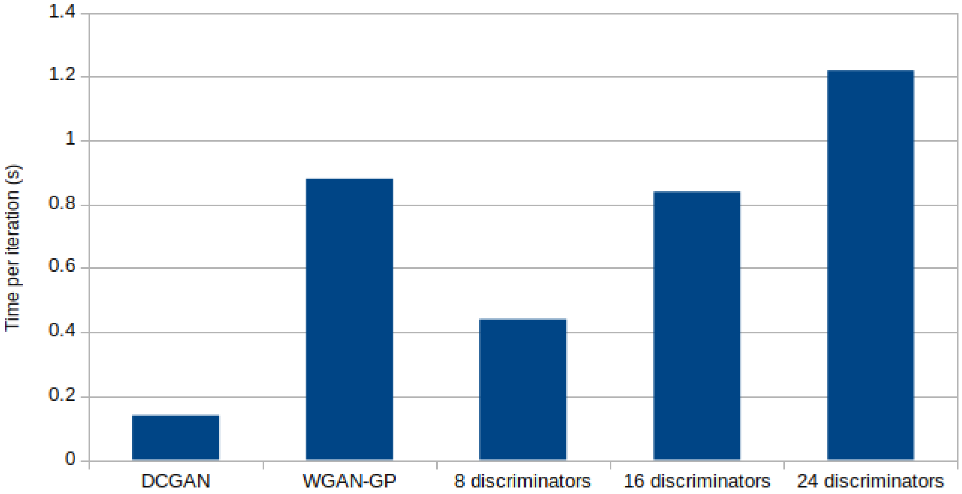


Figure 5.10 – Time in seconds per iteration of each method for serial updates of discriminators. The different multiple discriminators approaches considered do not present relevant difference in time per iteration.

on all methods evaluated. Serial implementations of discriminator updates with 8 and 16 discriminators were observed to run faster than WGAN-GP. Moreover, all experiments performed within this work were executed in single-GPU hardware, which indicates the multiple discriminator setting is a practical approach.

5.5.4 Effect of the number of discriminators on sample diversity

We repeat the experiments in [184] aiming to analyze how the number of discriminators affects the sample diversity of the corresponding generator when trained using the HV algorithm. The stacked MNIST dataset is employed and results reported in [178] are used for comparison. HV results for 8, 16, and 24 discriminators were obtained with 10k and 26k generator images, averaged

Model	Modes (Max 1000)	KL
DCGAN [187]	99.0	3.400
ALI [190]	16.0	5.400
Unrolled GAN [191]	48.7	4.320
VEEGAN [184]	150.0	2.950
PacDCGAN2 [178]	1000.0 ± 0.0	0.060 ± 0.003
HV - 8 disc. (10k)	679.2 ± 5.9	1.139 ± 0.011
HV - 16 disc. (10k)	998.0 ± 1.8	0.120 ± 0.004
HV - 24 disc. (10k)	998.3 ± 1.1	0.116 ± 0.003
HV - 8 disc. (26k)	776.8 ± 6.4	1.115 ± 0.007
HV - 16 disc. (26k)	1000.0 ± 0.0	0.088 ± 0.002
HV - 24 disc. (26k)	1000.0 ± 0.0	0.084 ± 0.002

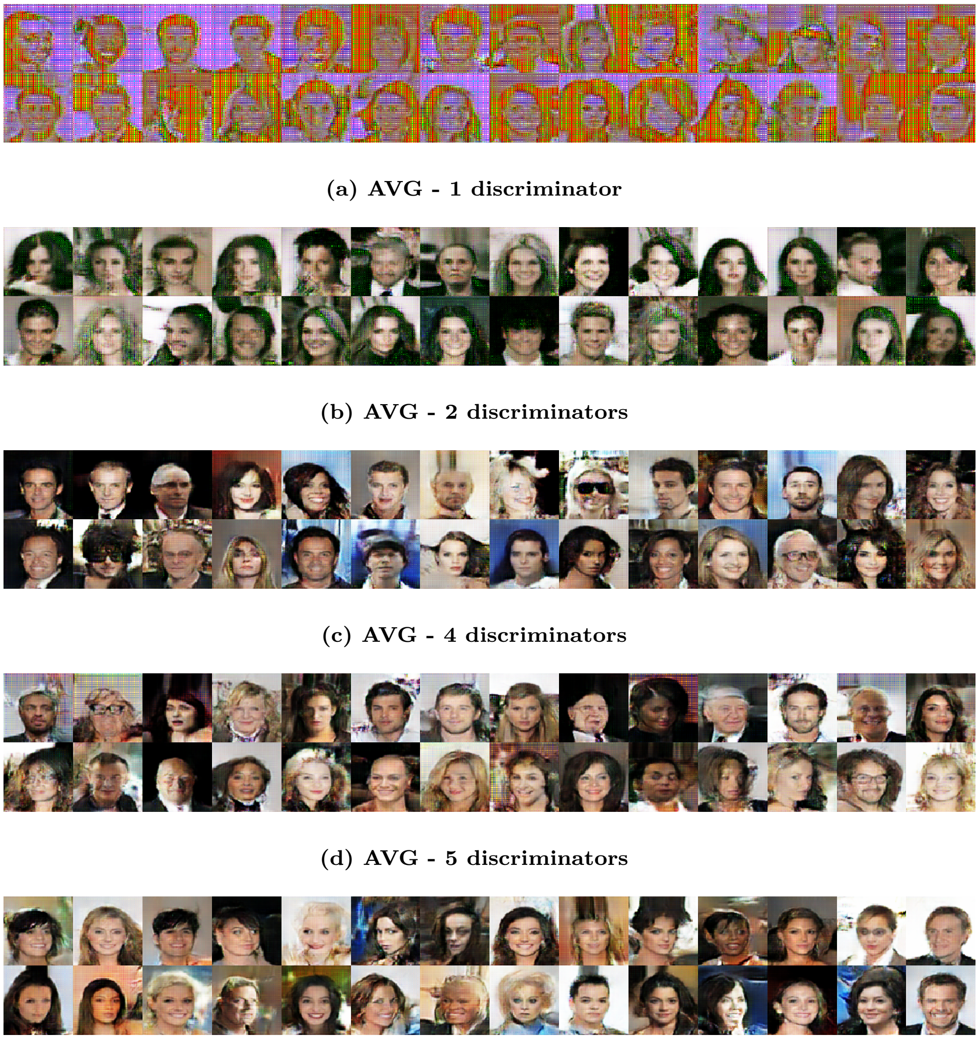
Table 5.3 – Number of covered modes and reverse KL divergence for stacked MNIST. We evaluate HV under a reduced test sample size (10k) with the goal of highlighting the effect provided by the increased number of discriminators on sample diversity.

over 10 runs. The number of covered modes along with the KL divergence between the generated mode distribution and test data are reported in Table 5.3.

As in previous experiments, results consistently improved as we increased the number of discriminators. All evaluated models using HV outperformed DCGAN, ALI, Unrolled GAN and VEEGAN. Moreover, HV with 16 and 24 discriminators achieved state-of-the-art coverage values. Thus, increasing each model's capacity by using more discriminators directly resulted in an improvement in the corresponding generator coverage.

5.5.5 Increasing the number of random projections

In this experiment we illustrate and confirm the results introduced in [165], showing the effect of using an increasing number of random projections to train a GAN. We trained models using average loss minimization with 1 to 6 discriminators on the CelebA dataset for 15 epochs. Samples from the generator obtained in the last epoch are shown in Fig. 5.11. Generated samples are closer to real data as the number of random projections (and discriminators, consequently) increases.



(a) AVG - 1 discriminator

(b) AVG - 2 discriminators

(c) AVG - 4 discriminators

(d) AVG - 5 discriminators

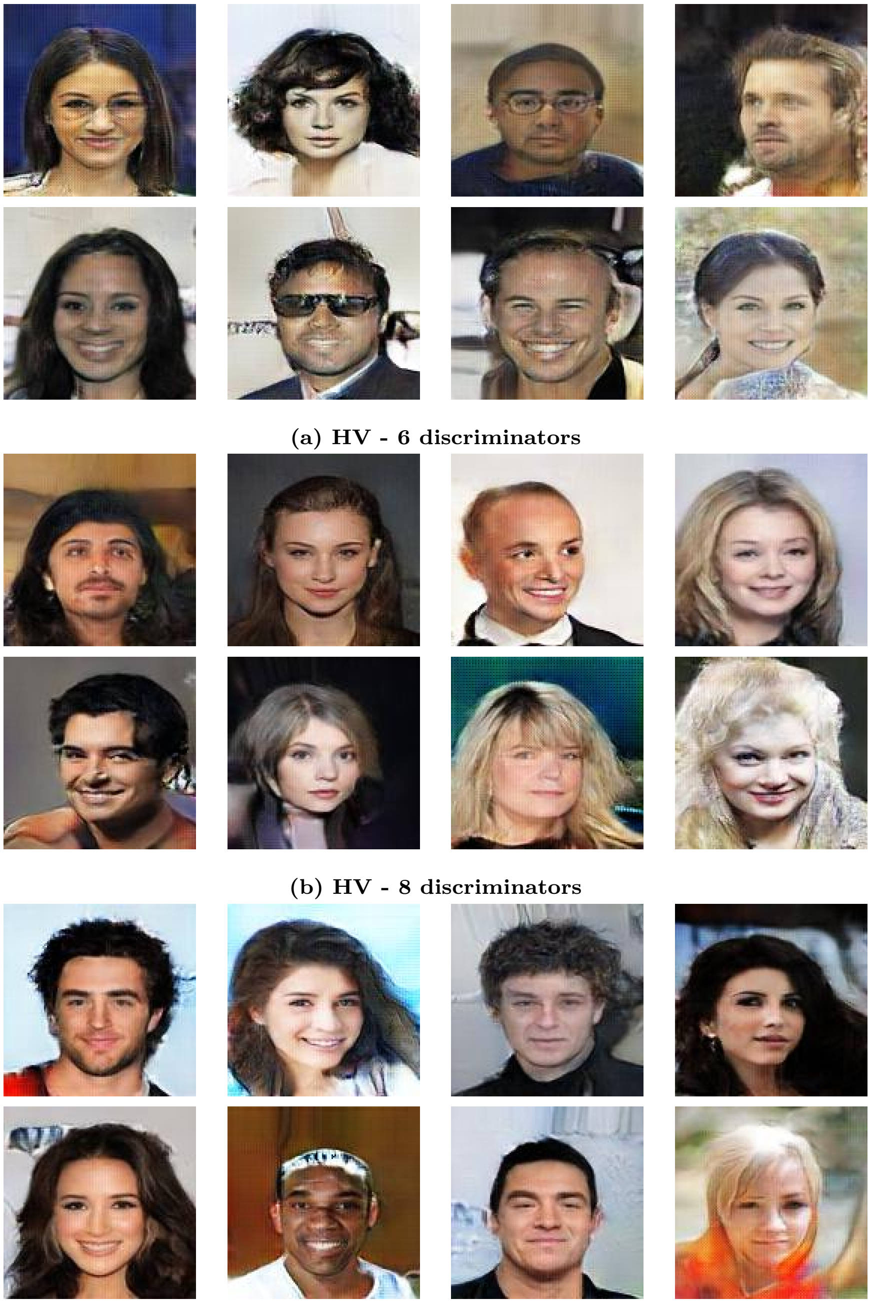
(e) AVG - 6 discriminators

Figure 5.11 – Samples drawn from models with an increasing number of random projections (i.e., discriminators) trained with AVG during 15 epochs.

5.5.6 Generating higher-resolution images

5.5.6.1 128x128 images: Upscaled CelebA

In this experiment, we verify whether the proposed multiple discriminators setting is capable of generating higher resolution images. For that, we employed the CelebA at a size of 128x128. We used a similar architecture for both generator and discriminators networks as described in the previous experiments. A convolutional layer with 2048 feature maps was added to both generator and discriminators architectures due to the increase in the image size. Adam optimizer with the



(a) **HV - 6 discriminators**

(b) **HV - 8 discriminators**

(c) **HV - 10 discriminators**

Figure 5.12 – 128x128 CelebA samples for HV trained during 24 epochs with 6, 8, and 10 discriminators.

same set of hyperparameters as for CIFAR-10 and CelebA 64x64 was employed. We trained models with 6, 8, and 10 discriminators during 24 epochs. Samples from each generator are shown in Figure 5.12.

5.5.6.2 256x256 images: Cats

We show the proposed multiple discriminators setting scales to higher resolution even in the small dataset regime, by reproducing the experiments presented in [192]. We used the same architecture for the generator. For the discriminator, we removed batch normalization from all layers and used stride equal to 1 at the last convolutional layer, after adding the initial projection step. The Cats dataset[2] was employed, we followed the same preprocessing steps, which, in our case, yielded 1740 training samples with resolution of 256x256. Our model is trained using 24 discriminators and Adam optimizer with the same hyperparameters as for CIFAR-10 and CelebA previously described experiments. In Figure 5.13 we show generator's samples after 288 training epochs. One epoch corresponds to updating over 27 minibatches of size 64.

[2] https://www.kaggle.com/crawford/cat-dataset

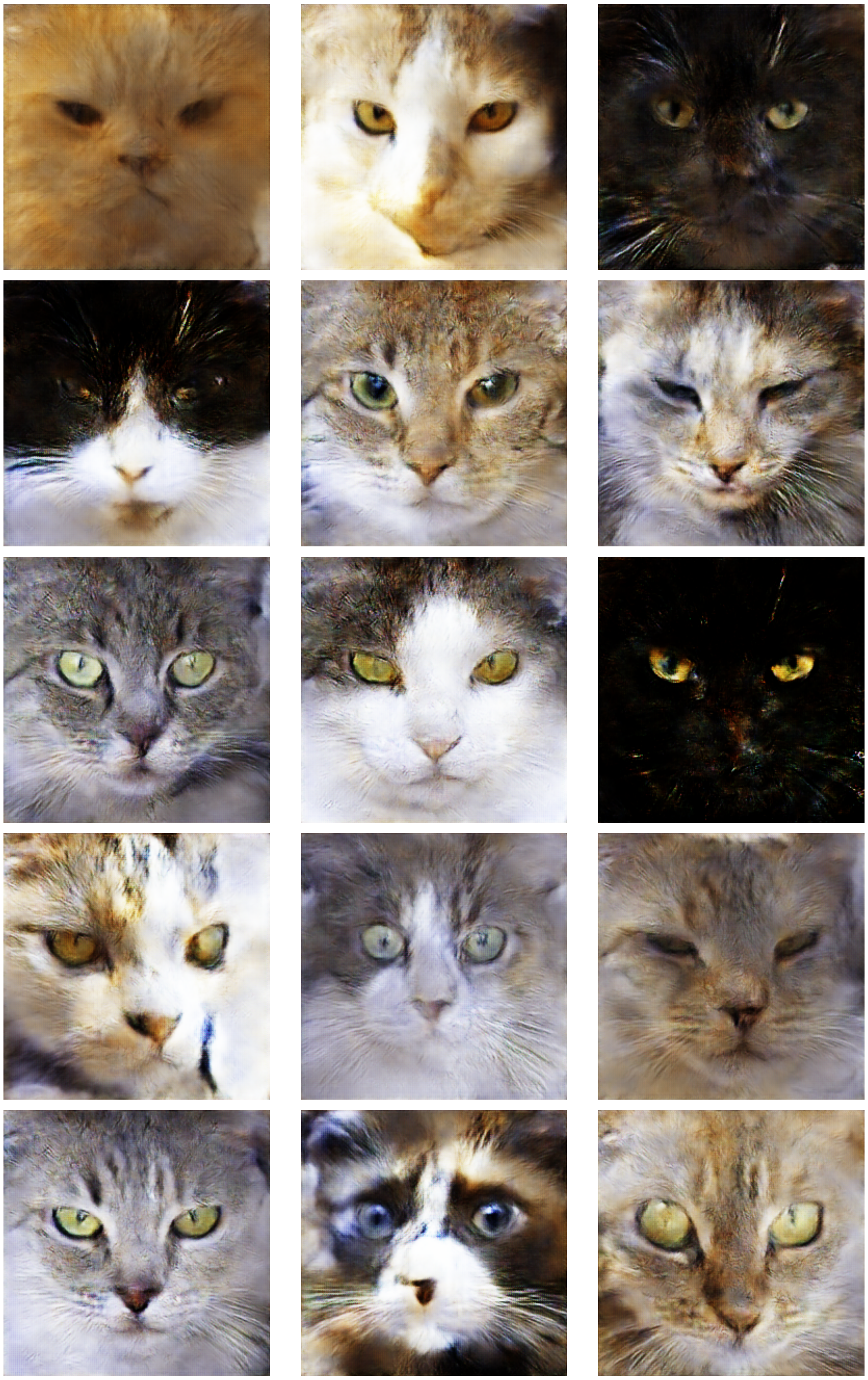


Figure 5.13 – Cats generated using 24 discriminators after 288 training epochs.

5.6 Conclusion

In this Chapter, we showed that employing multiple discriminators on GAN training is a practical approach for directly trading extra capacity - and thereby extra computational cost - for higher quality and diversity of generated samples. Such an approach is complementary to other advances in GANs training and can be easily used in tandem with other methods. We thus introduce a multi-objective optimization framework for studying multiple discriminator GANs, and showed strong similarities between previous work using such setting and the MGD algorithm. The proposed approach, namely a single-solution variation of the hypervolume maximization, was observed to consistently yield higher quality samples in terms of FID when compared to average loss and GMAN's aggregation rule. We further observed a higher number of discriminators to increase sample diversity and generator robustness.

Chapter 6

Using random projections to mitigate marginal shifts between domains

6.1 Preamble

This chapter is compiled from material extracted from the manuscript accepted to appear at the *Workshop on Distribution Shifts: Connecting Methods and Applications at the Conference on Neural Information Processing Systems*, 2021 [78].

6.2 Introduction

Different forms of distribution shifts often affect model prediction performance in machine learning applications. In recent years, new techniques have emerged to allow learning under naturally-occurring data variations, in settings such as domain adaptation and domain generalization [193, 29, 194, 85, 195]. Simultaneously, the vulnerability of neural networks to hand-crafted perturbations has also drawn attention due to the threat it poses to safety-critical applications. Thus, a myriad of techniques tailored to improve the robustness against artificially generated out-of-distribution examples has been proposed [196]. Although previous work has proposed to leverage advances in domain adaptation approaches to improve adversarial robustness [197, 198] and to mitigate the effect of distribution shifts by performing some type of adversarial training [199, 151], only a few contributions [200] attempted to devise strategies capable of dealing with both types of distribution shifts.

In this Chapter, we follow the contributions of Chapters 4 and 5 and employ random projections in the input of predictors with the aim of increasing the overlap in the support of different distributions. More specifically, we propose an efficient and unified framework to deal with both natural and artificial domain changes: *Randomly Projecting Out Distribution Shifts* (RPODS). Motivated by the earlier success of random projections for applications such as generative modeling [165, 77], data augmentation [201, 202], among others [203, 204, 205], we employ random data transformations as a means for distribution matching, i.e., we map input samples to a space where the overlap between domains is likely to be higher, thus decreasing the amount of available domain-specific information. As an extra practical contribution, the proposed approach further increases the robustness of a model to white-box adversarial attacks. That is, random projection layers are re-sampled prior to every prediction. As such, a subset of the model's parameters is always unknown to the attacker. Doing so limits the action of attacks that rely on previous knowledge about the model, while not harming the original accuracy, unlike methods that include adversarial examples at training time. The main contributions of this chapter are summarized in the following:

1. We propose RPODS, a principled and versatile approach to improve neural networks robustness to natural and artificial distribution shifts via random projections, without requiring domain labels and adversarial training;
2. We empirically confirm that random projections filter away domain-specific information by estimating the divergence between pairs of domains before and after projections;
3. We challenge the versatility of RPODS by performing experiments in two settings where distributions shifts are present, namely domain generalization and adversarial robustness, and show that RPODS outperforms approaches tailored to tackle either one of the settings.

This Chapter is organized as follows: In Section 6.3 we provide an overview of the related literature. We follow to the theoretical motivation behind our contribution and introduce the proposed approach to leverage random projections in Section 6.4. In Section 6.5 we provide empirical evidence supporting the use of random projections to mitigate distribution shifts as well as the evaluate the proposed approach on domain generalization and adversarial robustness tasks. Finally, we end this chapter by presenting the conclusions in Section 6.5.

6.3 Related work

In this Section we provide an overview of previous work that also relied on random convolutions to improve robustness of neural networks to *either* natural or adversarial perturbations, highlighting their main differences with respect to our work. Xu et al. [202] proposed a data augmentation based on multi-scale random convolutions for tasks under the domain generalization setting where a single domain is available at training time. The introduced RandConv consists in augmenting the training data with the output of convolutional layers which are randomly initialized and not updated during training. Best results were found by mixing the original input with the output of the random convolutions. The use of such approach to generate augmentations is motivated by the intuition that augmented images will present diverse types of texture, while preserving information which is relevant for detecting objects, such as global shapes. Previous work has also shown that techniques based on random convolutions can also be promising for improving robustness to adversarial perturbations via a data augmentation scheme [201]. Similarly to RandConv, in [201], a set of fixed random convolutions is computed offline, prior to training, and used to augment the original training set. Notice that, as RandConv, in this case the output size of the random convolutional layers must match the size of the original images.

It is important to highlight that both aforementioned strategies to leverage random convolutions for improving the robustness of neural networks are fundamentally different from our proposed approach in both motivation and implementation. While in [202] the use of random projections is intuitively motivated, we propose a principled approach by taking into account theoretical results that prove why the use of random projections is helpful for improving out-of-distribution generalization. Moreover, the implementation of our approach greatly differs from [201] and [202] mostly due to the following aspects: the original examples are not taken into account during training, and convolutional layers are re-sampled at every iteration in order to provide improve robustness to white-box adversarial attacks.

6.4 Using random projections to mitigate distribution shifts

In this section, we motivate the use of random projections to mitigate distribution shifts, and introduce the proposed approach that leverages such result on neural networks.

Similarly to Chapters 3 and 4, we consider settings where a predictor $h : \mathcal{X} \to \mathcal{Y}$ must present a good generalization performance on different domains, including those not available at training time. As in Chapter 4, we tackle the domain generalization setting [23, 22] and we are concerned with cases where multiple domains are available and the covariate shift assumption holds (i.e. the marginal distributions differ while data-conditional label distributions remain unchanged for all the considered domains). We assume that adversarial perturbations induce a shift in the marginal distribution of the original data[1].

Consider the input space $\mathcal{X}$ is the d-dimensional ball B^d of radius 1 centered at 0 and define the support of a domain $\mathcal{D}$, $supp(\mathcal{D}) \subset B^d$, as the set where the corresponding density is greater than some small threshold. Now let W represent a random projection and $\mathcal{D}_W$ represent the marginal of $\mathcal{D}$ along W and $\mathcal{X}_W$ denote the projection of the input space. The following result shows that the support of a projected domain occupies a higher fraction of the projected input space volume.

Theorem 5. *(Neyshabur et al. [165]) Assume $\mathcal{D}(x) = \sum_j \tau_j \mathcal{N}(x|\mu_j, \Sigma_j)$ is a mixture of Gaussians, such that there is no overlap between the supports or the projections of the components. If $supp(\mathcal{D}) \subset B^d$ and $Vol(supp(\mathcal{D})) > 0$, then, with high probability:*

$$\mathrm{Vol}\left(\mathrm{supp}\left(\mathcal{D}_W\right)\right)/\mathrm{Vol}\left(\mathcal{X}_W\right) > \mathrm{Vol}\left(\mathrm{supp}\left(\mathcal{D}\right)\right)/\mathrm{Vol}\left(\mathcal{X}\right). \tag{6.1}$$

Theorem 5 shows that random projections increase the overlap between the support of distributions over the same input space and thus can be used to mitigate covariate shifts. More specifically, in case two domains over $\mathcal{X}$, $\mathcal{D}^1$ and $\mathcal{D}^2$, are considered, the projection W acts in such a way that it likely increases the overlap between both domains. In the next section, we empirically confirm this observation by showing that the $\mathcal{A}$-distance [97], a proxy for the $\mathcal{H}$-divergence, that accounts for mismatches between distributions over the input space, is indeed decreased when estimated over projected inputs, i.e., $d_{\mathcal{A}}(\mathcal{D}^1, \mathcal{D}^2) > d_{\mathcal{A}}(\mathcal{D}^1_W, \mathcal{D}^2_W)$.

6.4.1 Adversarial robustness as domain generalization

Let S be set of allowed perturbations $S \subset \mathbb{R}^d$. In the case of images, S is such that it captures perceptual similarity between the original image and its respective perturbation. In order to obtain

[1] We consider the same notation introduced in Chapter 2: the input space is represented by $\mathcal{X} \subset \mathbb{R}^d$, while $\mathcal{Y}$ denotes the label space.

a predictor which is *robust* to adversarial perturbations, previous work [206] defined a modified version of the risk minimization objective as introduced in Chapter 2 to account for perturbations that incur in maximum values of loss. More specifically, the risk of a hypothesis h on a domain $\mathcal{D}$ is now defined as:

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\max_{\epsilon\in\mathcal{S}}\mathcal{L}[h(x+\epsilon),f(x)]\right]. \tag{6.2}$$

In comparison with the standard risk minimization framework the major practical difference between the *robust risk* defined in Equation 6.2 is that instead of computing the risk directly on samples from $\mathcal{D}$, the adversary is allowed to maximally perturb each sample x before a prediction is performed. Based on the aforementioned definition of robust risk, we claim the process of computing such quantity induces a new domain $\mathcal{D}^{adv}$. Sampling from $\mathcal{D}^{adv}$ yields adversarial perturbations x^{adv} rather than the original input x. In this work, we evaluate the proposed approach in a setting where no samples from $\mathcal{D}^{adv}$ are available at training time, i.e., no adversarial training is allowed. This setting can be seen as single-source domain generalization, where a model is trained on $\mathcal{D}$ and expected to also generalize well on $\mathcal{D}^{adv}$.

In practice, the process of sampling from $\mathcal{D}^{adv}$ given x can be seen as finding an adversary x^{adv} within a p-norm ball centered in x which will yield an incorrect prediction, i.e. $f(x^{adv})=y^{adv}\neq y$, or, formally:

$$\begin{aligned}\min_{x^{adv}}\quad & ||x^{adv}-x||_p\\ \text{s.t.}\quad & f(x)=y\\ & f(x^{adv})\neq y.\end{aligned} \tag{6.3}$$

Adversarial attacks to artificial neural networks can be either *black-box* or *white-box*, where the adversary, respectively, does not have, or has knowledge of the model architecture, parameters and its training data. In addition, the attacks could be *targeted* and *non-targeted*, which means that it aims to yield prediction mistakes to a specific, or to an arbitrary class except the correct one, respectively. In this work, we consider white-box and non-targeted adversarial examples.

6.4.2 Implementation

We consider applications of neural networks and implement our approach to Randomly Project Out Distribution Shifts (RPODS) using convolutional layers. More specifically, we consider a bank of

K projections such that $\mathcal{X}_W \subset \mathbb{R}^m$ and each random projection matrix $W_k \in R^{d \times m}$, $k = \{1, \ldots, K\}$, has entries drawn from a Gaussian distribution. In all of our experiments, we considered $\mathcal{N}(0, \sigma^2)$, where σ is set as per the scheme introduced in [207]. In order to prevent the resulting projections to be drastically distorted, we project the parameters of the convolutional layers to the L2 unitary ball. A model is then trained considering examples in the projected input space $\mathcal{X}_W$. Figure 6.1 illustrates the use of RPODS and shows examples from the PACS [120] dataset in the projected space.

6.4.2.1 Re-initializing the projections for improved robustness to white-box attacks

We further highlight that RPODS induce a further benefit in terms of improving a models' robustness to attacks that rely on knowledge about the model parameters (i.e., the white-box access model). By re-sampling the projection matrices at every iteration, in addition to having an input where distribution shifts are reduced, a part of the model parameters is constantly changing and, therefore, the attacker will never have access to the complete model when generating adversaries.

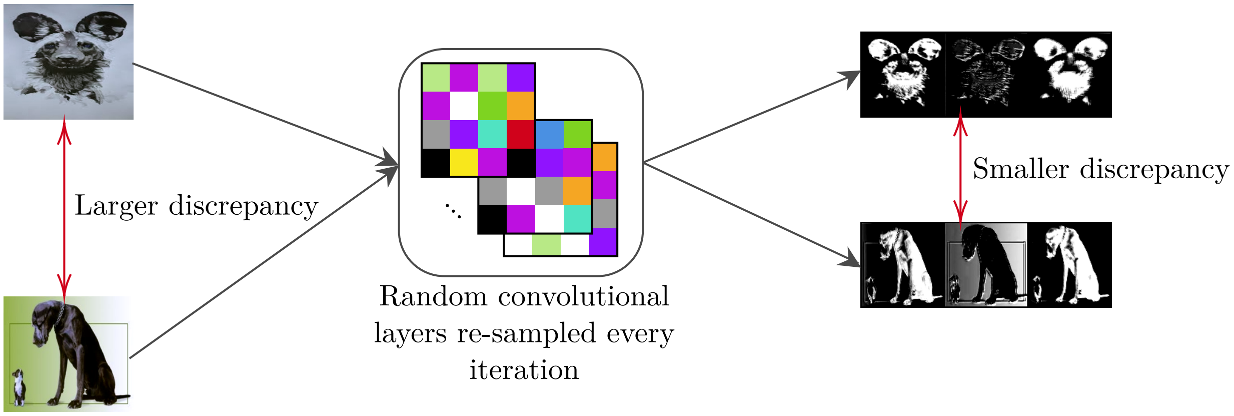


Figure 6.1 – Illustration of the proposed approach. Input images correspond to examples from the "Photo" and "Art Painting" domains from the PACS dataset and were projected via random convolutions to a space where discrepancies between domains are reduced.

6.5 Experiments

In this section, we empirically show that, as stated by Theorem 5, the use of RPODS in fact helps to mitigate distribution shifts, and evaluate the capability of RPODS to improve robustness to artificial and natural shifts in practical scenarios. In the case of natural domain shifts, we consider the domain generalization setting under a *leave-one-domain-out* scheme. We thus train a model with RPODS via empirical risk minimization with examples drawn from training distributions while evaluating it on an unseen domain. Finally, we show that RPODS are also able to improve robustness to adversarial attacks. For that, we train a model on the CIFAR-10 dataset and evaluate it on adversarial perturbations. In all cases, we compare RPODS with methods tailored to mitigate either natural or artificial shifts.

6.5.1 Random projections decrease domain divergences

We consider the PACS dataset and a ResNet-18 as backbone architecture. To evaluate whether randomly projecting distributions can in fact help to reduce distribution shifts, we estimate the $\mathcal{A}$-distance for each pair of domains within the PACS dataset, and compare the values obtained with raw inputs versus projected inputs using RPODS. To do so, we train a ResNet-18 to predict domain labels, and use the error rate on the test set to estimate the distances (refer to the Appendix for experimental details). Figures 6.2 and 6.3 show the $\mathcal{A}$-distance values for all pairs of domains for a model with and without RPODS, respectively. Each entry in the matrix depicted in the figure represents a value of distance computed considering a pair of domains, which are indicated by their initials (i.e. the “Sketch” domain is denoted by “S”).

We remark that, for all pairs of domains within the PACS dataset, the use of RPODS consistently decreases divergences between domains in comparison to raw data. This result is in-line with Theorem 5, providing further evidence that randomly projecting the input favours a decrease in domain discrepancies. Therefore, RPODS are suitable for domain generalization/adaptation settings, since previous work [17, 18, 29] showed that encodings resulting in smaller $\mathcal{A}$-distance between domains yield out-of-distribution generalization. Based on the examples of projected inputs provided in Figure 6.1, we argue that RPODS act by removing domain-specific information such as texture, and thus enforces a model to focus on higher-level features such as shape, which are more uniform across different domains.

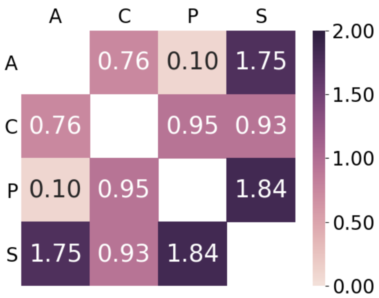


Figure 6.2 – $\mathcal{A}$-distance for pairs of domains on PACS estimated by a ResNet-18 with RPODS.

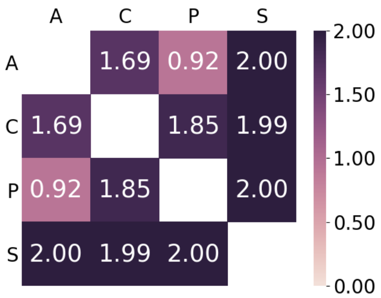


Figure 6.3 – $\mathcal{A}$-distance for pairs of domains on PACS estimated by a ResNet-18 on raw inputs (without RPODS).

In order to further highlight that the use of RPODS helps to mitigate domain divergences, we plot in Figure 6.4 the values of $\mathcal{A}$-distance obtained by a training a model on the original input space versus $\mathcal{A}$-distance values achieved by a model trained on top of the projected space (i.e. with RPODS) for each pair of domain. Each value is indicated by an "x", and makers lying below the dashed line indicated the cases where the $\mathcal{A}$-distance was higher for the model trained on the original input space. Notice that all points lie below the diagonal, indicating that, for all studied cases, RPODS were able to mitigate estimated domain shifts. Finally, we provide additional results illustrating the effect of RPODS on the input images. In Figure 6.5, we present examples of images from the class "dog" for all domains in the PACS dataset, along with their respective projected versions obtained with a bank of three random projections. By comparing images in the original and projected spaces, it is possible to notice that domain specific content (e.g. texture when comparing "Cartoon" and "Art painting") was removed by RPODS.

6.5.2 Domain generalization

To evaluate RPODS on the domain generalization setting, we once more consider the PACS dataset and a ResNet-18 architecture. Results in Figures 6.2 and 6.3 showed that RPODS reduce mismatches between data marginal distributions. Now, we are interested in verifying whether the projected input space preserves enough task-related information so that a model trained with RPODS is still capable of predicting class labels. For that, we compare the out-of-domain accuracy achieved by a model with RPODS to several approaches tailored to the domain generalization setting, as well as a model trained via standard empirical risk minimization (ERM). Following

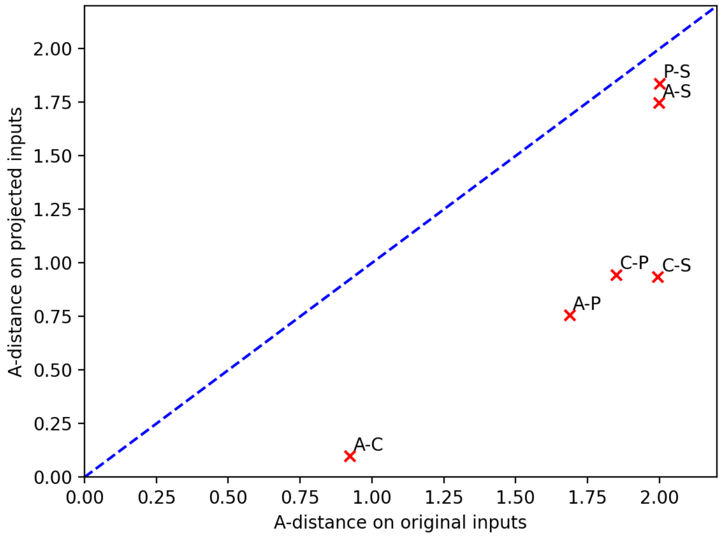


Figure 6.4 – Pair-wise domain $\mathcal{A}$-distance values for all domains within the PACS dataset. Each 'x' indicate the distance values for a pair of domains. The x-axis represents the distance estimated on the original input space, while the y-axis corresponds to distance values computed with RPODS. Points lying below the dashed line indicate a decrease in $\mathcal{A}$-distance when using RPODS.

recent work [208], we consider a ResNet-18 *trained from scratch* in order to favor a fair comparison with previous approaches, i.e. the impact of pre-training in final performances is ruled out.

In Table 6.1, we report the out-of-domain performance of models trained with RPODS on the source domains (e.g., results under column "Photo" correspond to models trained on "Art" + "Cartoon" + "Sketch"). We report the average performance across three independent training runs of the model when it presented its best in-domain accuracy (c.f. model selection protocol called *training domain validation set* in [172]). Baselines correspond to standard classifiers (denoted ERM) as well as recent methods specifically designed to tackle the domain generalization setting: self-challenge (SC) [209], Group DRO [210], GNN-Tag [208], and MLDG [152]. Further experimental details and results, including confidence intervals and other model selection criteria are presented in the Appendix. Results show that RPODS exceed the performance of the majority of the considered baselines in all domains and presents the highest average accuracy on PACS, showing that performing ERM on top of random projected input spaces improves out-of-distribution generalization.

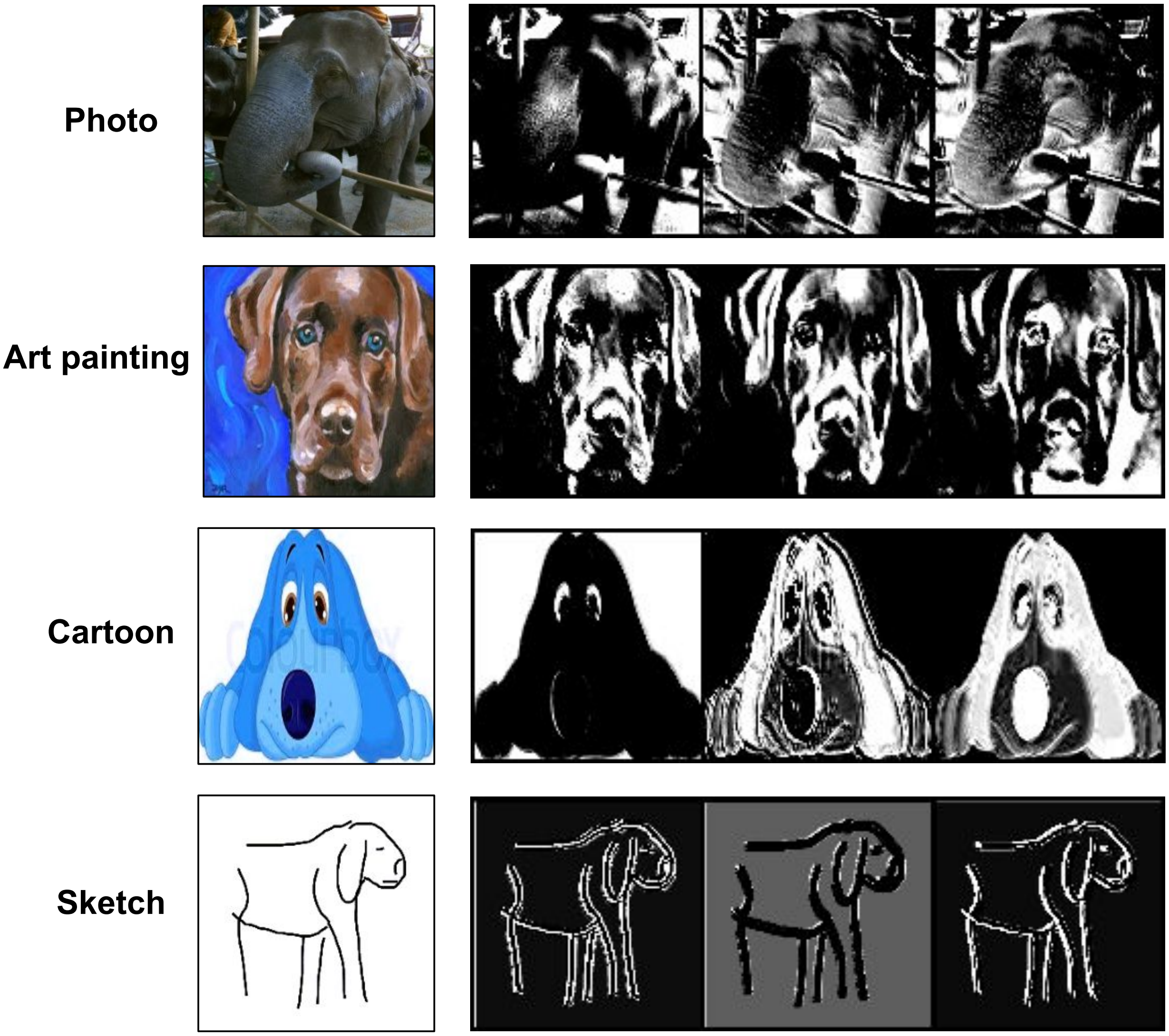


Figure 6.5 – Examples of randomly projected inputs obtained with a bank of 3 projections for each domain of the PACS dataset. The left-most column shows examples from the class "dog" in the original input space and the three remaining columns show the respective example in the projected space.

6.5.3 Adversarial robustness

Lastly, we evaluate the performance of RPODS against white-box adversarial perturbations. For that, we train a wide-ResNet [206] with RPODS on the CIFAR-10 dataset for 600 epochs, and report the robust accuracy of the model with best validation performance. We consider FGSM [211] and PGD [206] attacks under L_∞ budgets. In the following, we provide a brief description of the attack strategies considered in our evaluation.

Table 6.1 – Domain generalization results on PACS considering a leave-one-domain-out training scheme using the accuracy on the validation set of the training domains as model selection criterion. We also report the results of RPODS obtained by considering the "Oracle" criteria where accuracy on a partition of the unseen domain is considered. The * indicates results reported in [202].

Method	Selection	Photo	Art	Cartoon	Sketch
SC*	Training domain val. set	55.02	42.38	53.28	37.15
GroupDRO*	Training domain val. set	51.20	32.20	37.30	35.70
Episodic-DG*	Training domain val. set	41.13	29.83	42.15	37.69
Jigsaw*	Training domain val. set	42.34	30.37	45.65	29.14
MLDG*	Training domain val. set	47.30	29.30	40.30	28.80
ERM*	Training domain val. set	14.07	11.31	15.72	20.69
GNN-Tag*	Training domain val. set	53.23	33.26	49.16	54.15
Ours - RPODS	Training domain val. set	63.90 ± 0.78	42.63 ± 0.42	51.74 ± 1.74	56.67 ± 1.39
Ours - RPODS	Oracle	64.15 ± 0.11	38.34 ± 1.64	52.65 ± 0.52	58.18 ± 2.26

Fast Gradient Sign Method (FGSM). FGSM [211] computes adversarial attacks by perturbing each pixel of a clean sample x by ϵ on the direction of the gradient of the training loss $\mathcal{L}$ (usually the categorical cross-entropy) given the true label y with respect to the input:

$$x^{adv} = x + \epsilon \cdot \text{sign}[\nabla_x \mathcal{L}(h(x), y)]. \tag{6.4}$$

In this case, ϵ corresponds to the attack budget, i.e. the allowed distortion on x to yield x^{adv}. A high ϵ will increase the attacker's success rate, but it will be easier to detect.

Projected Gradient Descent (PGD). PGD [206] is a multi-step variant of FGSM and computes an adversarial attack given an input x over a total of T iterations according to the following update rule:

$$x^t = \Pi_{x+S}\left(x^{t-1} + \alpha \cdot \text{sign}\left(\nabla_x \mathcal{L}(h(x^{t-1}), y)\right)\right), \tag{6.5}$$

where Π_{x+S} is a projection operator onto the set S of available perturbations and α is the step size. The adversarial attack x^{adv} is obtained after this update is applied T times, i.e., $x^{adv} = x^T$.

We compare RPODS performance with approaches that *have access to adversarial examples at training time*, namely: adversarial training (AT) [206], adversarial logit pairing (ALP) [212], triplet loss adversarial training (TLA) [213], and TRADES [214]. In Table 6.2 we report the performance obtained by each one of these approaches as well as an undefended model, denoted as ERM. The column "General" in the Table indicates whether the evaluated approach was designed to specifically improve adversarial robustness, i.e., if it adversarial examples are considered at training time.

Table 6.2 – Adversarial robustness evaluation in term of accuracy (%) considering PGD and FGSM attackers under a L_∞ budget of $\frac{8}{255}$ for the CIFAR-10 dataset. The number of steps employed for each attack is represented within parenthesis.

	General	Clean	PGD (7)	PGD (20)	FGSM (1)
ERM	✓	95.01	0.00	0.00	13.35
AT	✗	87.14	55.63	49.79	45.72
ALP	✗	89.79	60.29	51.89	48.50
TLA	✗	86.21	58.88	53.87	51.59
TRADES ($1/\lambda = 1$)	✗	88.64	-	49.14	48.90
TRADES ($1/\lambda = 6$)	✗	84.92	-	56.61	56.43
RPODS	✓	89.70	75.62	46.35	47.49

Results show that RPODS achieve better accuracy on clean samples than most of the baselines and competitive robust accuracy for both attacks. Moreover, when compared with the undefended model (ERM), we observe that RPODS greatly improve the robust accuracy despite the fact it does not have access to adversarial examples at training time.

6.5.4 Experimental details

6.5.4.1 A-distance estimation

In order to estimate the $\mathcal{A}$-distance, we consider a hypothesis class corresponding to all models parameterized by a ResNet-18. We train both models with SGD with a learning set to equal to 0.001 and weight decay parameter equal to 0.00001. We report the accuracy on the validation partition of each domain after 10 training epochs.

6.5.4.2 Domain generalization

We implemented RPODS and run the experiments on the domain generalization setting using DomainBed [172] with the following hyperparameters:

- Batch size: 32
- Iterations: 5000
- Learning rate: 5e-4
- Dropout: 0.0
- Number of random projections: 3

- Random projection kernel size: 8
- Random projection stride: 1
- Weight decay: 0.0

6.5.4.3 Adversarial robustness

We trained a ResNet with SGD using the hyperparameters reported below. Attacks were implemented using *FoolBox*[2].

- Batch size: 64
- Epochs: 600
- Initial learning rate: 0.1
- Schedule: Decay the learning by a factor of 10 at epochs $[10, 150, 250, 350]$.
- Number of random projections: 3
- Random projection kernel size: 3
- Random projection stride: 1
- Weight decay: 0.0005
- Dropout probability: 0.3

6.6 Conclusion

We introduced RPODS – a simple and efficient approach to mitigate the effects of distribution shifts on neural networks performance. In practice, RPODS project the input space via a bank of random projections, implemented as a convolutional layer added to the input of a model, with weights re-sampled at every iteration. We show that RPODS improve out-of-distribution generalization in scenarios where distribution shifts stem from different sources. More specifically, experiments on the PACS dataset showed that RPODS improve upon a number of approaches tailored to the domain generalization setting, improving the average accuracy on unseen domains by almost 6.8% with respect to the best performing baseline. We also evaluated RPODS in a setting where domain shifts were given by adversarial perturbations and showed that, despite its simplicity, RPODS greatly improved robustness to white-box attacks on the CIFAR-10 dataset in comparison

[2] `https://foolbox.readthedocs.io/en/stable/index.html`

to the undefended model. Notably, models employing RPODS are competitive when compared to adversarial training approaches, specifically designed to attenuate the effects of adversarial perturbations.

Chapter 7

Conclusions and future research directions

7.1 Summary of contributions

In this dissertation, we made contributions towards designing general learning systems capable of succeeding in real-world scenarios. We considered multiple settings and tasks, and provided empirical evidence that our findings also generalize to real-world applications of machine learning, such as passive brain-computer interfaces. In Chapter 3, we investigated and developed strategies to quantify distribution shifts and evaluated on a EEG-based mental workload assessment scenario considering the WAUC database, a dataset we introduced in Chapter 2. In Chapter 4, we proposed an approach to improve neural networks' generalization ability in the scenario where domain shifts are observed, as well as devised generalization guarantees for the risk on unseen distributions. The proposed method, G2DM, is based on multiple domain discriminators, random projections, and a considers a multi-objective approach to aggregated loss functions. In Chapter 5, we perform an in-depth investigation of such architecture and study different approaches to aggregate the losses provided by the discriminators. Notably, we consider different application: improving the stability of the training of GANs for generative modeling and show that gradient-based multi-objective optimization yields better perceptual quality as well as diversity of generated samples in several cases. We finalize our quest for general and robust learning systems in Chapter 6, by building on the findings of Chapters 4 and 5, and propose to use the random projection layers that were previously used in order to improve the training stability of adversarial games with a completely different purpose. We show that such projections can be used to improve the robustness of neural networks to domain shifts, and proposed RPODS, an simple and versatile approach to mitigate the effects of both natural and adversarial distribution shifts.

7.2 Conclusions

7.2.1 Cross-subject generalization on BCIs

We presented in Chapter 3 the first steps towards better understanding the cross-subject variability phenomena on passive EEG-based BCIs from a statistical learning perspective. We looked at this problem through the lens of domain adaptation and proposed strategies to estimate distributional shifts between conditional and marginal distributions corresponding to the data generating process of features and labels from different subjects. To evaluate the proposed approach, the WAUC dataset, introduced in Chapter 2, was used and binary mental workload assessment from EEG power spectral features was performed. Our analysis showed that feature normalization, as well as data collection conditions such as the equipment used to induce physical workload, had a relevant impact on the estimated values of conditional shift. In case the goal is to improve out-of-distribution performance, normalization procedures that decrease the overall cross-subject conditional shift should be prioritized since they yield smaller generalization gaps. Our analysis showed that z-score normalization provided the best strategy for normalizing EEG power spectral density features. Moreover, such normalized feature spaces should be considered in case representation learning methods based on domain adaptation are used to learn domain-invariant classifiers on top of features.

In case end-to-end representation learning approaches are employed on BCIs, we show in Chapter 4 that the introduced approach, G2DM, can also be considered for obtaining improved cross-subject generalization without requiring any calibration step. The empirical evaluation of G2DM on the SEED dataset for affective state prediction showed that the proposed approach is capable of successfully leveraging the available subject labels (which in this case comes with no additional effort at the data collection) and presents an improvement of more than in comparison to ERM and, more importantly, it outperforms domain adaptation approaches (which resemble a BCI calibrated for a specific subject).

7.2.2 Domain Generalization

In Chapter 4, we made contributions to the domain generalization setting and showed that out-of-distribution generalization can be achieved in the neighborhood of the set of mixtures of distributions observed at training time. We first show in Lemma 1 that the $\mathcal{H}$-divergence between any pair of domains within the convex hull of the training distributions can be bounded the maximum $\mathcal{H}$-divergence between the source domains. We built on this result to derive a generalization bound in Theorem 4 for *any* domain and show that the risk measured on such unseen domain depends on a convex sum of the risks on the sources, the maximum $\mathcal{H}$-divergence between sources, the $\mathcal{H}$-divergence between the considered unseen domain and its projection to the convex hull, and the mismatch between labelling functions in case the covariate shift assumption does not hold. Aiming to minimize the terms of the introduced bound, we devise G2DM, an adversarial approach so that pairwise domain divergences are estimated and minimized. G2DM contains several practical innovations in comparison to previous adversarial approaches such as the use of random projection layers prior to domain discriminators. We show that the representations learned by G2DM are capable of discarding domain-specific information, which indicates that such representations are well-suited for problems where distribution shifts are likely to be observed. We confirm this hypothesis on multiple scenarios, such as object recognition tasks on the PACS and VLCS benchmarks, as well as the challenging EEG-based affective state prediction. In the case objective recognition, our experiments showed that fine-tuning pre-trained convolutional architectures such as ResNet and AlexNet using G2DM instead of ERM yields higher out-of-distribution accuracy for most of the considered cases. Moreover, our results on affective state prediction showed that even in cases where a pre-trained architecture is not employed, G2DM presented an improved performance in comparison to multiple baselines, including domain adaptation techniques that have access to a sample from the unseen domain.

We further contribute to the domain generalization setting in Chapter 6. We proposed RPODS, a versatile and efficient strategy for improving the robustness of neural networks. RPODS do not rely on domain labels and leverages convolutional layers randomly initialized and re-sampled at every iteration in order to learn representations in a space where the overlap between distributions is larger and prevent full access of the model to a potential attacker. We provided empirical evidence to support the claim that such random projections are able to achieve this goal by showing that the use of RPODS consistently decreases divergences between domains for the PACS dataset when compared to raw data. This indicates that RPODS act by removing domain-specific information such as texture, and enforce a model to focus on higher-level features such as shape, which are more related to class labels in the considered case. The experiments on the PACS dataset confirmed this hypothesis by showing that RPODS improved the performance on most of the considered evaluation scenarios. Despite its simplicity, RPODS greatly improved robustness to white-box attacks on the CIFAR-10 dataset in comparison to the undefended model and, more importantly, it achieved competitive performance when compared to approaches specifically designed to attenuate the effects of adversarial perturbations.

7.2.3 Generative Modeling

In Chapter 5 we proposed to use the same contributions of G2DM for a different application: generative modeling. We considered the training of GANs with multiple discriminators and showed that the training of the generator can be seen as multi-objective problem where each objective corresponds to the loss of a discriminator. We thus proposed the use of gradient-based multi-objective optimization techniques to update the parameters of the generator and exploited well known methods such as the multiple gradient descent algorithm. However, due to MGD's prohibitively high cost in the case of large neural networks, we propose to use a more efficient alternative, namely, the hypervolume maximization algorithm which optimizes a weighted loss such that the generator's training will assign greater importance to feedback from discriminators against which it performs poorly. The proposed approach was observed to consistently yield higher quality samples in terms of FID when compared to average loss and other aggregation rules for the losses. We further observed a higher number of discriminators to increase sample diversity and generator robustness to hyper-parameters values. Such approach for training GANs was also shown to succeed in the generation of higher-resolution images and to yield competitive performance in terms of metrics such as FID and Inception Score to other methods to stabilize the training such as WGAN-GP and SNGAN.

7.2.4 Multi-objective optimization

In terms of multi-objective optimization, our contributions of Chapter 5 revealed that employing gradient-based approaches for solving multi-objective problems should be considered when multiple conflicting loss functions are required to be minimized. We showed that despite the increased computational cost and challenges such as noisy gradient estimates and non-convexity of the objectives, the multiple gradient descent algorithm is capable of descent directions that yield better performance than naive strategies such as linear scalarization of the losses. We also proposed a way to employ the more efficient hypervolume maximization approach to train GANs and showed that with a proper adaptation of the nadir point, this algorithm consists of a better strategy to find solutions in the central region of the Pareto front, which, in practice, induces better convergence for all the discriminators, since all loss functions will be assigned overall equal importance. Finally, we showed that previous approaches proposed to train GANs with multiple discriminators can be seen as gradient-based multi-objective optimization methods with different strategies to find a descent direction for the objectives.

7.3 Future Research Directions

In the following, we highlight further research questions within the topics considered in the thesis, as well as limitations of our contributions which can also be addressed by future investigation.

7.3.1 Cross-subject generalization on BCIs

Our contributions to understanding and mitigating cross-subject disparities on BCIs can be further extended by employing the developed strategies to estimate distributional shifts in order to better inform the development of methods for applications involving EEG signals. This could be achieved by, for example, performing feature selection in such a way that the features included are the ones which yielded lower values of overall distribution shifts. Moreover, we believe our proposed estimators can be used to shed light on the reasons behind the success of representation learning approaches such as [215] which showed promising performances on unseen subjects in BCIs applications despite not including explicit mechanisms to mitigate distribution shifts.

7.3.2 Domain Generalization

In future work, we intend to investigate if the assumptions on the data generating process for the domain generalization setting can yield PAC-like results for domain complexity in a meta-distribution-agnostic fashion, i.e., we intend to assess questions such as: how many source domains are needed to guarantee low meta-risk with high probability? In terms of future investigations regarding our proposed algorithm G2DM, exploring different strategies to estimate pairwise $\mathcal{H}$-divergence values can be investigated. As an example, a single discriminator could be employed to distinguish whether two examples belong to the same domain. Notice that, although more efficient, this strategy does not directly estimate $\mathcal{H}$-divergence values and this aspect should also be considered in further investigations. In addition, we believe that investigating strategies to tackle the need of using domain labels in G2DM should also be considered in order to enlarge the scope of applications where it can be considered. Future work within the domain generalization setting include exploring the use of RPODS in situations where robustness to natural and adversarial distribution shifts is simultaneously required, as well as other out-of-distribution generalization settings such as single-source domain generalization and domain adaptation. Moreover, it is worth to mention that investigating the performance of RPODS in more favorable scenarios for the attacker is also a future research direction. As an example of such a scenario, the attacker could be allowed to perform multiple queries to the random projections' parameters when computing adversarial perturbations.

7.3.3 Generative Modeling

Our contributions to improving the training of generative adversarial networks unlocked two main future research avenues. First, we hypothesize further investigation of the role of the norm of the descent direction, $||\sum_{k=1}^{K}\alpha_k \nabla l_k||$, is necessary since this quantity indicates how "close" to achieve Pareto stationarity a solution is. Given that, this quantity could be used as a penalty added to the generator's objective in order to speed-up the convergence to a Pareto stationary point, which could be helpful for decreasing the amount of discriminators necessary to reach certain performance level. Moreover, another avenue of future investigation consists in developing strategies for parallelizing the training of the generator in order to the decrease the computational cost overhead resulting from having multiple discriminators. For example, by having the parameters of each discriminator on a single GPU, computing the generator's objective could be almost as fast as in the single-discriminator case, which could allow employment of our proposed multi-discriminator setting for training architectures capable of learning high dimensional distributions, such as BigGAN [216].

7.3.4 Multi-objective optimization

The promising results obtained for training GANs with gradient-based multi-objective algorithms ked us to conclude these techniques can also be applied to other machine learning problems were multiple conflicting losses are required to be minimized. As an example, we consider settings such as multi-task learning, domain adaptation, and continual learning as instances where multi-objective optimization could play an important role to provide issues such as catastrophic forgetting in continual learning methods [217]. Moreover, although hypervolume maximization presented a better compromise between computational cost and performance in comparison to MGD, an analysis of the convergence of this algorithm has not been performed yet. Therefore, we believe this analysis should be considered as part of future investigations. Finally, we highlight that the hypervolume maximization algorithm as proposed in Chapter 5 is tailored to applications where each objective varies within the same scale and we propose as a future research direction investigating strategies to alleviate such a limitation.

Chapter 8

Sur des réseaux de neurones robustes et génératifs avec des applications aux interfaces cerveau-ordinateur et à la reconnaissance d'objets

8.1 Introduction

8.1.1 Succès et pièges de l'apprentissage automatique

Il a été démontré que les techniques d'apprentissage automatique permettent l'automatisation de plusieurs tâches [1, 2, 3, 4] en apprenant à partir de collections de points de données liés à l'objectif souhaité. Ces dernières années, la croissance rapide de la quantité de données disponibles rendue possible par l'exploitation d'outils tels qu'Amazon Mechanical Turk et de plateformes telles que les réseaux sociaux [5], a facilité le développement de techniques capables d'augmenter les degrés de automatisation des systèmes d'apprentissage[1]. En particulier, suivis d'une avancée sur le matériel disponible pour exécuter des programmes informatiques en parallèle, les réseaux de neurones profonds [8, 9] sont apparus comme une alternative pour tirer parti des ensembles de données disponibles pour développer des systèmes d'apprentissage qui nécessitent un minimum de pré-traitement sur les entrées [8] et atteint des niveaux de performances sans précédent sur des tâches dans une variété de domaines d'application [10, 11, 12, 13].

[1] Veuillez vous référer aux travaux de Birhane et al. [6, 7] et Crawford [5] pour une étude approfondie des dommages causés par de telles approches pour acquérir des ensembles de données et comment elles renforcent les relations de déséquilibre de pouvoir actuelles et affectent négativement les populations sous-représentées.

L'entraînement des réseaux de neurones et de nombreux systèmes d'apprentissage automatique est généralement effectué en suivant le paramètre minimisation des risques empirique (*empirical risk minimization*, ERM) [14]. L'hypothèse principale dans ce cadre est que tous les exemples utilisés pour l'apprentissage et le test des prédicteurs sont tirés indépendamment d'une distribution fixe, c'est-à-dire l'i.i.d. hypothèse. Un certain nombre de garanties de généralisation ont été dérivées de cette hypothèse et ces résultats ont induit plusieurs algorithmes pour la résolution de problèmes d'apprentissage supervisé [15]. Malgré le succès des applications d'apprentissage automatique reposant sur le framework ERM, des limitations importantes dans ce cadre peuvent être soulignées: i) l'i.i.d. hypothèse est *invérifiable* [16] étant donné que l'on n'a pas accès à la distribution des données, et ii) elle ne tient pas compte des changements de distribution qui se produisent souvent dans la pratique. Des exemples représentatifs de ces décalages de distribution incluent des changements dans les conditions d'acquisition de données, telles que l'éclairage des images pour la segmentation d'objets, ou de nouvelles sources de données telles que des locuteurs invisibles lors de la reconnaissance vocale.

Un certain nombre de paramètres alternatifs ont ensuite été introduits afin de mieux faire face à des cas plus réalistes où *généralisation hors distribution*[2] est obligatoire. La minimisation des risques dans le cadre de *domaine adaptation*, par exemple, détend une partie de l'i.i.d. hypothèse en autorisant une distribution source (ou domaine)[3] ainsi qu'une distribution cible différente observé au moment du test. Les résultats d'adaptation de domaine introduits dans [17] ont montré que l'écart de généralisation en termes de différence de risque entre les deux distributions considérées pour un prédicteur fixe est borné par une notion de distance mesurée entre les domaines d'entraînement et de test. Bien qu'il soit moins restrictif que le cadre précédent, le cas d'adaptation de domaine est toujours limité dans la mesure où seules les paires de distributions observées pendant l'entraînement devraient générer un faible risque, et les changements au-delà de ces domaines entraîneront probablement de mauvaises performances. De plus, les algorithmes conçus pour ce paramètre **reposent sur l'accès au moment de l'apprentissage à un échantillon non étiqueté de la distribution cible** afin que les représentations puissent être apprises induisant une invariance entre les domaines train et cible [18]. Il s'agit d'un facteur limitant pour les applications pratiques où les données du domaine cible peuvent être inaccessibles; par exemple, un service de reconnaissance vocale ne peut pas être (re)formé sur les données obtenues à partir de chaque nouveau locuteur qu'il observe.

Malgré le succès des stratégies d'adaptation de domaine dans plusieurs scénarios d'application [19, 20, 21], nous allons plus loin dans ce cadre et considérons un cadre plus général qui est souvent appelé dans la littérature *généralisation de domaine* [22, 23]. Dans ce cas, on suppose qu'un ensemble de distributions sur l'espace d'entrée est disponible au moment de l'apprentissage. Au moment du

[2] Généraliser "en dehors" des distributions d'entraînement, c'est-à-dire maintenir le niveau de performance sur l'invisible distributions proches du niveau observé sur les distributions disponibles au moment de l'apprentissage.

[3] Nous utilisons les termes *domaine*, *distribution de données* et *source de données* de manière interchangeable dans tout le texte.

test, cependant, les deux distributions observées, ainsi que de nouveaux domaines invisibles, peuvent apparaître, et un faible risque doit être obtenu quel que soit le domaine sous-jacent. *les stratégies de généralisation de domaine visent à trouver un espace de représentation qui donne de bonnes performances sur de nouvelles distributions, inconnues au moment de l'apprentissage.*

En plus des changements de distribution naturels qui sont expliqués par les stratégies d'adaptation de domaine et de généralisation de domaine, des travaux antérieurs ont identifié que les réseaux de neurones sont également vulnérables aux perturbations artificiellement fabriquées à la main, connues sous le nom d'exemples contradictoires [24]. Ce type de perturbation est particulièrement nocif pour les applications du monde réel critiques pour la sécurité telles que les voitures autonomes [25], en raison du fait qu'il est conçu pour être imperceptible par les humains tout en créant un modèle appris pour classer mal l'entrée attaquée citekurakin2016adversarial. Malgré la prise de conscience par la communauté de l'apprentissage automatique de la susceptibilité potentielle des systèmes d'apprentissage actuels à de telles perturbations artificielles, ainsi que des problèmes causés par les changements de domaine naturels, la littérature manque encore de contributions visant à traiter simultanément les deux types de vulnérabilités. De même, un autre aspect d'amélioration des systèmes d'apprentissage actuels qui devrait être pris en compte est leur spécificité pour aborder des paramètres et des tâches particuliers [26]. Bien que différents problèmes d'apprentissage puissent être formulés de la même manière, les algorithmes développés pour les résoudre sont généralement liés à une application particulière comme objectif principal et ne sont pas explorés comme des alternatives pour résoudre d'autres problèmes. Par exemple, les formulations minimax, où plusieurs sous-ensembles des paramètres d'un modèle sont appris en optimisant différentes fonctions de perte, apparaissent comme la formalisation de plusieurs tâches dans la littérature [27, 28, 29, 30], cependant, peu d'algorithmes sont considérés comme des stratégies viables pour plusieurs de ces applications.

8.1.2 Domaines d'application abordés dans la thèse

L'électroencéphalographie (EEG) est une modalité principale pour surveiller les changements dans les états du cerveau. En enregistrant et en traitant les signaux EEG, il est possible de traduire l'activité neuronale et de l'utiliser, par exemple, pour le contrôle prothétique [31]. Les systèmes capables d'enregistrer, de traiter et de prendre des décisions sur la base des informations neuronales sont appelés interfaces cerveau-ordinateur (*Brain-Computer Interface*, BCI). Au cours des dernières décennies, l'intérêt pour les BCIs basés sur l'EEG s'est considérablement accru en raison de son potentiel d'impact positif [32] sur la vie de plusieurs personnes, par exemple en permettant une rééducation post-AVC plus engageante [33]. Les changements dans l'état du cerveau peuvent être déduits des enregistrements EEG et utilisés dans les BCIs via l'extraction d'un certain nombre de caractéristiques, notamment la densité spectrale de puissance [34, 35], la cohérence [36, 37, 38], et plus récemment mesures de modulation d'amplitude [39]. Des travaux récents [40] ont montré une augmentation de l'application des réseaux de neurones aux BCIs impliquant divers ensembles de tâches allant de l'imagerie motrice à la prédiction de l'état affectif. Malgré le succès observé des réseaux de neurones dans de telles applications, les BCIs basés sur l'EEG manquent de généralisabilité entre différents sujets, ou même entre différentes sessions d'enregistrement acquises à partir du même sujet [41].

Les facteurs anatomiques et environnementaux sont attribués comme les principales causes de la différence typique des réponses neuronales entre les individus sous le même stimulus [42, 43, 44]. De plus, de tels décalages entre les conditions d'entraînement et de test pourraient être dus à différents équipements de collecte de données, ainsi qu'à des changements dans le positionnement des électrodes au cours d'une session expérimentale. Un moyen standard de gérer la forte variabilité inter-sujets sur les applications basées sur l'EEG consiste à *calibrer* le modèle avant de l'appliquer à un individu invisible en collectant un certain nombre d'exemples étiquetés de ce sujet particulier et en recyclant le modèle en tenant compte de cela. nouvel échantillon [45]. Cependant, des travaux récents [46, 47] ont mis en évidence que l'étape de calibrage pourrait être trop coûteuse et lente. Améliorer la généralisation inter-sujets des BCIs actuels est donc essentiel pour permettre l'application de ces modèles dans des conditions réelles et des applications à fort impact telles que la surveillance de la charge de travail mentale. Une alternative à l'étalonnage des BCIs avant de les utiliser sur un nouveau sujet/une nouvelle condition consiste à utiliser des stratégies pour apprendre des modèles qui sont moins enclins à s'appuyer sur des informations spécifiques au sujet. Pour y parvenir, des travaux récents ont envisagé des techniques de gestion des décalages entre les distributions de entraînement et de test, telles que les approches d'adaptation de domaine [48, 49, 50, 51].

Malgré leur succès sur les tâches de vision par ordinateur lorsque des ensembles de données à grande échelle sont disponibles [63, 64, 65, 66], les réseaux de neurones présentent une diminution des performances face aux changement de distribution lorsqu'ils sont utilisés pour des tâches du monde réel [67]. Dans le cas des tâches de reconnaissance d'objets, par exemple, les changements dans les caractéristiques de bas niveau telles que la texture et l'éclairage sont suffisants pour embrouiller un modèle à un point où il n'est pas capable de produire des prédictions fiables [67, 68]. Les travaux précédents [69, 70, 71] ont montré que les réseaux de neurones peuvent être biaisés vers la capture de caractéristiques qui ne sont pas nécessairement *causant* des changements dans la classe d'objets. À titre d'exemple, considérons un scénario où l'objectif est de classer les chiens et les chats à partir de photos. Dans le cas de distracteurs tels que des jouets en forme d'os sont présents dans la majorité des images de chiens, cela pourrait être le cas où un réseau neuronal apprendra à distinguer les chiens des chats en recherchant de tels objets dans une image, plutôt que plus caractéristiques représentatives et généralisables telles que la forme des oreilles.

En plus des tâches de reconnaissance d'objets, les réseaux de neurones ont également des difficultés dans les applications à l'imagerie médicale où les conditions de collecte de données peuvent différer de l'entraînement au temps de test. Des changements de domaine peuvent être observés, par exemple, une fois que les données de nouveaux sujets ou équipements sont utilisées après l'apprentissage d'un modèle [72, 73]. Étant donné qu'une exigence majeure des modèles déployés consiste à ce qu'ils soient capables de généraliser même à des conditions non observées auparavant, c'est-à-dire qu'un modèle de segmentation tumorale ne peut pas être recyclé une fois que les données d'un nouveau patient sont observées, la généralisation à travers les domaines est devenue une direction de recherche pertinente.

8.1.3 Contributions de thèse

Dans cette section, nous donnons un aperçu des contributions de cette thèse, séparées par chapitre.

- Chapitre 2: Nous nous concentrons sur la fourniture de ressources permettant le développement de différentes stratégies d'évaluation de la charge de travail mentale dans des conditions réelles. Pour cela, nous avons introduit WAUC, une base de données multimodale d'Workload Aévaluation Ude l'aCactivité physique. L'étude a impliqué 48 participants qui ont effectué la NASA Revised Multi-Attribute Task Battery II dans trois conditions de niveau d'activité différentes. L'activité physique a été manipulée en changeant la vitesse d'un vélo stationnaire ou d'un tapis roulant.
- Chapitre 3: Nous proposons une stratégie pour estimer deux types d'écarts entre plusieurs distributions de données, à savoir les décalages marginaux et conditionnels, observés sur les données collectées auprès de différents sujets. En plus de élucider les hypothèses valables pour un ensemble de données particulier, les estimations des changements statistiques obtenus avec l'approche proposée peuvent être utilisées pour étudier d'autres aspects d'un pipeline d'apprentissage automatique, tels que l'évaluation quantitative de l'efficacité des stratégies d'adaptation de domaine. En particulier, dans ce chapitre, nous considérons les enregistrements EEG de l'ensemble de données WAUC en distinguant les individus qui ont effectué l'expérience en courant sur un tapis roulant ou en pédalant sur un vélo stationnaire. Nous avons exploré les effets de différentes stratégies de normalisation couramment utilisées pour atténuer la variabilité inter-sujets et avons montré les effets que différents schémas de normalisation ont sur les changements statistiques et leur relation avec la précision de la prédiction de la charge de travail mentale évaluée sur des participants invisibles au moment du train.
- Chapitre 4: Nous abordons le problème hors distribution en nous concentrant sur le cadre de généralisation de domaine: une formalisation où le processus de génération de données au moment du test peut produire des échantillons de distributions jamais vues auparavant. Notre travail s'appuie sur le lemme suivant: en minimisant une notion de discordance entre toutes les paires d'un ensemble de domaines donnés, nous minimisons également la discordance entre toutes paires de mélanges de domaines. En utilisant ce résultat, nous dérivons une borne de généralisation pour notre cadre. Nous montrons ensuite qu'un faible risque sur des domaines invisibles peut être obtenu en représentant les données dans un espace où (i) les

distributions d'entraînement sont indiscernables et (ii) les informations pertinentes pour la tâche à accomplir sont préservées. La minimisation des termes dans notre borne donne une formulation contradictoire qui estime et minimise les écarts par paires. Nous validons notre stratégie proposée sur des benchmarks de généralisation de domaine standard, surpassant un certain nombre de méthodes récemment introduites.

- Chapitre 5: Nous revisitons le paramètre de discriminateur multiple introduit au chapitre 4 pour l'entraînement de réseaux contradictoires génératifs en encadrant la minimisation simultanée des pertes fournies par différents modèles comme un problème d'optimisation multi-objectifs. Plus précisément, nous évaluons les performances de la descente de gradient multiple et de l'algorithme de maximisation de l'hypervolume sur un certain nombre d'ensembles de données différents. De plus, nous soutenons que les méthodes proposées précédemment et la maximisation de l'hypervolume peuvent toutes être considérées comme des variations de descente à gradient multiple dans lesquelles la direction de mise à jour peut être calculée efficacement.
- Chapitre 6: Nous proposons une approche unifiée et polyvalente pour atténuer les changements de distribution naturels et artificiels via l'utilisation de projections aléatoires. Nous montrons que de telles projections, lorsqu'elles sont mises en œuvre sous forme de couches convolutives avec des poids aléatoires placés à l'entrée d'un modèle, sont capables d'augmenter le chevauchement entre les différentes distributions qui peuvent apparaître au moment de l'apprentissage/du test. Nous évaluons l'approche proposée dans des contextes où différents types de changements de distribution se produisent.

8.1.4 Organisation de thèse

Dans cette thèse, il y a des contributions à différents aspects du pipeline Machine Learning. Nous commençons par introduire un nouvel ensemble de données pour l'évaluation de la charge de travail mentale basée sur l'EEG dans le chapitre 2, ainsi que le reste des antécédents requis pour les chapitres suivants. Nous procédons ensuite à l'introduction de notre première contribution algorithmique sur le domaine de la généralisation hors distribution dans le Chapitre 3 où nous proposons une approche pour estimer les décalages de distribution donnés à des échantillons de différents domaines et la validons sur l'ensemble de données WAUC introduit au Chapitre 2. Nous suivons nos prochaines contributions pour améliorer la généralisation hors distribution sur les réseaux de neurones en présentant au Chapitre 4 de nouveaux résultats théoriques pour le paramètre de généralisation de domaine et en introduisant un nouvel algorithme basé sur ces résultats qui repose sur l'utilisation de discriminateurs multiples et de projections aléatoires implémentées sous forme de convolutions avec des poids aléatoires. Dans le Chapitre 5, nous montrons que l'approche proposée pour la généralisation de domaine peut également être utilisée dans d'autres applications avec des objectifs d'apprentissage similaires. Nous considérons la modélisation générative des distributions de probabilité et montrons que l'algorithme peut être utilisé pour former des réseaux antagonistes génératifs. Nous explorons également en profondeur les différentes stratégies pour agréger les différentes fonctions de perte impliquées dans l'approche et proposons d'envisager des stratégies d'optimisation multi-objectifs basées sur le gradient pour l'apprentissage. Nous présentons ensuite notre contribution algorithmique finale dans le Chapitre 6 où nous proposons d'utiliser des convolutions aléatoires, comme dans les Chapitres 4 et 5, pour augmenter la robustesse des réseaux de neurones aux changements de distribution naturels et accusatoires en proposant une approche qui ne repose pas sur des étiquettes de domaine ni ne nécessite un entraînement accusatoire. Enfin, nous concluons la thèse au Chapitre 7 en résumant nos principales conclusions et contributions et en introduisant les futures directions d'investigation dérivées des travaux présentés tout au long de cette thèse.

8.2 Chapitre 2: Contexte

8.2.1 Minimisation des risques empiriques

Soit les données représentées par $\mathcal{X} \subset \mathbb{R}^d$, où d correspond à la dimension de l'espace d'entrée (ou caractéristique), et $\mathcal{Y}$ désigne le espace étiquette. Dans ce cas, les exemples correspondent aux couples $(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}$, tels que $y = f(x)$, et $f : \ mathcalX \to \mathcal{Y}$ est une fonction d'étiquetage déterministe. Un domaine est défini comme un tuple $\langle \mathcal{D}, f \rangle$ où $\mathcal{D}$ correspond à une distribution de probabilité sur $\mathcal{X}$[4] et f est la fonction d'étiquetage respective. Soit h une hypothèse définie comme une application $h : \mathcal{X} \to \mathcal{Y}$, telle que $h \in \mathcal{H}$, où $\mathcal{H}$ est un ensemble d'hypothèses candidates, et enfin définir le risque R associé à une hypothèse donnée h sur le domaine $\langle \mathcal{D}, f \rangle$ comme:

$$R[h] = \mathbb{E}_{x \sim \mathcal{D}} \mathcal{L}[h(x), f(x)], \tag{8.1}$$

où la perte $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to R_+$ quantifie à quel point une hypothèse h est différente de la vraie fonction d'étiquetage f pour une instance donnée (x, y).

8.2.2 Apprentissage supervisé comme minimisation des risques empiriques

Le problème de l'apprentissage supervisé peut être défini comme la recherche de l'hypothèse de risque minimum h^* au sein de la classe $\mathcal{H}$:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} R[h]. \tag{8.2}$$

Cependant, le calcul de $R[h]$ est généralement insoluble puisqu'on n'a pas accès à $\mathcal{D}$. En pratique, on ne dispose pour calculer h^* qu'un ensemble d'échantillons observés à partir de $\mathcal{D}$.

[4] Notez que dans le cas où nous considérons un modèle stochastique où la fonction d'étiquetage n'est pas déterministe, la distribution $\mathcal{D}$ serait définie sur $\mathcal{X} \times \mathcal{Y}$.

Étant donné le caractère insoluble du cadre de minimisation des risques décrit ci-dessus, la ERM est un cadre alternatif pratique courant pour l'apprentissage supervisé. Dans ce cas, un échantillon X de taille N est observé à partir de $\mathcal{D}$, soit $X = \{x_1, x_2, \ldots, x_N\}$, **où tout x_n, $n = \{1, \cdots, N\}$ sont supposés être indépendamment échantillonnés dans le même domaine $\mathcal{D}$.** Cette exigence est communément appelée dans la littérature sur l'apprentissage automatique comme l'hypothèse "indépendante et identiquement distribuée" (i.i.d.).

Le risque empirique d'une hypothèse h calculée sur un échantillon X, $\hat{R}_X[h_X]$, est ainsi défini comme:

$$\hat{R}_X[h_X] = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}[h_X(x_i), f(x_i)]. \tag{8.3}$$

L'erreur de généralisation de h_X (ou écart de généralisation) sur le domaine $\mathcal{D}$ est définie comme la différence entre les risques réels et empiriques:

$$\varepsilon = |R[h_X] - \hat{R}_X[h_X]|. \tag{8.4}$$

Idéalement, $\hat{R}_X[h_X] \approx 0$ et $\varepsilon \approx 0$, auquel cas h_X est capable d'atteindre un faible risque sur des échantillons de $\mathcal{D}$ qui ont été pas observé au moment de l'entraînement.

8.2.3 Adaptation de domaine

Nous considérons maintenant des scénarios où l'i.i.d. l'hypothèse ne tient pas et les exemples ne devraient pas être distribués de manière identique. Nous introduisons le paramètre d'adaptation de domaine où les échantillons du *domaine source* $\langle \mathcal{D}_S, f_S \rangle$ sont considérés au moment de l'entraînement, mais au moment du test, on s'attend à ce que les échantillons soient tirés d'un *domaine cible* $\langle \mathcal{D}_T, f_T \rangle$. La discussion dans [96] a établi les fondements théoriques pour l'étude des propriétés de généralisation inter-domaines pour les problèmes d'adaptation de domaine, et nous présentons maintenant les résultats de la littérature sur l'adaptation de domaine qui sont pertinents pour cette thèse.

Une borne pour le risque d'une hypothèse h sur le domaine cible $R_T[h]$ a été introduite [17]. Ce résultat montre que $R_T[h]$ dépend de $R_S[h]$, du risque de h sur le domaine source, d'une notion de divergence entre les deux domaines, ainsi que du risque minimum qui peut être atteint par certains $h \in \mathcal{H}$ sur $\mathcal{D}_S$ et $\mathcal{D}_T$. Nous reformulons ce résultat dans le théorème 1. Une extension de ce résultat présenté dans le Théorème 1 a été introduite dans [98] afin de remplacer λ par un terme qui explique explicitement un éventuel décalage entre les fonctions d'étiquetage de source et domaines cibles. Pour ce faire, la divergence entre la source et la cible est calculée sur une classe d'hypothèses $\tilde{\mathcal{H}}$ définie comme $\tilde{\mathcal{H}} = \{sign(|h(x) - h'(x)| - t) | h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$. On énonce ce résultat dans le Théorème 2.

8.2.4 Formules Minimax

Dans cette thèse, nous avons considéré des applications distinctes (c. Dans ces cas, trouver des valeurs optimales pour ces groupes de paramètres θ et ϕ peut être formulé comme le problème d'optimisation suivant défini dans 2.13. Ce problème peut être résolu en trouvant un point-selle de $\mathcal{L}$ où il n'est pas possible d'améliorer simultanément sa valeur par rapport à θ et ϕ. Ce type de problème apparaît fréquemment dans les applications d'apprentissage automatique telles que les d'équité [100], la méthode critique d'acteur pour les d'apprentissage par renforcement [101], les d'optimisation robuste [102], la modélisation générative modeling [27] et l'apprentissage de représentations invariantes de domaine [28].

Dans cette thèse, nous apportons des contributions pour améliorer l'optimisation d'une variation du problème défini dans l'Éq. 2.13 et envisager son application à la fois à la modélisation générative et à l'apprentissage de représentations invariantes de domaine pour la généralisation de domaine. Dans le cas d'applications à la modélisation générative, nous considérons les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GANs) [27]. Ce cadre offre une nouvelle approche de la modélisation générative, utilisant des schémas d'apprentissage de la théorie des jeux pour apprendre implicitement une densité de probabilité. Les GANs sont généralement composés d'un modèle discriminateur D avec des paramètres θ, tels que $D(x) : \mathbb{R}^d \rightarrow [0,1]$, où d est la dimensionnalité de l'espace d'entrée, et un générateur G avec des paramètres ϕ, tel que $G(z) : \mathbb{R}^m \rightarrow \mathbb{R}^n$, où m est la taille d'un vecteur de bruit d'entrée z. $D(x)$ reçoit un échantillon de la distribution de données p_{data} ou un échantillon du générateur $G(z)$, $z \sim p_z$. Au cours de l'entraînement, son objectif est d'apprendre à distinguer ces deux types d'intrants différents. Le générateur, quant à lui, vise à *tromper* le discriminateur en apprenant à produire des échantillons aussi proches que possible de la distribution des données.

L'autre exemple de problèmes tels que 2.13 que nous avons pris en compte sont les réseaux de neurones conflictuels de domaine (*Domain Adversarial Neural Networks*, DANN). Les DANNs proposés dans [28, 18] comme approche contradictoire pour l'adaptation de domaine non supervisée, un cadre où un échantillon non étiqueté du domaine cible est disponible au moment de l'entraînement, via l'apprentissage de représentations invariantes de domaine. De la même manière que le cadre GAN, dans le cas des DANNs, le modèle d'encodeur (ou "générateur") vise à tromper le discriminateur afin d'apprendre des représentations qui sont *invariantes* à l'information de domaine, et, en pratique, cela est obtenu en mettant à jour ϕ dans une direction qui augmente la perte de discrimination de domaine $\mathcal{L}_D$. D'autre part, les paramètres du discriminateur de domaine θ sont mis à jour dans une direction qui diminue $\mathcal{L}_D$ afin que ce modèle puisse être un bon estimateur de la $\mathcal{H}$-divergence entre la source et les domaines cibles.

8.2.4.1 Optimisation multi-objectifs

Dans cette section, nous fournissons quelques définitions concernant l'optimisation multi-objectifs de la littérature antérieure qui seront utiles dans le Chapitre 5. La notation en caractères gras est utilisée pour désigner les fonctions à valeur vectorielle. Un problème d'optimisation multi-objectif est défini comme 2.20 [107].

8.2.4.2 Pareto-dominance

Soient x_1 et x_2 deux vecteurs de décision. x_1 domine x_2 (noté $x_1 \prec x_2$) si et seulement si $f_i(x_1) \leq f_i(x_2)$ pour tout $i \in \{1, 2, \ldots, K\}$ et $f_j(x_1) < f_j(x_2)$ pour certains $j \in \{1, 2, \ldots, K\}$. Dans le cas où un vecteur de décision x n'est dominé par aucun autre vecteur dans Ω, x est *une solution non dominée.*

8.2.4.3 Pareto-optimalité

Un vecteur de décision $x^* \in \Omega$ est dit Pareto-optimal si et seulement s'il n'y a pas de $x \in \Omega$ tel que $x \prec x^*$, c'est-à-dire que x^* est une solution non dominée. L'ensemble Pareto-optimal (EP) est défini comme l'ensemble de toutes les solutions Pareto-optimales $x \in \Omega$, c'est-à-dire PS $= \{x \in \Omega \mid \text{x is Pareto optimal}\}$. L'ensemble de tous les vecteurs objectifs $\mathbf{F}(x)$ tels que x est Pareto-optimal est appelé front de Pareto (PF), c'est-à-dire $PF = \{\mathbf{F}(x) \in \mathbb{R}^K \mid x \in PS\}$.

8.2.4.4 Pareto-stationnarité

La Pareto-stationnarité est une condition nécessaire à l'optimalité de Pareto. Pour f_k dérivable partout pour tout k, $\mathbf{F}$ est Pareto-stationnaire à x s'il existe un ensemble de scalaires $\alpha_k, k \in \{1, \ldots, K\}$, tel que 2.21.

8.2.4.5 Descente de gradient multiple

Descente de gradient multiple (MGD) [108, 109, 110, 111] a été proposée pour le cas sans contrainte d'optimisation multi-objectif de $\mathbf{F}(x)$ en supposant convexe, continuellement différentiable et lisse $f_k(x)$, $\forall k = \{1, \ldots, K\}$. Pour chaque itération, MGD vise à trouver une direction qui diminuera simultanément au maximum tous les objectifs. Dans [108], il a été montré que cela peut être réalisé en trouvant une direction w^* telle que la dérivée directionnelle vers w^* est négative pour tous les objectifs, c'est-à-dire $w^* \cdot \nabla f_k(x) \leq 0$, $\forall k = \{1, \ldots, K\}$. Dans le cas de MGD, le problème de trouver une telle direction de descente commune est résolu en définissant l'enveloppe convexe de tous les $\nabla f_k(x)$, et en trouvant l'élément de norme minimum en son sein.

8.2.4.6 Maximisation de l'hypervolume (HV)

Soit S l'ensemble des solutions d'un problème d'optimisation multi-objectifs. L'hypervolume H de S est défini comme 2.24. $\mathcal{H}(S)$ peut être compris comme la taille de l'espace couvert par $\{\mathbf{F}(x) \mid x \in S\}$ [113].

L'hypervolume a été introduit à l'origine comme une métrique quantitative pour la couverture et la convergence des fronts Pareto-optimaux obtenus par des algorithmes basés sur la population [114]. Les méthodes basées sur la maximisation directe de $\mathcal{H}$ présentent une convergence favorable même dans des scénarios difficiles, tels que la minimisation simultanée de 50 objectifs [113]. Dans le contexte de l'apprentissage automatique, la HV à solution unique a été appliquée aux réseaux de neurones en tant que perte de substitution pour l'erreur quadratique moyenne [115], c'est-à-dire que la perte fournie par chaque exemple dans un lot d'entraînement est traitée comme un coût unique et le L'approche multi-objectifs vise à minimiser les coûts sur tous les exemples. Les auteurs montrent qu'une telle méthode fournit un entraînement de type boost peu coûteux.

8.2.5 Ensembles de données

Dans cette section, nous avons présenté les ensembles de données utilisés tout au long de cette thèse. Nous commençons par décrire les ensembles de données d'images, à savoir PACS et VLCS, que nous avons utilisés dans les tâches de reconnaissance d'objets impliquant des décalages de distribution dans les chapitres 4 et 6. Ensuite, nous présentons les ensembles de données WAUC et SEED, les ensembles de données EEG utilisés dans les chapitres 3 et 4. En plus des bases de données décrites dans cette section, nous avons également utilisé tout au long de nos expériences des ensembles de données standard dans la littérature sur l'apprentissage automatique, tels que MNIST [116], CIFAR-10 [117], ImageNet [118], CelebA [119], et Cats.

Il est important de souligner qu'en plus d'utiliser des ensembles de données déjà disponibles dans la littérature, nous avons également collecté au cours de ce travail une base de données multimodale d'enregistrements psychophysiologiques de 48 sujets. Dans cette thèse, nous avons utilisé les enregistrements EEG de cette base de données dans le Chapitre 3 et les décrivons dans la Section 2.6.3. Pour une description complète et une analyse de toutes les modalités collectées, le lecteur intéressé est renvoyé à [74].

8.2.6 Conclusion

Dans ce chapitre, nous avons présenté les principales définitions, paramètres, ensembles de données et résultats précédents utilisés dans les chapitres suivants de cette thèse. Nous avons introduit le cadre empirique de minimisation des risques, ainsi que les cadres d'adaptation de domaine et d'adaptation de domaine multi-sources auxquels il sera fait référence en particulier dans les chapitres 3, 4, 6. Ce chapitre contient également dans la section 2.4 une brève introduction à un problème d'optimisation standard en apprentissage automatique qui apparaît dans les chapitres 4 et 5 et relie deux des contributions majeures cette thèse pour le cadre de généralisation de domaine et la modélisation générative. Nous avons également introduit des définitions et des algorithmes de base dans le domaine de l'optimisation multi-objectifs qui sont des composants essentiels du contenu présenté dans les chapitres 4 et 5. Enfin, nous avons présenté les principaux ensembles de données utilisés tout au long de nos expériences, en mettant particulièrement l'accent sur l'ensemble de données WAUC qui est utilisé dans le Chapitre 3 et a été collecté par les auteurs de la thèse et a inspiré les contributions ultérieures à généralisation de la diffusion.

8.3 Chapitre 3: Caractérisation des changements de distribution sur les données EEG

8.3.1 Introduction

Des travaux antérieurs sur l'adaptation de domaine ont montré que différentes techniques reposent sur des hypothèses distinctes sur les distributions d'entraînement et de test [134, 98]. Par exemple, une exigence courante est l'hypothèse *covariate shift*, qui considère que les distributions des étiquettes y conditionnées sur les données x, $p(y|x)$, ne se déplacent pas entre les conditions d'entraînement et de test et seules les distributions marginales $p(x)$ se déplacent vers [134]. Dans le cas des applications des BCIs passives basées sur l'EEG, cependant, des travaux antérieurs ont soutenu que $p(y|x)$ est susceptible de changer entre différents sujets [135, 44, 43, 42]. Par conséquent, l'hypothèse de décalage de covariable ne peut pas être considérée comme acquise puisque, étant donné les vecteurs de caractéristiques x_1 et x_2 respectivement acquis à partir de deux sujets distincts et représentés dans un espace de caractéristiques partagé, $p_1(y|x_1) \neq p_2(y|x_2)$ même dans le cas où $x_1 = x_2$.

Vérifier si les hypothèses sous-jacentes d'une approche d'adaptation de domaine particulière tiennent dans la pratique est une étape fréquemment négligée par les approches d'adaptation de domaine [137]. Dans ce chapitre, nous affirmons qu'il est nécessaire d'évaluer la structure sous-jacente d'un ensemble de données particulier afin de vérifier quels types de changements de distribution existent et quelles hypothèses pourraient être considérées en toute sécurité (ou non), lors de l'utilisation de stratégies d'adaptation de domaine. À cette fin, nos principales contributions sont:

(i) Nous introduisons une méthode pour estimer l'inadéquation entre les sujets entre les distributions d'étiquettes conditionnelles; (ii) Nous appliquons une notion de divergence introduite dans [97] pour estimer l'inadéquation entre les distributions marginales de paires de sujets; (iii) Nous étudions si les pratiques courantes dans la littérature EEG pour atténuer la variabilité inter-sujets, telles que la normalisation des caractéristiques spectrales, sont capables d'atténuer les changements de distribution conditionnels et marginaux. Compte tenu de la pertinence d'atténuer la variabilité inter-sujets sur l'évaluation de la charge de travail mentale basée sur l'EEG, nous validons empiriquement notre méthode proposée sur l'ensemble de données WAUC, précédemment introduit dans le Chapitre 2 [74]. Nous avons examiné un sous-ensemble de cet ensemble de données composé de 18 sujets et étudié comment différentes manières de moduler l'activité physique affectent les changements statistiques inter-sujets sur les corrélats EEG de la charge de travail mentale.

8.3.2 Approches pratiques de l'estimation des changements de distribution

Dans cette section, nous proposons des stratégies pratiques pour estimer les décalages conditionnels et marginaux dans le cas où plusieurs domaines (sujets) sont disponibles. La quantification de cette inadéquation nous permettra de:

- Élucider sur les stratégies d'adaptation de domaine à utiliser pour un scénario donné en vérifiant si, par exemple, l'hypothèse de changement de covariable est vérifiée.
- Comme ces quantités sont liées à la performance d'une hypothèse particulière sur des sujets invisibles, nous pouvons utiliser leurs estimations calculées en tenant compte de différents espaces de caractéristiques et en déduire laquelle obtiendrait de meilleures performances sur des sujets invisibles.

Un décalage conditionnel est observé entre les sujets lorsque la fonction d'étiquetage (ou, dans le cas stochastique, la distribution conditionnelle des étiquettes compte tenu des caractéristiques d'entrée) diffère entre les sujets, c'est-à-dire que pour M sujets, nous avons $f_i(x) \neq f_j(x)$, $\forall i, j = \{1, \cdots, M\}$. Afin de caractériser le décalage conditionnel inter-sujets d'un ensemble de données de M sujets, nous considérons la quantité suivante sur la borne de généralisation présentée dans le corollaire 2 pour toutes les paires de sujets selon l'Eq. 3.3. En pratique, il n'est pas possible de calculer une telle quantité car on n'a pas accès aux vraies fonctions d'étiquetage et au calcul des espérances dans l'Eq. 3.3 est intraitable. Nous proposons donc d'estimer de telles valeurs en apprenant une règle d'étiquetage pour chacun des domaines, et de rendre compte de la qualité de la classification des exemples de l'autre domaine. En supposant que nous soyons capables d'apprendre un bon prédicteur pour les étiquettes de chaque domaine, une telle approche est capable de rendre compte de la proximité des véritables fonctions d'étiquetage des différents domaines.

Notre approche proposée pour estimer le décalage marginal entre les sujets à partir d'un groupe de M domaines (sujets) repose sur l'estimation des divergences de domaines par paires, c'est-à-dire que nous calculons $d_{\mathcal{H}}[\mathcal{D}_i, \mathcal{D}_j]$ $\forall i, j = \{1, \cdots, M\}$. Dans le cas de scénarios où les jeux de données EEG sont pris en compte, l'estimation des décalages marginaux inter-domaines consiste à obtenir des modèles capables de discriminer par paires des caractéristiques extraites des enregistrements de différents sujets. De la même manière que la stratégie proposée pour estimer les valeurs de décalage conditionnel inter-sujets, nous introduisons une matrice hermitienne H qui tient compte des décalages marginaux entre tous les sujets. Chaque entrée de H correspond au taux d'erreur moyen de la classification des sujets par paire. En pratique, nous utilisons une validation croisée 5 fois pour estimer les taux d'erreur. Une valeur agrégée du décalage marginal peut également être obtenue via la norme de Frobenius rééchelonnée de H.

8.3.3 Expérimentations et discussions

Dans nos expérimentation, nous visons à répondre aux questions principales suivantes: i) Est-ce que différents schémas de normalisation de caractéristiques produisent différentes valeurs de décalages distributionnels? ii) L'estimation des changements de distribution peut-elle indiquer à quel point il est difficile d'apprendre des BCIs qui se généralisent bien sur des sujets invisibles? iii) Pour un espace de fonction fixe, nos résultats sont-ils cohérents sur deux partitions du WAUC contenant des sujets dont les niveaux d'activité physique étaient modulés par le vélo ou le tapis roulant?

Notre pipeline de prétraitement et d'extraction de caractéristiques consistait à sous-échantillonner l'enregistrement EEG à 250 Hz, à le filtrer avec un filtre passe-bande de 0,5 à 45 Hz et à calculer les caractéristiques sur des périodes de 4 secondes avec 3 secondes de chevauchement entre les fenêtres consécutives. Considérant une session expérimentale de 10 minutes, après sous-échantillonnage et épochage des données, nous avons obtenu un total approximatif de 600 points par sujet×session. Comme la littérature a montré que l'augmentation de la charge de travail mentale induisait des changements dans les bandes alpha, bêta et thêta dans le cortex frontal [139, 140], nous avons considéré les caractéristiques de densité spectrale de puissance (*Power Spectral Density*, PSD) dans les bandes de fréquences standard.

Nous envisageons de normaliser les caractéristiques par rapport à la première période de référence (ligne de base-1) met en évidence les changements dans le PSD dus à des stimuli mentaux et physiques. À son tour, la normalisation par rapport aux statistiques d'enregistrement recueillies au cours de la deuxième ligne de base met en évidence des modifications provenant uniquement des changements de charge de travail mentale, car seule l'effort physique a été modulé au cours de cette étape. En plus des stratégies de normalisation susmentionnées, nous effectuons également des expériences avec des caractéristiques obtenues après blanchiment par sujet des données. Nous évaluons quantitativement l'impact de la normalisation sur les performances de la charge de travail mentale sous l'angle des changements conditionnels et marginaux, ainsi que des performances de classification inter-sujets.

À la lumière de nos résultats et de notre discussion, nous soulignons les observations que nous avons trouvées les plus pertinentes pour être prises en compte dans les recherches futures. Dans le cas où l'objectif est d'améliorer les performances hors distribution, les procédures de normalisation qui diminuent le décalage conditionnel inter-sujet global doivent être prioritaires car elles génèrent des écarts de généralisation plus petits. Pour concevoir des BCIs passifs dans le but de surveiller la charge de travail mentale sous activité physique, notre analyse a montré que la normalisation du score z constituait la meilleure stratégie pour normaliser les caractéristiques de densité spectrale de puissance EEG. De plus, de tels espaces de caractéristiques normalisés devraient être pris en compte dans le cas où un apprentissage de représentation basé sur l'adaptation de domaine est utilisé pour apprendre des classificateurs invariants de domaine. Notez qu'il y a une mise en garde qui doit également être prise en compte: les résultats présentés dans les tableaux 3.1 et 3.2 indiquent systématiquement (c'est-à-dire, à travers l'équipement pour moduler l'activité physique et les procédures de normalisation) que l'amélioration les performances hors distribution via la normalisation des caractéristiques entraînent une diminution de la précision du modèle calculée sur des données invisibles provenant des sujets de l'entraînement.

8.4 Chapitre 4: Généralisation à de nouveaux domaines via l'appariement de distribution

8.4.1 Introduction

Dans ce chapitre, nous apportons des contributions dans le sens de la conception de machines d'apprentissage plus générales. Nous prenons un peu plus loin de l'adaptation de domaine et considérons un cadre plus général qui est souvent appelé dans la littérature de la généralisation de domaine [22]. Dans ce cas, on suppose qu'un ensemble de distributions sur l'espace d'entrée est disponible au moment de l'apprentissage. Au moment du test, cependant, les deux distributions observées, ainsi que de nouveaux domaines non vus, peuvent apparaître, et un faible risque doit être obtenu quel que soit le domaine sous-jacent. Plus important encore, contrairement à l'adaptation de domaine dans laquelle l'objectif est de trouver une représentation qui aligne les distributions de données d'apprentissage avec un domaine cible spécifique, *les stratégies de généralisation de domaine visent à trouver un espace de représentation qui donne de bonnes performances sur de nouvelles distributions, inconnues au moment de l'apprentissage.* Les travaux récents sur la généralisation de domaine ont inclus l'utilisation de l'augmentation de données [150, 151] au moment de l'entraînement, le méta-apprentissage pour simuler le changement de domaine [152], ajoutant une tâche auto-supervisée pour encourager un encodeur à apprendre représentations robustes [153, 85], et apprentissage des représentations invariantes de domaine [154], entre autres approches.

Nos contributions dans ce chapitre se situent dans le cadre de la généralisation du domaine et peuvent être divisées en trois catégories principales: théorique, algorithmique et applicative. Nous argumentons et prouvons d'abord que, étant donné un ensemble de distributions sur des données, si les distances mesurées entre une paire de telles distributions sont petites, la distance entre les mélanges obtenus à partir du même ensemble de distributions l'est également. Cela conduit au développement d'une borne sur le risque mesuré par rapport à n'importe quelle distribution, et montre en outre qu'une généralisation peut être attendue si l'on considère les distributions au voisinage de l'"enveloppe convexe"[5] défini par l'ensemble des domaines accessibles lors de l'apprentissage. Inspirés par ces résultats, nous introduisons une approche pour qu'un encodeur soit appliqué pour mapper les données de l'espace d'entrée à un espace où les indices dépendant du domaine sont filtrés tandis que les informations pertinentes pour la tâche d'intérêt sont préservées. Ce faisant, contrairement aux approches d'adaptation de domaine standard, aucune donnée provenant des distributions de test n'est considérée comme étant observée. Nous évaluons l'algorithme proposé dans des problèmes où différents ensembles d'hypothèses sont susceptibles de tenir, montrant sa polyvalence ainsi que ses performances améliorées par rapport à plusieurs lignes de base.

8.4.2 Apprentissage des représentations agnostiques de domaine pour la généralisation de domaine

Nous introduisons des hypothèses sur le processus de génération de données adaptées au contexte de généralisation du domaine, qui, selon nous, sont plus générales que l'i.i.d. standard. exigences et plus susceptibles de tenir dans la pratique. En d'autres termes, étant donné un échantillon de données, il est plus probable que nos hypothèses se maintiennent par rapport à l'i.i.d. plus restrictif. Nous prouvons ensuite dans le Lemme 1 qu'il est possible de borner la $\mathcal{H}$-divergence entre les domaines dans l'enveloppe convexe des sources et utilisons ce résultat pour prouver une généralisation bornée pour le risque sur des domaines invisibles dans le Théorème 4. Ce théorème montre qu'une généralisation peut être attendue pour des domaines au voisinage d'une notion d'enveloppe convexe des distributions observées au moment de l'apprentissage.

Motivés par les résultats précédents, nous proposons de concevoir des algorithmes qui minimisent les termes de la borne dans (4.7) qui peuvent être estimés même si seules les données sources sont observées, c'est-à-dire ϵ ainsi que la risques sur l'échantillon de train. Nous visons donc à apprendre un encodeur $E : \mathcal{X} \to \mathcal{Z}$, où $\mathcal{Z} \subset \mathbb{R}^d$ préserve les informations pertinentes pour séparer les classes tout en supprimant le domaine -des indices spécifiques de telle sorte qu'il est plus difficile de distinguer les exemples de différents domaines par rapport à l'espace d'origine $\mathcal{X}$.

L'approche proposée (nommé G2DM, *Generalizing to unseen Domains via Distribution Matching*) contient trois modules principaux, tous paramétrés par des réseaux de neurones: un encodeur E avec des paramètres ϕ, un classifieur de tâches C avec des paramètres θ_C, et un ensemble de $\mathcal{H}$ -estimateurs de divergence D_k avec les paramètres θ_k, $k \in [N_S]$. Intuitivement, E tente de minimiser une perte de classification $\mathcal{L}_C(\cdot;\theta_C)$ (entropie croisée standard dans notre cas) et des

[5] c'est-à-dire l'ensemble de tous les mélanges obtenus à partir de distributions données.

$\mathcal{H}$-divergences empiriques, ce qui est réalisé par la maximisation des pertes de discrimination de domaine, dénommées $\mathcal{L}_k$. Chaque discriminateur de domaine, quant à lui, vise à minimiser $\mathcal{L}_k$. La procédure pour estimer ϕ, θ_T et tous les θ_k peut être ainsi formulée comme le jeu minimax multijoueur dans 4.13. Nous proposons en outre d'augmenter l'approche accusatoire décrite en utilisant des stratégies initialement utilisées pour stabiliser l'entraînement de réseaux génératifs avec plusieurs discriminateurs [165, 77]. À savoir, nous incluons une couche de projection aléatoire dans l'entrée de chaque discriminateur de domaine dans le but de rendre les exemples de différentes distributions plus difficiles à distinguer. De plus, nous utilisons l'hypervolume log négatif au lieu de la sommation dans le jeu représenté dans 4.13 afin d'attribuer plus de préférence aux solutions qui diminuent uniformément toutes les divergences par paires, même dans les cas où il y a un échange. dans leur minimisation.

8.4.3 Expérimentations et discussions

Nous concevons notre évaluation empirique pour valider G2DM dans des conditions où différentes hypothèses sont satisfaites. Dans le premier scénario, nous avons choisi des conditions expérimentales telles que l'hypothèse de changement de covariable soit vérifiée. Pour cela, nous employons G2DM sur des tâches de reconnaissance d'objets. Dans ce cas, nous visons à répondre aux questions de recherche suivantes: i) G2DM peut-il être plus performant que l'ERM standard sous i.i.d. hypothèses en utilisant uniquement les informations des domaines sources? ii) Où se situent les performances de G2DM par rapport aux stratégies de généralisation de domaine proposées précédemment? iii) G2DM applique-t-il effectivement la correspondance de distribution entre les domaines source et invisible? Et iv) Quel est l'effet sur les performances résultantes données par les différents modèles d'accès pour tester les distributions pendant l'entraînement?

Nous évaluons ensuite si G2DM est capable d'atteindre de bonnes performances hors domaine même dans le scénario difficile où l'hypothèse de changement de covariable est susceptible d'être violée. Pour cela, nous considérons une tâche du monde réel qui consiste à classer les séries chronologiques EEG pour la prédiction de l'état affectif, un domaine en plein essor dans le domaine des systèmes homme-machine. Dans les applications impliquant des données EEG, les sujets sont souvent considérés comme des domaines distincts avec des fonctions d'étiquetage différentes, comme cela est également discuté et montré au Chapitre 3 [72].

Les résultats montrent que G2DM surpasse l'ERM en termes de performances moyennes dans les domaines invisibles pour les deux benchmarks, et soutiennent l'affirmation selon laquelle l'exploitation des informations du domaine source comme le fait G2DM fournit une amélioration de la généralisation aux distributions invisibles par rapport à la simple prise en compte de l'i.i.d. l'exigence est satisfaite. Nous observons également que G2DM était capable de mieux correspondre à la plupart des distributions sources, produisant ainsi des ϵ plus petits qui favorisent la généralisation comme prédit par le Théorème 4. Notamment, nous soulignons que bien que notre approche proposée n'ait pas accès aux données du domaine invisible au moment de l'apprentissage et, par conséquent, ne met pas directement en œuvre une stratégie pour diminuer la divergence entre le domaine invisible et l'enveloppe convexe des sources. Enfin, les expériences sur les BCIs pour la prédiction de l'état affectif ont démontré que les performances de G2DM sont équivalentes, voire meilleures, que certaines des stratégies d'adaptation de domaine considérées.

8.5 Chapitre 5: Entraînement multi-objectif des Réseaux Antagonistes Génératifs avec plusieurs discriminateurs

8.5.1 Introduction

Les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GANs) [27] offrent une nouvelle approche de la modélisation générative, utilisant des schémas d'entraînement théoriques des jeux pour apprendre implicitement une densité de probabilité donnée. Avant l'émergence des architectures GANs, la modélisation générative réaliste restait insaisissable. Tout en offrant un réalisme sans précédent, l'entraînement des GANs reste toujours lourde de problèmes de stabilité. Les défauts couramment rapportés impliquent le manque de signal de gradient utile fourni par le discriminateur et l'effondrement de mode, c'est-à-dire le manque de diversité dans les échantillons du générateur. Dans ce chapitre, nous appliquons une architecture similaire à celle introduite au Chapitre 4 pour l'apprentissage des représentations invariantes pour le problème d'apprentissage des GANs. Comme indiqué précédemment dans le Chapitre 2, les deux problèmes peuvent être formulés comme une optimisation minimax, et nous tirons parti de ce fait pour montrer que la même approche peut être utilisée à la fois pour la modélisation générative (dans ce chapitre) et l'apprentissage des représentations (Chapitre 4). De plus, dans ce chapitre, nous proposons d'utiliser l'optimisation multi-objectifs pour résoudre de tels problèmes et d'introduire un algorithme qui est une meilleure alternative par rapport aux approches multi-objectifs proposées précédemment.

Plus précisément, nous nous appuyons sur le cadre de [165] et proposons de reformuler leur minimisation de perte moyenne pour stabiliser l'entraînement des GANs en traitant le signal de perte fourni par chaque discriminateur comme une fonction objective indépendante. Pour y parvenir, nous minimisons simultanément les pertes en utilisant des techniques d'optimisation multi-objectifs. À savoir, nous exploitons des méthodes bien connues dans la littérature d'optimisation telles que l'algorithme de descente de gradient multiple (*Multiple Gradient Descent*, MGD) [109], précédemment introduit dans le Chapitre 2. Cependant, en raison du coût prohibitif de MGD dans le cas de grands réseaux de neurones, nous proposons d'utiliser des alternatives plus efficaces telles que la maximisation de l'hypervolume (cf. Chapitre 2) dans la région définie entre une borne supérieure fixe et partagée sur les pertes, que nous appelons le *point nadir* $\boldsymbol{\eta}^*$, et chacune des pertes composantes. Contrairement à l'approche décrite dans [165], où la perte moyenne est minimisée lors de l'entraînement du générateur, la maximisation de l'hypervolume optimise une perte pondérée, et l'entraînement du générateur attribuera une plus grande importance au retour des discriminateurs contre lesquels il effectue pauvrement.

8.5.2 Formation multi-objectifs des GANs avec plusieurs discriminateurs

Nous introduisons une variante du jeu GAN dans laquelle le générateur résout le problème multi-objectif suivant dans 5.4. L'entraînement se déroule de la manière habituelle [27], c'est-à-dire avec des mises à jour alternées entre les discriminateurs et le générateur. Des mises à jour de chaque discriminateur sont effectuées pour minimiser la perte décrite dans l'Eq. 5.2. Un choix naturel pour les mises à jour de notre générateur est l'algorithme MGD, décrit au Chapitre 2. Cependant, le calcul de la direction de la descente la plus raide w^* avant chaque étape de mise à jour des paramètres, comme requis dans MGD, peut être prohibitif pour les grands réseaux de neurones. Par conséquent, nous proposons un schéma alternatif pour l'optimisation multi-objectifs et affirmons que notre proposition et les méthodes précédemment publiées peuvent toutes être considérées comme exécutant une version de calcul plus efficace de la règle de mise à jour MGD, sans avoir à résoudre un programme quadratique, c'est-à-dire en calculant w^*, à chaque itération.

Des travaux antérieurs [112] ont montré que trouver l'ensemble des solutions candidates pour un problème d'optimisation multi-objectif qui maximise l'hypervolume H (c.f Eq. 2.24) donne des solutions Pareto-optimales. Puisque MGD converge vers un ensemble de points Pareto-stationnaires, c'est-à-dire un sur-ensemble des solutions Pareto-optimales, la maximisation de l'hypervolume produit un sous-ensemble des solutions obtenues à l'aide de MGD. Nous exploitons cette propriété et définissons la perte du générateur comme le log-hypervolume négatif, tel que défini dans l'équation. 5.5. Dans la figure 5.1, nous fournissons un exemple illustratif pour le cas où $K = 2$. La région en surbrillance correspond à $e^{\mathcal{V}}$. Puisque le point nadir $\boldsymbol{\eta}^*$ est fixe, $\mathcal{V}$ sera maximisé, et par conséquent $\mathcal{L}_G$ minimisé, si et seulement si chaque l_k est minimisé . On observe que le gradient de $\mathcal{L}_G$ par rapport à ses paramètres peut être obtenu en calculant une somme pondérée des gradients des pertes fournies par chaque discriminateur, dont les poids sont définis comme l'inverse de la distance au point nadir Composants. Cette formulation attribuera naturellement plus d'importance à des pertes plus élevées dans le gradient final, ce qui est une autre propriété utile de la maximisation de l'hypervolume.

8.5.3 Expérimentations et discussions

Nous avons réalisé quatre séries d'expériences visant à comprendre les phénomènes suivants: (i) Comment les méthodes alternatives d'entraînement des GANs avec discriminateurs multiples se comportent-elles par rapport au MGD; (ii) les performances des méthodes alternatives les unes par rapport aux autres en termes de qualité et de couverture des échantillons; (iii) Comment le nombre variable de discriminateurs affecte les performances compte tenu des méthodes étudiées; et (iv) si le réglage du discriminateur multiple est pratique étant donné le coût supplémentaire impliqué dans l'entraînement d'un ensemble de discriminateurs.

Les expériences réalisées sur MNIST montrent que HV présente un compromis utile entre *coût de calcul* et *qualité de l'échantillon* par rapport à des approches similaires telles que la minimisation de la perte moyenne (faible qualité et coût) et MGD (haute qualité et coût). Nos résultats indiquent que l'augmentation du nombre de discriminateurs augmente par conséquent la robustesse du générateur aux réglages d'hyperparamètres. De plus, les expériences réalisées sur le CIFAR-10 indiquent que la méthode décrite produit des échantillons de générateurs de meilleure qualité et plus diversifiés, tels que mesurés par plusieurs mesures quantitatives. De plus, la qualité de l'image et la diversité des échantillons s'améliorent une fois de plus à mesure que nous augmentons le nombre de discriminateurs.

Premièrement, nous avons exploité la dimensionnalité relativement faible du MNIST et l'avons utilisé comme banc d'essai pour comparer la MGD avec les autres approches, c'est-à-dire la minimisation de la perte moyenne, la perte moyenne pondérée et la HV, proposées dans ce travail. De plus, plusieurs initialisations et combinaisons *slack* ont été évaluées afin d'étudier comment la variation du nombre de discriminateurs affecte la robustesse à ces facteurs. Ensuite, des expériences ont été réalisées avec une version agrandie de CIFAR-10 à la résolution de 64x64 pixels tout en augmentant le nombre de discriminateurs. La mise à l'échelle a été effectuée dans le but de mener des expériences utilisant la même architecture décrite dans [165]. Nous avons évalué les performances de HV par rapport aux méthodes de référence en termes de qualité d'échantillon résultante. Des expériences supplémentaires ont été réalisées avec CIFAR-10 à sa résolution d'origine afin de fournir une comparaison claire avec les réglages bien connus du discriminateur unique. Nous avons analysé plus en détail l'impact de HV sur la diversité des échantillons générés à l'aide de l'ensemble de données MNIST empilé [184] et avons également comparé le coût de calcul et les performances pour les cas à discriminateur unique ou multiple. De plus, nous fournissons des échantillons générés pour illustrer l'impact du nombre de projections aléatoires sur la qualité du modèle génératif appris. Enfin, nous effectuons des expériences avec des ensembles de données à haute résolution, à savoir CelebA et Cats mis à l'échelle, et montrons que l'approche multi-objectifs proposée pour l'entraînement des GANs est également capable d'apprendre des distributions de grande dimension. Dans l'ensemble, nos expériences ont montré que la maximisation de l'hypervolume offre un meilleur compromis entre le coût de calcul et l'amélioration des performances par rapport aux approches à objectif unique et multi-objectifs pour l'entraînement des GANs.

8.6 Chapitre 6: Utiliser des projections aléatoires pour atténuer les changements marginaux entre les domaines

8.6.1 Introduction

Différentes formes de décalage de distribution affectent souvent les performances de prédiction du modèle dans les applications d'apprentissage automatique. Ces dernières années, de nouvelles techniques sont apparues pour permettre l'apprentissage sous des variations de données naturelles, dans des contextes tels que l'adaptation de domaine et la généralisation de domaine [193, 29, 194, 85, 195]. Simultanément, la vulnérabilité des réseaux de neurones aux perturbations artisanales a également attiré l'attention en raison de la menace qu'elle représente pour les applications critiques pour la sécurité. Ainsi, une myriade de techniques adaptées pour améliorer la robustesse contre des exemples hors distribution générés artificiellement a été proposée [196]. Bien que des travaux antérieurs aient proposé de tirer parti des avancées dans les approches d'adaptation de domaine pour améliorer la robustesse de l'adversaire [197, 198] et d'atténuer l'effet des changements de distribution en effectuant un certain type d'entraînement accusatoire [199, 151], seules quelques contributions [200] a tenté de concevoir des stratégies capables de gérer les deux types de changements de distribution.

Dans ce chapitre, nous suivons les contributions des Chapitres 4 et 5 et utilisons des projections aléatoires dans l'entrée des prédicteurs dans le but d'augmenter le chevauchement dans le support de différentes distributions. Plus précisément, nous proposons un cadre efficace et unifié pour traiter les changements de domaine à la fois naturels et artificiels: *Randomly Projecting Out Distribution Shifts* (RPODS). Motivés par le succès antérieur des projections aléatoires pour des applications telles que la modélisation générative [165, 77], l'augmentation des données [201, 202], entre autres [203, 204, 205], nous utilisons des transformations de données aléatoires comme un moyen d'appariement de la distribution, c'est-à-dire que nous mappons les échantillons d'entrée à un espace où le chevauchement entre les domaines est susceptible d'être plus élevé, diminuant ainsi la quantité d'informations spécifiques au domaine disponibles. En tant que contribution pratique supplémentaire, l'approche proposée augmente encore la robustesse d'un modèle aux attaques accusatoires en boîte blanche. C'est-à-dire que les couches de projection aléatoires sont ré-échantillonnées avant chaque prédiction. Ainsi, un sous-ensemble des paramètres du modèle est toujours inconnu de l'attaquant. Cela limite l'action des attaques qui reposent sur des connaissances antérieures sur le modèle, sans nuire à la précision d'origine, contrairement aux méthodes qui incluent des exemples contradictoires au moment de l'entraînement.

8.6.2 Utiliser des projections aléatoires pour atténuer les changements de distribution

De manière similaire aux chapitres 3 et 4, nous considérons des paramètres où un prédicteur $h : \mathcal{X} \to \mathcal{Y}$ doit présenter une bonne performance de généralisation sur différents domaines, y compris ceux qui ne sont pas disponibles au moment de l'entraînement. Comme dans le Chapitre 4, nous abordons le paramètre de généralisation de domaine [23, 22] et nous nous intéressons aux cas où plusieurs domaines sont disponibles et où l'hypothèse de décalage de covariable est vérifiée (c'est-à-dire que les distributions marginales diffèrent tandis que les données- les distributions conditionnelles d'étiquettes restent inchangées pour tous les domaines considérés). Nous supposons que les perturbations contradictoires induisent un changement dans la distribution marginale des données originales.

Le Théorème 5 montre que les projections aléatoires augmentent le chevauchement entre le support des distributions sur le même espace d'entrée et peuvent donc être utilisées pour atténuer les décalages de covariables. Plus précisément, dans le cas où deux domaines sur $\mathcal{X}$, $\mathcal{D}^1$ et $\mathcal{D}^2$, sont considérés, la projection W agit de telle manière que cela augmente probablement le chevauchement entre les deux domaines. Dans la section suivante, nous confirmons empiriquement cette observation en montrant que la $\mathcal{A}$-distance [97], un proxy de la $\mathcal{H}$-divergence, qui explique les discordances entre les distributions sur l'espace d'entrée, est en effet diminué lorsqu'il est estimé sur les entrées projetées, c'est-à-dire $d_{\mathcal{A}}(\mathcal{D}^1, \mathcal{D}^2) > d_{\mathcal{A}}(\mathcal{D}^1_W, \mathcal{D}^2_W)$.

Nous considérons les applications des réseaux de neurones et implémentons notre approche des décalages de distribution à projection aléatoire (*Randomly Projecting Out Distribution Shifts*, RPODS) en utilisant des couches convolutives. Plus précisément, nous considérons une banque de K projections telle que $\mathcal{X}_W \subset \mathbb{R}^m$ et chaque matrice de projection aléatoire $W_k \in R^{d\times m}$, $k = \{1, \ldots, K\}$, a des entrées tirées d'une distribution gaussienne. Dans toutes nos expériences, nous avons considéré $\mathcal{N}(0, \sigma^2)$, où σ est défini selon le schéma introduit dans [207]. Afin d'éviter que les projections ne soient considérablement déformées, nous projetons les paramètres des couches convolutives sur la boule unitaire L2. Un modèle est ensuite entraîné en considérant des exemples dans l'espace d'entrée projeté $\mathcal{X}_W$. La figure 6.1 illustre l'utilisation de RPODS et montre des exemples de l'ensemble de données PACS [120] dans l'espace projeté. Nous soulignons en outre que les RPODS induisent un avantage supplémentaire en termes d'amélioration de la robustesse d'un modèle aux attaques qui reposent sur la connaissance des paramètres du modèle (c'est-à-dire le modèle d'accès en boîte blanche). En ré-échantillonnant les matrices de projection à chaque itération, en plus d'avoir une entrée où les décalages de distribution sont réduits, une partie des paramètres du modèle change constamment et, par conséquent, l'attaquant n'aura jamais accès au modèle complet lors de la génération d'adversaires.

8.6.3 Expérimentations et discussions

Nous montrons empiriquement que, comme indiqué par le Théorème 5, l'utilisation de RPODS aide en fait à atténuer les changements de distribution et à évaluer la capacité de RPODS à améliorer la robustesse aux changements artificiels et naturels dans des scénarios pratiques. Dans le cas de déplacements naturels de domaine, nous considérons le cadre de généralisation de domaine sous un schéma *leave-one-domain-out*. Nous entraînons ainsi un modèle avec RPODS via une minimisation empirique des risques avec des exemples tirés des distributions d'entraînement tout en l'évaluant sur un domaine non vu. Enfin, nous montrons que les RPODS sont également capables d'améliorer la robustesse aux attaques adverses. Pour cela, nous entraînons un modèle sur l'ensemble de données CIFAR-10 et l'évaluons sur des perturbations contradictoires. Dans tous les cas, nous comparons les RPODS avec des méthodes adaptées pour atténuer les changements naturels ou artificiels. Nos résultats confirment empiriquement que les projections aléatoires filtrent les informations spécifiques au domaine en estimant la divergence entre les paires de domaines avant et après les projections. De plus, des expériences sur l'ensemble de données PACS ont montré que les RPODS améliorent un certain nombre d'approches adaptées au paramètre de généralisation de domaine, améliorant la précision moyenne sur les domaines invisibles de près de 6,8% par rapport à la ligne de base la plus performante. Nous avons également évalué RPODS dans un contexte où les changements de domaine étaient provoqués par des perturbations contradictoires et avons montré que, malgré sa simplicité, RPODS a considérablement amélioré la robustesse aux attaques en boîte blanche sur l'ensemble de données CIFAR-10 par rapport au modèle non défendu.

8.7 Chapitre 7: Conclusions

8.7.1 Généralisation inter-sujets sur les BCIs

Nous avons présenté au Chapitre 3 les premières étapes vers une meilleure compréhension des phénomènes de variabilité inter-sujets sur les BCIs passives basées sur l'EEG dans une perspective d'apprentissage statistique. Nous avons examiné ce problème à travers le prisme de l'adaptation de domaine et proposé des stratégies pour estimer les changements de distribution entre les distributions conditionnelles et marginales correspondant au processus de génération de données de caractéristiques et d'étiquettes de différents sujets. Pour évaluer l'approche proposée, l'ensemble de données WAUC, présenté au Chapitre 2, a été utilisé et une évaluation binaire de la charge de travail mentale à partir des caractéristiques spectrales de puissance EEG a été réalisée. Notre analyse a montré que la normalisation des caractéristiques, ainsi que les conditions de collecte de données telles que l'équipement utilisé pour induire la charge de travail physique, avaient un impact pertinent sur les valeurs estimées du changement conditionnel. Dans le cas où l'objectif est d'améliorer les performances hors distribution, les procédures de normalisation qui diminuent le décalage conditionnel inter-sujet global doivent être prioritaires car elles génèrent des écarts de généralisation plus petits. Notre analyse a montré que la normalisation du score z fournissait la meilleure stratégie pour normaliser les caractéristiques de densité spectrale de puissance EEG. De plus, de tels espaces de caractéristiques normalisés devraient être pris en compte dans le cas où des méthodes d'apprentissage de représentation basées sur l'adaptation de domaine sont utilisées pour apprendre des classificateurs invariants de domaine en plus des caractéristiques.

Dans le cas où des approches d'apprentissage de représentation de bout en bout sont utilisées sur les BCIs, nous montrons au Chapitre 4 que l'approche introduite, G2DM, peut également être envisagée pour obtenir une généralisation inter-sujet améliorée sans nécessiter aucune étape de calibrage. L'évaluation empirique de G2DM sur l'ensemble de données SEED pour la prédiction de l'état affectif a montré que l'approche proposée est capable d'exploiter avec succès les étiquettes de sujet disponibles (qui dans ce cas ne nécessite aucun effort supplémentaire pour la collecte de données) et présente une amélioration de plus que dans par rapport à l'ERM et, plus important encore, il surpasse les approches d'adaptation de domaine (qui ressemblent à un BCI calibré pour un sujet spécifique).

8.7.2 Généralisation de domaine

Dans le Chapitre 4, nous avons apporté des contributions au cadre de généralisation de domaine et montré que la généralisation hors distribution peut être réalisée au voisinage de l'ensemble des mélanges de distributions observés au moment de l'apprentissage. Nous montrons d'abord dans le lemme 1 que la $\mathcal{H}$-divergence entre n'importe quelle paire de domaines dans l'enveloppe convexe des distributions d'apprentissage peut être bornée par la $\mathcal{H}$-divergence maximale entre les domaines sources. Nous nous sommes basés sur ce résultat pour dériver une borne de généralisation dans le Théorème 4 pour le domaine *any* et montrer que le risque mesuré sur un tel domaine invisible dépend d'une somme convexe des risques sur les sources, le maximum La $\mathcal{H}$-divergence entre les sources, la $\mathcal{H}$-divergence entre le domaine invisible considéré et sa projection sur l'enveloppe convexe, et la discordance entre les fonctions d'étiquetage dans le cas où l'hypothèse de décalage de covariable ne tient pas. Dans le but de minimiser les termes de la borne introduite, nous concevons G2DM, une approche contradictoire afin que les divergences de domaine par paires soient estimées et minimisées. G2DM contient plusieurs innovations pratiques par rapport aux approches antagonistes précédentes telles que l'utilisation de couches de projection aléatoire avant les discriminateurs de domaine. Nous montrons que les représentations apprises par G2DM sont capables d'ignorer les informations spécifiques au domaine, ce qui indique que de telles représentations sont bien adaptées aux problèmes où des changements de distribution sont susceptibles d'être observés. Nous confirmons cette hypothèse sur plusieurs scénarios, tels que les tâches de reconnaissance d'objets sur les benchmarks PACS et VLCS, ainsi que la prédiction de l'état affectif basée sur l'EEG. Nos résultats ont montré que G2DM présentait une performance améliorée par rapport à plusieurs lignes de base, y compris les techniques d'adaptation de domaine qui ont accès à un échantillon du domaine invisible.

Nous contribuons en outre au paramètre de généralisation de domaine dans le Chapitre 6. Nous avons proposé RPODS, une stratégie polyvalente et efficace pour améliorer la robustesse des réseaux de neurones. Les RPODS ne reposent pas sur des étiquettes de domaine et exploitent des couches convolutives initialisées et ré-échantillonnées de manière aléatoire à chaque itération afin d'apprendre des représentations dans un espace où le chevauchement entre les distributions est plus important et empêcher l'accès complet du modèle à un attaquant potentiel. Nous avons fourni des preuves empiriques pour étayer l'affirmation selon laquelle de telles projections aléatoires sont capables d'atteindre cet objectif en montrant que l'utilisation de RPODS diminue systématiquement les divergences entre les domaines pour l'ensemble de données PACS par rapport aux données brutes. Cela indique que RPODS agit en supprimant les informations spécifiques au domaine telles que la texture et applique un modèle pour se concentrer sur des fonctionnalités de niveau supérieur telles que la forme, qui sont davantage liées aux étiquettes de classe dans le cas considéré. Les expériences sur l'ensemble de données PACS ont confirmé cette hypothèse en montrant que RPODS améliorait les performances sur la plupart des scénarios d'évaluation considérés. Malgré sa simplicité, RPODS a considérablement amélioré la robustesse aux attaques en boîte blanche sur l'ensemble de données CIFAR-10 par rapport au modèle non défendu et, plus important encore, il a atteint des performances compétitives par rapport aux approches spécifiquement conçues pour atténuer les effets des perturbations contradictoires.

8.7.3 Modélisation générative

Au Chapitre 5 nous avons proposé d'utiliser les mêmes apports de G2DM pour une application différente: la modélisation générative. Nous avons considéré l'apprentissage de GAN avec plusieurs discriminateurs et nous avons montré que l'apprentissage du générateur peut être vu comme un problème multi-objectif où chaque objectif correspond à la perte d'un discriminateur. Nous avons ainsi proposé l'utilisation de techniques d'optimisation multi-objectifs basées sur le gradient pour mettre à jour les paramètres du générateur et exploité des méthodes bien connues telles que l'algorithme de descente de gradient multiple. Cependant, en raison du coût prohibitif de MGD dans le cas de grands réseaux de neurones, nous proposons d'utiliser une alternative plus efficace, à savoir, l'algorithme de maximisation d'hypervolume qui optimise une perte pondérée de sorte que l'apprentissage du générateur attribuera une plus grande importance au retour des discriminateurs contre qu'il fonctionne mal. Il a été observé que l'approche proposée produisait systématiquement des échantillons de meilleure qualité en termes de FID par rapport à la perte moyenne et à d'autres règles d'agrégation pour les pertes. Nous avons en outre observé un nombre plus élevé de discriminateurs pour augmenter la diversité des échantillons et la robustesse du générateur aux valeurs des hyperparamètres. Il a également été démontré qu'une telle approche pour l'entraînement des GANs réussissait à générer des images haute résolution et à générer des performances compétitives en termes de métriques telles que FID et Inception Score par rapport à d'autres méthodes pour stabiliser l'entraînement telles que WGAN-GP et SNGAN.

8.7.4 Optimisation multi-objectif

En termes d'optimisation multi-objectifs, nos contributions du Chapitre 5 ont révélé que l'utilisation d'approches basées sur les gradients pour résoudre des problèmes multi-objectifs doit être envisagée lorsque plusieurs fonctions de perte conflictuelles doivent être minimisées. Nous avons montré que malgré le coût de calcul accru et les défis tels que les estimations de gradient bruyantes et la non-convexité des objectifs, l'algorithme de descente à gradient multiple est capable de directions de descente qui donnent de meilleures performances que les stratégies naïves telles que la scalarisation linéaire des pertes. Nous avons également proposé un moyen d'utiliser l'approche de maximisation d'hypervolume plus efficace pour former les GANs et avons montré qu'avec une bonne adaptation du point nadir, cet algorithme consiste en une meilleure stratégie pour trouver des solutions dans la région centrale du front de Pareto, qui, en pratique, induit une meilleure convergence pour tous les discriminateurs, puisque toutes les fonctions de perte se verront attribuer une importance globalement égale. Enfin, nous avons montré que les approches précédentes proposées pour former des GANs avec plusieurs discriminateurs peuvent être considérées comme des méthodes d'optimisation multi-objectifs basées sur le gradient avec différentes stratégies pour trouver une direction de descente pour les objectifs.

Bibliography

[1] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.-F. Li, W.-W. Tu, Q. Yang, and Y. Yu, "Taking human out of learning applications: A survey on automated machine learning," *arXiv preprint arXiv:1810.13306*, 2018.

[2] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[3] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[4] D. Zhang and J. J. Tsai, *Advances in machine learning applications in software engineering.* Igi Global, 2006.

[5] K. Crawford, *The Atlas of AI.* Yale University Press, 2021.

[6] A. Birhane and F. Cummins, "Algorithmic injustices: Towards a relational ethics," *arXiv preprint arXiv:1912.07376*, 2019.

[7] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1537–1547.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning.* MIT press, 2016.

[10] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[11] D. Bourilkov, "Machine and deep learning applications in particle physics," *International Journal of Modern Physics A*, vol. 34, no. 35, p. 1930019, 2019.

[12] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon *et al.*, "Highly accurate protein structure prediction for the human proteome," *Nature*, vol. 596, no. 7873, pp. 590–596, 2021.

[13] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge *et al.*, "Skillful precipitation nowcasting using deep generative models of radar," *arXiv preprint arXiv:2104.00954*, 2021.

[14] V. Vapnik, "Principles of risk minimization for learning theory," in *Advances in neural information processing systems*, 1992, pp. 831–838.

[15] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[16] J. Langford, “Tutorial on practical prediction theory for classification,” *Journal of machine learning research*, vol. 6, no. Mar, pp. 273–306, 2005.

[17] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in *Advances in neural information processing systems*, 2007, pp. 137–144.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] A. Jaech, L. Heck, and M. Ostendorf, “Domain adaptation of recurrent neural networks for natural language understanding,” *arXiv preprint arXiv:1604.00117*, 2016.

[20] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *ICML*, 2011.

[21] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó, “One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks,” *NeuroImage: Clinical*, vol. 21, p. 101638, 2019.

[22] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” in *International Conference on Machine Learning*, 2013, pp. 10–18.

[23] G. Blanchard, G. Lee, and C. Scott, “Generalizing from several related classification tasks to a new unlabeled sample,” *Advances in neural information processing systems*, vol. 24, pp. 2178–2186, 2011.

[24] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *International Conference on Learning Representations (Workshop track)*, 2017.

[25] A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim, “Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction,” in *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2019, pp. 132–137.

[26] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, “Perceiver io: A general architecture for structured inputs & outputs,” *arXiv preprint arXiv:2107.14795*, 2021.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[28] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[29] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, “Generalizing to unseen domains via distribution matching,” *arXiv preprint arXiv:1911.00804*, 2019.

[30] A. Singh, “Adversarial incremental learning,” *arXiv preprint arXiv:2001.11152*, 2020.

[31] C. Guger, W. Harkam, C. Hertnaes, and G. Pfurtscheller, “Prosthetic control by an EEG-based brain-computer interface (BCI),” in *Proc. aaate 5th european conference for the advancement of assistive technology*. Citeseer, 1999, pp. 3–6.

[32] G. Santhanam, S. I. Ryu, M. Y. Byron, A. Afshar, and K. V. Shenoy, “A high-performance brain–computer interface,” *nature*, vol. 442, no. 7099, pp. 195–198, 2006.

[33] M. de Castro-Cros, M. Sebastian-Romagosa, J. Rodríguez-Serrano, E. Opisso, M. Ochoa, R. Ortner, C. Guger, and D. Tost, “Effects of gamification in bci functional rehabilitation,” *Frontiers in neuroscience*, vol. 14, p. 882, 2020.

[34] L. Reinerman-Jones, G. Matthews, D. Barber, and J. Abich, "Psychophysiological metrics for workload are demand-sensitive but multifactorial," in *Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1. Sage Publications Sage CA: Los Angeles, CA, 2014, pp. 974–978.

[35] G. Matthews, L. Reinerman-Jones, D. Barber, and J. Abich, "The psychometrics of mental workload: multiple measures are sensitive but divergent," *Human Factors*, vol. 57, no. 1, pp. 125–143, 2015.

[36] A. Kok, "Event-related-potential (ERP) reflections of mental resources: a review and synthesis," *Biological psychology*, vol. 45, no. 1-3, pp. 19–56, 1997.

[37] P. Nikolov, "The effect of concurrent cognitive-visuomotor multitasking and task difficulty on dynamic functional connectivity in the brain," Master's thesis, Virginia Commonwealth University, 2013.

[38] A. Seemüller, E. Müller, and F. Rösler, "EEG-power and-coherence changes in a unimodal and a crossmodal working memory task with visual and kinesthetic stimuli," *International Journal of Psychophysiology*, vol. 83, no. 1, pp. 87–95, 2012.

[39] A. Drouin-Picaro, I. Albuquerque, J.-F. Gagnon, D. Lafond, and T. H. Falk, "EEG coupling features: Towards mental workload measurement based on wearables," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on.* IEEE, 2017, pp. 28–33.

[40] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[41] S. Sanei and J. A. Chambers, *EEG signal processing.* John Wiley & Sons, 2013.

[42] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, C.-T. Lin, and T.-P. Jung, "A subject-transfer framework for obviating inter-and intra-subject variability in EEG-based drowsiness detection," *NeuroImage*, vol. 174, pp. 407–419, 2018.

[43] D. Wu, C.-H. Chuang, and C.-T. Lin, "Online driver's drowsiness estimation using domain adaptation with model fusion," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 2015, pp. 904–910.

[44] D. Wu, V. J. Lawhern, and B. J. Lance, "Reducing bci calibration effort in rsvp tasks using online weighted adaptation regularization with source domain selection," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 2015, pp. 567–573.

[45] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, 2015.

[46] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[47] P. Aricò, G. Borghini, G. Di Flumeri, N. Sciaraffa, and F. Babiloni, "Passive BCI beyond the lab: current trends and future directions," *Physiological measurement*, vol. 39, no. 8, p. 08TR02, 2018.

[48] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications.* Springer, 2017, pp. 153–171.

[49] H. Daume III and D. Marcu, "Domain adaptation for statistical classifiers," *Journal of artificial Intelligence research*, vol. 26, pp. 101–126, 2006.

[50] L.-M. Zhao, X. Yan, and B.-L. Lu, “Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition,” in *Proceedings of the 35th AAAI Conference on Artificial Intelligence.* sn, 2021.

[51] H. Chen, M. Jin, Z. Li, C. Fan, J. Li, and H. He, “MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition,” *arXiv preprint arXiv:2107.07740*, 2021.

[52] I. Albuquerque, A. Tiwari, J.-F. Gagnon, D. Lafond, M. Parent, S. Tremblay, and T. Falk, “On the analysis of EEG features for mental workload assessment during physical activity,” in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 2018, pp. 538–543.

[53] J. Zhang, Z. Yin, and R. Wang, “Pattern classification of instantaneous cognitive task-load through gmm clustering, laplacian eigenmap, and ensemble svms,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 4, pp. 947–965, 2016.

[54] M. A. Almogbel, A. H. Dang, and W. Kameyama, “EEG-signals based cognitive workload detection of vehicle driver using deep learning,” in *2018 20th International Conference on Advanced Communication Technology (ICACT).* IEEE, 2018, pp. 256–259.

[55] Z. Yin and J. Zhang, “Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights,” *Neurocomputing*, vol. 260, pp. 349–366, 2017.

[56] M. Stikic, R. R. Johnson, V. Tan, and C. Berka, “EEG-based classification of positive and negative affective states,” *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 99–112, 2014.

[57] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, “Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos,” in *International Conference on Brain Informatics.* Springer, 2010, pp. 89–100.

[58] Y. Liu, O. Sourina, and M. K. Nguyen, “Real-time EEG-based emotion recognition and its applications,” in *Transactions on computational science XII.* Springer, 2011, pp. 256–277.

[59] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza, “Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors,” *Scientific reports*, vol. 8, no. 1, pp. 1–15, 2018.

[60] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, “Personalized emotion recognition by personality-aware high-order learning of physiological signals,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–18, 2019.

[61] D. Wu, X. Han, Z. Yang, and R. Wang, “Exploiting transfer learning for emotion recognition under cloud-edge-client collaborations,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 479–490, 2020.

[62] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, “Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition,” *arXiv preprint arXiv:2109.09559*, 2021.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[64] T. Postadjian, A. Le Bris, H. Sahbi, and C. Mallet, “Investigating the potential of deep neural networks for large-scale classification of very high resolution satellite images.” *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, 2017.

[65] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, “Imagenet training in minutes,” in *Proceedings of the 47th International Conference on Parallel Processing*, 2018, pp. 1–10.

[66] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” *IEEE Transactions on geoscience and remote sensing*, vol. 55, no. 2, pp. 645–657, 2016.

[67] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *International conference on machine learning.* PMLR, 2014, pp. 647–655.

[68] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in *European conference on computer vision.* Springer, 2016, pp. 443–450.

[69] L. A. Gatys, A. S. Ecker, and M. Bethge, “Texture and art with deep neural networks,” *Current opinion in neurobiology*, vol. 46, pp. 178–186, 2017.

[70] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness,” *arXiv preprint arXiv:1811.12231*, 2018.

[71] H. Wang, Z. He, Z. C. Lipton, and E. P. Xing, “Learning robust representations by projecting superficial statistics out,” *arXiv preprint arXiv:1903.06256*, 2019.

[72] I. Albuquerque, J. Monteiro, O. Rosanne, A. Tiwari, J.-F. Gagnon, and T. H. Falk, “Cross-subject statistical shift estimation for generalized electroencephalography-based mental workload assessment,” in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 3647–3653.

[73] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, “Domain generalization via model-agnostic learning of semantic features,” in *Advances in Neural Information Processing Systems*, 2019, pp. 6447–6458.

[74] I. Albuquerque, A. Tiwari, M. Parent, R. Cassani, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, “Wauc: a multi-modal database for mental workload assessment under physical activity,” *Frontiers in Neuroscience*, vol. 14, 2020.

[75] I. Albuquerque, J. Monteiro, O. Rosanne, and T. H. Falk, “Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment,” *Under Review at the IEEE Transactions on Human-Machine Systems*, 2021.

[76] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, “Generalizing to unseen domains via distribution matching,” *Under Review at the IEEE Transactions on Systems, Man, and, Cybernetics: Systems*, 2021.

[77] I. Albuquerque, J. Monteiro, T. Doan, B. Considine, T. Falk, and I. Mitliagkas, “Multi-objective training of generative adversarial networks with multiple discriminators,” in *International Conference on Machine Learning*, 2019, pp. 202–211.

[78] I. Albuquerque, J. a. Monteiro, and T. Falk, “Randomly projecting out distribution shifts for improved robustness,” *Workshop on Distribution Shifts: Connecting Methods and Applications at the Conference on Neural Information Processing Systems*, 2021.

[79] X. Liu, J. Monteiro, I. Albuquerque, Y. Lai, C. Jiang, S. Zhang, T. H. Falk, and J. Liang, “Single-shot real-time compressed ultrahigh-speed imaging enabled by a snapshot-to-video autoencoder,” *Photonics Research*, vol. 9, no. 12, pp. 2464–2474, 2021.

[80] O. Rosanne, I. Albuquerque, R. Cassani, J.-F. Gagnon, S. Tremblay, and T. H. Falk, “Adaptive filtering for improved EEG-based mental workload assessment of ambulant users,” *Frontiers in Neuroscience*, vol. 15, p. 341, 2021.

[81] M. Parent, I. Albuquerque, A. Tiwari, R. Cassani, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "Pass: a multimodal database of physical activity and stress for mobile passive body/brain-computer interface research," *Frontiers in Neuroscience*, vol. 14, p. 1274, 2020.

[82] A. Tiwari, I. Albuquerque, M. Parent, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H Falk, "Multi-scale heart beat entropy measures for mental workload assessment of ambulant users," *Entropy*, vol. 21, no. 8, p. 783, 2019.

[83] I. Albuquerque, J. Monteiro, T. H. Falk, V. Pavlovic, F. Ephrem, and D. Lucaci, "Multimodal assessment of human innovation perception based on eye tracking, electroencephalography and electrocardiography," in *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE).* IEEE, 2018, pp. 1–4.

[84] I. Albuquerque, J. Monteiro, and T. H. Falk, "Learning to navigate image manifolds induced by generative adversarial networks for unsupervised video generation," *arXiv preprint arXiv:1901.11384*, 2019.

[85] I. Albuquerque, N. Naik, J. Li, N. Keskar, and R. Socher, "Improving out-of-distribution generalization via multi-task self-supervised pretraining," *arXiv preprint arXiv:2003.13525*, 2020.

[86] I. Albuquerque, O. Rosanne, J.-F. Gagnon, S. Tremblay, and T. H. Falk, "Fusion of spectral and spectro-temporal EEG features for mental workload assessment under different levels of physical activity," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER).* IEEE, 2019, pp. 311–314.

[87] A. Tiwari, I. Albuquerque, M. Parent, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "A comparison of two ECG inter-beat interval measurement methods for HRV-based mentalworkload prediction of ambulant users," *CMBES Proceedings*, vol. 42, 2019.

[88] A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, M. Parent, S. Tremblay, and T. H. Falk, "Mental workload assessment during physical activity using non-linear movement artefact robust electroencephalography features," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 4149–4154.

[89] M. Parent, A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "A multimodal approach to improve the robustness of physiological stress prediction during physical activity," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 4131–4136.

[90] T. Doan, J. Monteiro, I. Albuquerque, B. Mazoure, A. Durand, J. Pineau, and R. D. Hjelm, "On-line adaptative curriculum learning for GANs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3470–3477.

[91] J. Monteiro, I. Albuquerque, Z. Akhtar, and T. H. Falk, "Generalizable adversarial examples detection based on bi-model decision mismatch," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 2839–2844.

[92] H. Banville, I. Albuquerque, A. Hyvärinen, G. Moffat, D.-A. Engemann, and A. Gramfort, "Self-supervised representation learning from electroencephalography signals," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP).* IEEE, 2019, pp. 1–6.

[93] J. Monteiro, I. Albuquerque, J. Alam, R. D. Hjelm, and T. Falk, "An end-to-end approach for the verification problem: learning the right distance," in *International Conference on Machine Learning.* PMLR, 2020, pp. 7022–7033.

[94] O. Rosanne, I. Albuquerque, J.-F. Gagnon, S. Tremblay, and T. H. Falk, "Performance comparison of automated EEG enhancement algorithms for mental workload assessment of ambulant users," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER).* IEEE, 2019, pp. 61–64.

[95] R. Cassani, I. Albuquerque, J. Monteiro, and T. H. Falk, "Ama: An open-source amplitude modulation analysis toolkit for signal processing applications," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP).* IEEE, 2019, pp. 1–4.

[96] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[97] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30.* VLDB Endowment, 2004, pp. 180–191.

[98] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *International Conference on Machine Learning*, 2019, pp. 7523–7532.

[99] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems*, 2018, pp. 8559–8570.

[100] A. Rezaei, R. Fathony, O. Memarrast, and B. Ziebart, "Fairness for robust log loss classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5511–5518.

[101] D. Pfau and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *arXiv preprint arXiv:1610.01945*, 2016.

[102] B. Chen, J. Wang, L. Wang, Y. He, and Z. Wang, "Robust optimization for transmission expansion planning: Minimax cost vs. minimax regret," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 3069–3077, 2014.

[103] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[104] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6629–6640.

[105] M. Fréchet, "Sur la distance de deux lois de probabilité," *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES*, vol. 244, no. 6, pp. 689–692, 1957.

[106] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[107] K. Deb, *Multi-objective optimization using evolutionary algorithms.* John Wiley & Sons, 2001, vol. 16.

[108] S. Schäffler, R. Schultz, and K. Weinzierl, "Stochastic method for the solution of unconstrained vector optimization problems," *Journal of Optimization Theory and Applications*, vol. 114, no. 1, pp. 209–222, 2002.

[109] J.-A. Désidéri, "Multiple-gradient descent algorithm (MGDA) for multiobjective optimization," *Comptes Rendus Mathematique*, vol. 350, no. 5-6, pp. 313–318, 2012.

[110] S. Peitz and M. Dellnitz, "Gradient-based multiobjective optimization with uncertainties," in *NEO 2016.* Springer, 2018, pp. 159–182.

[111] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 527–538.

[112] M. Fleischer, "The measure of pareto optima applications to multi-objective metaheuristics," in *International Conference on Evolutionary Multi-Criterion Optimization.* Springer, 2003, pp. 519–533.

[113] J. Bader and E. Zitzler, "HypE: An algorithm for fast hypervolume-based many-objective optimization," *Evolutionary computation*, vol. 19, no. 1, pp. 45–76, 2011.

[114] N. Beume, B. Naujoks, and M. Emmerich, "SMS-EMOA: Multiobjective selection based on dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.

[115] C. S. Miranda and F. J. Von Zuben, "Single-solution hypervolume maximization and its use for improving generalization of neural networks," *arXiv preprint arXiv:1602.01164*, 2016.

[116] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[117] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[118] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[119] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[120] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5542–5550.

[121] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1657–1664.

[122] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[123] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[124] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop.* IEEE, 2004, pp. 178–178.

[125] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* IEEE, 2010, pp. 129–136.

[126] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology.* Elsevier, 1988, vol. 52, pp. 139–183.

[127] G. A. Borg, "Psychophysical bases of perceived exertion," *Med sci sports exerc*, vol. 14, no. 5, pp. 377–381, 1982.

[128] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery II (MATB-II) software for human performance and workload research: A user's guide," 2011.

[129] G. Ruffini, S. Dunne, E. Farrés, Í. Cester, P. C. Watts, S. Ravi, P. Silva, C. Grau, L. Fuentemilla, J. Marco-Pallares *et al.*, "ENOBIO dry electrophysiology electrode; first human trial plus wireless electrode system," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 6689–6693.

[130] L. Shu, T. Xu, and X. Xu, "Multilayer sweat-absorbable textile electrode for EEG measurement in forehead site," *IEEE Sensors Journal*, vol. 19, no. 15, pp. 5995–6005, 2019.

[131] R. Cassani, H. Banville, and T. H. Falk, "Mules: An open source EEG acquisition and streaming server for quick and simple prototyping and recording," in *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion.* ACM, 2015, pp. 9–12.

[132] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[133] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[134] S. Ben-David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 129–136.

[135] D. Wu, "Online and offline domain adaptation for reducing bci calibration effort," *IEEE Transactions on human-machine Systems*, vol. 47, no. 4, pp. 550–563, 2016.

[136] F. D. Johansson, R. Ranganath, and D. Sontag, "Support and invertibility in domain-invariant representations," *arXiv preprint arXiv:1903.03448*, 2019.

[137] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Domain generalization via invariant representation under domain-class dependency," 2018.

[138] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *Journal of Machine Learning Research*, vol. 9, no. Aug, pp. 1757–1774, 2008.

[139] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.

[140] M. A. Hogervorst, A.-M. Brouwer, and J. B. van Erp, "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload," *Frontiers in neuroscience*, vol. 8, p. 322, 2014.

[141] S. Pati, E. Toth, and G. Chaitanya, "Quantitative EEG markers to prognosticate critically ill patients with covid-19: a retrospective cohort study," *Clinical Neurophysiology*, vol. 131, no. 8, p. 1824, 2020.

[142] J. Bogaarts, D. Hilkman, E. D. Gommer, V. van Kranen-Mastenbroek, and J. P. Reulen, "Improved epileptic seizure detection combining dynamic feature normalization with EEG novelty detection," *Medical & biological engineering & computing*, vol. 54, no. 12, pp. 1883–1892, 2016.

[143] Y. Bai, G. Huang, Y. Tu, A. Tan, Y. S. Hung, and Z. Zhang, "Normalization of pain-evoked neural responses using spontaneous EEG improves the performance of EEG-based cross-individual pain prediction," *Frontiers in computational neuroscience*, vol. 10, p. 31, 2016.

[144] H. A. Shedeed and M. F. Issa, "Brain-EEG signal classification based on data normalization for controlling a robotic arm," *Int. J. Tomogr. Simul*, vol. 29, no. 1, pp. 72–85, 2016.

[145] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[146] A. Cruz, G. Pires, A. Lopes, C. Carona, and U. J. Nunes, "A self-paced bci with a collaborative controller for highly reliable wheelchair driving: Experimental tests with physically disabled individuals," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 2, pp. 109–119, 2021.

[147] N. Sulaiman, M. N. Taib, S. A. M. Aris, N. H. A. Hamid, S. Lias, and Z. H. Murat, "Stress features identification from EEG signals using EEG asymmetry & spectral centroids techniques," in *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2010, pp. 417–421.

[148] R. Zhang, P. Xu, L. Guo, Y. Zhang, P. Li, and D. Yao, "Z-score linear discriminant analysis for eeg based brain-computer interfaces," *PloS one*, vol. 8, no. 9, p. e74433, 2013.

[149] S. Ladouce, D. I. Donaldson, P. A. Dudchenko, and M. Ietswaart, "Mobile EEG identifies the re-allocation of attention during real-world activity," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[150] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Dx7fbCW

[151] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Advances in Neural Information Processing Systems*, 2018, pp. 5334–5344.

[152] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[153] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[154] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 624–639.

[155] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[156] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7313–7324.

[157] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, "Invariant risk minimization games," in *International Conference on Machine Learning*. PMLR, 2020, pp. 145–155.

[158] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[159] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[160] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," in *Advances in Neural Information Processing Systems*, 2018, pp. 998–1008.

[161] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, “Episodic training for domain generalization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1446–1455.

[162] A. Schoenauer-Sebag, L. Heinrich, M. Schoenauer, M. Sebag, L. F. Wu, and S. J. Altschuler, “Multi-domain adversarial learning,” *arXiv preprint arXiv:1903.09239*, 2019.

[163] M. Dredze, A. Kulesza, and K. Crammer, “Multi-domain learning by confidence-weighted parameter combination,” *Machine Learning*, vol. 79, no. 1-2, pp. 123–149, 2010.

[164] J. Hoffman, M. Mohri, and N. Zhang, “Algorithms and theory for multiple-source adaptation,” in *Advances in Neural Information Processing Systems*, 2018, pp. 8246–8256.

[165] B. Neyshabur, S. Bhojanapalli, and A. Chakrabarti, “Stabilizing gan training with multiple random projections,” *arXiv preprint arXiv:1705.07831*, 2017.

[166] S. Sun, H. Shi, and Y. Wu, “A survey of multi-source domain adaptation,” *Information Fusion*, vol. 24, pp. 84–92, 2015.

[167] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, “Moment matching for multi-source domain adaptation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1406–1415.

[168] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” 2007.

[169] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 2009, pp. 248–255.

[170] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[171] Y. Li, K. Dzirasa, L. Carin, D. E. Carlson *et al.*, “Targeting EEG/LFP synchrony with neural nets,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4620–4630.

[172] I. Gulrajani and D. Lopez-Paz, “In search of lost domain generalization,” in *International Conference on Learning Representations*, 2020.

[173] Y. Li, D. E. Carlson *et al.*, “Extracting relationships by multi-domain matching,” in *Advances in Neural Information Processing Systems*, 2018, pp. 6798–6809.

[174] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *International Conference on Machine Learning*, 2015, pp. 97–105.

[175] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[176] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” *arXiv preprint arXiv:1701.07875*, 2017.

[177] D. Berthelot, T. Schumm, and L. Metz, “BEGAN: Boundary equilibrium generative adversarial networks,” *arXiv preprint arXiv:1703.10717*, 2017.

[178] Z. Lin, A. Khetan, G. Fanti, and S. Oh, “Pacgan: The power of two samples in generative adversarial networks,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 324–335, 2020.

[179] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *International Conference on Learning Representations*, 2018.

[180] I. Durugkar, I. Gemp, and S. Mahadevan, “Generative multi-adversarial networks,” *arXiv preprint arXiv:1611.01673*, 2016.

[181] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[182] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, “Theory of the hypervolume indicator: optimal μ-distributions and the choice of the reference point,” in *Proceedings of the tenth ACM SIGEVO workshop on Foundations of genetic algorithms*. ACM, 2009, pp. 87–102.

[183] ——, “Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications,” *Theoretical Computer Science*, vol. 425, pp. 75–103, 2012.

[184] A. Srivastava, L. Valkoz, C. Russell, M. U. Gutmann, and C. Sutton, “VEEGAN: Reducing mode collapse in GANs using implicit variational learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3310–3320.

[185] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[186] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[187] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.

[188] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of Wasserstein GANs,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5769–5779.

[189] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[190] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, “Adversarially learned inference,” *arXiv preprint arXiv:1606.00704*, 2016.

[191] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, “Unrolled generative adversarial networks,” *arXiv preprint arXiv:1611.02163*, 2016.

[192] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard gan,” in *International Conference on Learning Representations*, 2018.

[193] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” *arXiv preprint arXiv:2103.02503*, 2021.

[194] J. Wang, C. Lan, C. Liu, Y. Ouyang, W. Zeng, and T. Qin, “Generalizing to unseen domains: A survey on domain generalization,” *arXiv preprint arXiv:2103.03097*, 2021.

[195] J. Monteiro, X. Gibert, J. Feng, V. Dumoulin, and D.-S. Lee, “Domain conditional predictors for domain adaptation,” in *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, ser. Proceedings of Machine Learning Research, L. Bertinetto, J. F. Henriques, S. Albanie, M. Paganini, and G. Varol, Eds., vol. 148. PMLR, 11 Dec 2021, pp. 193–220. [Online]. Available: https://proceedings.mlr.press/v148/monteiro21a.html

[196] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “A survey on adversarial attacks and defences,” *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[197] P. Bashivan, B. Richards, and I. Rish, “Adversarial feature desensitization,” *arXiv preprint arXiv:2006.04621*, 2020.

[198] K. Han, B. Xia, and Y. Li, "Adversarial domain adaptation to defense with adversarial perturbation removal," *Pattern Recognition*, p. 108303, 2021.

[199] L. Zhao, T. Liu, X. Peng, and D. Metaxas, "Maximum-entropy adversarial data augmentation for improved generalization and robustness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[200] M. Awais, F. Zhou, H. Xu, L. Hong, P. Luo, S.-H. Bae, and Z. Li, "Adversarial robustness for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8568–8577.

[201] N. X. Vinh, S. Erfani, S. Paisitkriangkrai, J. Bailey, C. Leckie, and K. Ramamohanarao, "Training robust models using random projection," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 531–536.

[202] Z. Xu, D. Liu, J. Yang, C. Raffel, and M. Niethammer, "Robust and generalizable visual representation learning via random convolutions," in *International Conference on Learning Representations*, 2020.

[203] C. Hegde, M. Wakin, and R. Baraniuk, "Random projections for manifold learning," *Advances in neural information processing systems*, vol. 20, pp. 641–648, 2007.

[204] A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng, "On random weights and unsupervised feature learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 1089–1096.

[205] G.-A. Thanei, C. Heinze, and N. Meinshausen, "Random projections for large-scale regression," in *Big and complex data analysis*. Springer, 2017, pp. 51–68.

[206] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[207] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[208] M. Narayanan, V. Rajendran, and B. Kimia, "Shape-biased domain generalization via shock graph embeddings," in *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[209] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 124–140.

[210] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[211] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[212] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018.

[213] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," in *Advances in Neural Information Processing Systems*, 2019, pp. 480–491.

[214] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*, 2019, pp. 7472–7482.

[215] R. Hefron, B. Borghetti, C. Schubert Kabban, J. Christensen, and J. Estepp, “Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks,” *Sensors*, vol. 18, no. 5, p. 1339, 2018.

[216] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” in *International Conference on Learning Representations*, 2018.

[217] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.