Identification of Importance-Weighting and Geodesics on Statistical Manifolds

Author:
Masanari Kimura

Supervisor:
Prof. Hideitsu Hino

Doctor of Philosophy

Department of Statistical Science
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies, SOKENDAI

March 2024

A dissertation submitted to Department of Statistical Science,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies, SOKENDAI,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Advisory Committee

1. Prof. Kenji Fukumizu
   The Institute of Statistical Mathematics,
   SOKENDAI

2. Assoc.Prof. Keisuke Yano
   The Institute of Statistical Mathematics,
   SOKENDAI

3. Assoc.Prof. Mirai Tanaka
   The Institute of Statistical Mathematics,
   SOKENDAI

4. Prof. Takashi Takenouchi
   National Graduate Institute for Policy Studies

Acknowledgements

An immense thank you to my PhD supervisors and examiners of the advisory committee: Prof. Hideitsu Hino, Prof. Kenji Fukumizu, Assoc. Prof. Keisuke Yano, Assoc. Prof. Mirai Tanaka and Prof. Takashi Takenouchi. Support and guidance throughout the project from you all has been invaluable. Hideitsu Hino in particular you have been a fantastic primary supervisor. Thank you to all the academics who helped me get to this stage. My pre-doc supervisor, Assoc. Prof. Kei Wakabayashi, greatly helped me with my first steps in the field of informatics at the University of Tsukuba. Dr. Yoshitaka Ushiku is among the researchers whom I hold in the highest regard, and I want to express my sincere gratitude, as without your support, my doctoral challenge would not have been achievable.

I have gained great experience from the researchers who have worked with me as a industrial researcher, and thanks to everyone from the ZOZO Research. Dr. Yuki Saito and Dr. Ryotaro Shimizu in particular thanks for you all your help, support and friendship. Hiroki Naganuma, affiliated with the Université de Montréal, is collaborating with me on a joint research project, and his exceptional insight and implementation skills have been a tremendous source of inspiration for me.

I would like to express my gratitude to my collaborators and friends for a very valuable experience. Dr. Kento Nozawa kindly discussed with me a fruitful discussion on machine learning. Mr. Makoto Hiramatsu and Mr. Kosuke Takenaka have always shared useful knowledge with me to help my understanding of the computational resources. Mr. Ryohei Izawa and Mr. Ryosuke Sato have been good friends and worked with me in a friendly competition. Dr. Aaron C. Bell and Dr. Motaz Sabri are former colleagues who have always been friendly and willing to talk about subjects I was not familiar with.

Thanks of course to my family for their support. Finally, thank you Nozomi for always being so supportive and helping me every step of the way.

THE GRADUATE UNIVERSITY FOR ADVANCED STUDIES, SOKENDAI

Abstract

School of Multidisciplinary Sciences
Department of Statistical Science

Doctor of Philosophy

Identification of Importance-Weighting and Geodesics on Statistical Manifolds

by Masanari Kimura

It is known that the set of probability distributions forms a Riemannian manifold, namely the statistical manifold, and the research area that explores the geometric properties of such manifolds is called information geometry. In the framework of information geometry, several statistical concepts can be identified with geometric objects. Such identifications can lead to a deeper understanding of statistical procedures, improvements to existing methods, and geometrically natural generalizations of algorithms. In this study, we consider the identification of the importance-weighting for probability distributions with the selection of curves on the statistical manifold, following this framework. We introduce that this identification leads to generalization of two statistical concepts. One is the generalization of a family of algorithms called importance-weighted empirical risk minimization, designed to adapt to covariate shifts where the marginal distributions of input variables in the training and test data differ. We also demonstrate that, by appropriately choosing the parameters introduced by the generalization, the proposed method can outperform existing algorithms by numerical experiments. The other is a generalization of a function measuring the difference between two probability distributions called skew divergence. We explore various properties of this generalized divergence and introduce its inclusion of several existing divergences. In particular, the value of the generalized divergence maintains continuity with respect to the parameters introduced during generalization. Therefore, it is expected that by optimizing these parameters, existing algorithms dependent on skew divergence can be improved.

Contents

List of Figures

List of Tables

Symbols

$\mathbb{N}$	The set of all natural numbers.
$\mathbb{R}$	The set of all real numbers.
$\emptyset$	The empty set.
A^c	The complement of set A.
$A \cup B$	The union of two sets A and B.
$A \cap B$	The intersection of two sets A and B.
$A \setminus B$	The set difference of set A and set B.
$\mathrm{Id}(\boldsymbol{x})$	Identity map, $\mathrm{Id}(\boldsymbol{x}) = \boldsymbol{x}$.
δ_{ij}	Kronecker delta, $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$.
$(\mathcal{M}, g)$	The Riemannian manifold endowed with a Riemannian metric g.
(g, ∇, ∇^*)	The dualistic structure on $\mathcal{M}$ with a metric g, and dual connections ∇ and ∇^*.
$(\mathcal{M}, g, \nabla, \nabla^*)$	The statistical manifold endowed with a dualistic structure (g, ∇, ∇^*).
$\mathcal{F}(\mathcal{M})$	The set of all differentiable functions on the manifold $\mathcal{M}$.
$T_p\mathcal{M}$	The tangent space of manifold $\mathcal{M}$ at point $p \in \mathcal{M}$.

Chapter 1

Introduction

Our interest lies in considering the geometric properties of a manifold constructed by a set of probability distributions. On the basis of such a geometric framework, it is possible to identify statistical concepts with geometric objects. It is known that this identification allows us to achieve a deeper understanding of statistical procedures, improvements to existing methods, and the induction of geometrically natural generalized algorithms. In this study, we demonstrate the equivalence between the weighting of probability distributions and the selection of curves on a manifold. Furthermore, we show that this identification leads to the following generalizations: i) a generalization of algorithms designed to handle differences in the distribution of input variables between training and testing data, that is the covariate shift adaptation, and ii) a generalization of a stable version of a divergence, namely the skew divergence, used to measure differences between probability distributions. The first generalization concerns a problem setting in which the probability distributions of the data during training and testing are different in machine learning tasks, which have attracted a great deal of attention in recent years. In particular, the validity of the framework called supervised learning is supported by the identity of these distributions, and it is very important to consider shifts in the distributions. The second generalization is about a function called divergence that evaluates differences between probability distributions, which is used very often in statistics and machine learning. Skew divergence takes into account the numerical instability of the most commonly used divergence called Kullback–Leibler divergence or KL-divergence, which has many known applications. We expect that the

identification of these important concepts with geometrical concepts could lead to the realization of various beneficial applications.

1.1 Background

1.1.1 Identification of Statistical Concepts and Geometric Objects

A set of probability distributions forms a Riemannian manifold, namely, the statistical manifold. Indeed, when a probability distribution with a parameter $\boldsymbol{\theta}$ denoted as $p(\cdot;\boldsymbol{\theta})$ is uniquely determined with a specific $\boldsymbol{\theta}$, the set $\{p(x;\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$ forms a Riemannian manifold with $\boldsymbol{\theta}$ values as coordinates. By studying the geometric properties of this statistical manifold, we can expect to gain various insights. The framework for analyzing the geometric properties of statistical manifolds, known as information geometry [8, 9, 14, 19], is notably used in various applied studies. See Appendix C for more details of information geometry. Within the framework of information geometry when considering statistical procedures, we can identify several statistical concepts with geometric objects, which leads to a deeper understanding of statistical procedures, improvements to existing methods, and geometrically natural generalizations of algorithms. Examples of the identified statistical and geometric concepts are as follows.

- The E- and M-steps of the EM algorithm can be identified with special projections called the e- and m-projections between the data manifold and the model manifold [10, 116].
- Natural gradient descent characterizes the steepest direction on a statistical manifold using the inverse matrix of the Fisher information matrix [11].
- Jeffreys prior is one of the non-informative priors that remain invariant under parameter transformations, and it can be considered as the volume element of statistical manifolds. Furthermore, Jeffreys' prior can be generalized within the framework of information geometry [194, 282].

In this study, we demonstrate the equivalence between the weighting of probability distributions and the selection of curves on a manifold. Furthermore, we show that this identification leads to the following two generalizations:

- a generalization of algorithms designed to handle differences in the distribution of input variables between training and testing data, known as the covariate shift adaptation, and

- a generalization of a stable version of a divergence, known as the skew divergence, used to measure differences between probability distributions.

In the following, we introduce the background for each problem setting.

1.1.2 Generalization of Covariate Shift Adaptation Methods

Machine learning [37, 134, 140, 209] is a subfield of artificial intelligence that involves the use of algorithms and statistical models to enable computer systems to learn and improve their performance on a specific task or problem from data without being explicitly programmed. More formally, machine learning refers to the process in which a computer program learns from experience to improve its performance on tasks [134, 165, 201, 343]. Machine learning has advanced markedly over the past two decades and involves numerous applications such as computer vision [80, 103, 126, 200, 268], natural language processing [70, 225, 272, 332], and recommender systems [198, 239, 338]. Especially in recent years, there has been a growing interest in advanced applications such as generative models capable of producing images and artwork that are indistinguishable from those created by humans [64, 335], as well as high-capacity chat systems capable of information retrieval, question answering, and text generation [190, 342].

Machine learning tasks are categorized into the following three groups on the basis of their learning approach and behavior; Supervised learning is an approach that uses pairs of input data and output labels to learn the rules between inputs and outputs. Many fundamental learning tasks are classified under supervised learning [48, 217]. Unsupervised learning aims to extract a certain structure form solely from input data. For example, clustering [43, 127, 319], dimensionality reduction [117, 249], and self-organizing maps [149, 150] are categorized into unsupervised learning. Reinforcement learning allows an agent to acquire knowledge within an interactive environment through experimentation and learning from the feedback generated by its own actions and encounters [16, 138, 281]. Furthermore, there exists a paradigm called semi-supervised learning [296, 345], which is a combination of supervised and unsupervised learning. In this study, we particularly focus on supervised learning.

Supervised learning aims to acquire a function that best approximates the relationship between input vectors and outputs using pairs of input vectors and corresponding outputs as labeled training data. The goal of supervised learning is to predict outputs for novel data using a model that has been trained on a labeled dataset. The learning algorithms most commonly used are, for example, support vector machines [113], linear regression [204], logistic regression [119], decision trees [251] and neural networks [98, 105, 112] for classification and regression.

The procedures of these machine learning algorithms all depend on the provided training dataset. In particular, all approaches of supervised, unsupervised, and reinforcement learning rely on input vectors. Here, we consider why the fit to the training data leads to a fit to the test data. This involves considering the assumptions necessary for learning from the given data to enable predictions of unknown data. Intuitively, this assumes that the tendencies of instances present in the test dataset are similar to those of present in the training dataset. For example, consider the task of predicting the risk of a certain disease based on information about a patient. If the disease is more likely to occur in older individuals than in younger ones, it is desirable that the distribution of patient ages in the training dataset is similar to that in the test dataset. Alternatively, if the disease is more prone to affect males than females, a similar expectation about patient gender can also be made. Intuitively, an algorithm trained on a predominantly female dataset might not be reliable for making predictions on unfamiliar datasets in which males constitute the majority.

From the above example, the similarity in distribution between the training and test data seems to be a necessary assumption in supervised learning. As a stronger assumption, the well-known independent and identically distributed (i.i.d.) assumption is often imposed, requiring the independence of each instance.

Supervised learning, which optimizes with respect to the empirical risk of the training data, is often formulated within the framework of empirical risk minimization (ERM). The reason why it can be concluded that ERM works well is that this framework exhibit statistically desirable properties namely consistency and unbiasedness (see Chapter 2). These properties are based on the i.i.d. assumption between the training and test data.

However, such assumptions are often violated in real-world problem settings. We consider the task of disease risk prediction again as an example. For instance, when the hospital providing the training dataset differs from that holding the data for prediction, it can be anticipated that the identical distributions assumption is more likely to be violated. In one hospital, owing to its geographical location, there might be a higher proportion of elderly patients, whereas in another hospital, the presence of a renowned female physician might result in a larger number of female patients. In such situations, it is common for patient distributions to vary between hospitals, often leading to the violation of the i.i.d. assumption. As another example, we consider chat systems based on large-scale language models, which are currently attracting significant attention. Although these chat systems exhibit high performance, they have been reported to lack political neutrality [196, 197], have ethical bias [156, 244], or have performance gap among languages [163, 180]. These reports also stem from differences in the distribution of data. The behavior of such chat systems relies on the dominant presence of political

and ethical biases within the training dataset, as well as the proportion of English text within the dataset.

The situation where the distributions of the training and test datasets differ is referred to as distribution shift, also known as dataset shift [159, 206, 236, 241]. Similar problem formulations include out-of-distribution (OOD) generalization [212, 259, 328], distributionally robust optimization (DRO) [75, 91, 242], domain adaptation [25, 232, 304], transfer learning [226, 289, 307, 346], active learning [115, 257, 258], continual learning [187, 228], and sampling bias correction [28, 122] among others. These related studies are reviewed in Appendix B.

The covariate shift assumption [260] is one of the problem settings within the context of distribution shift. Covariate shift assumes that the marginal distributions of input vectors between the training and test data are different. The phenomenon of covariate shift can frequently occur owing to factors such as limited resources during data collection, sampling bias, and experimental design. See Chapter 2 for the detailed definition of the covariate shift assumption. An illustrative example involves a model aimed at classifying and distinguishing images of dogs from those of other animals. During training, the model learns to identify distinct dog features using labeled datasets. However, the training data might lack a full spectrum of dog breeds, leading to gaps in knowledge. Consequently, when the model is used, it may struggle to accurately identify dog breeds not encountered in the training data owing to the covariate shift phenomenon. Other examples are emotion recognition [109], image generation [254], 3D pose estimation [321], sleep stage classification [141], and other tasks [110, 256, 270, 294, 302, 306].

The most fundamental strategy for covariate shift adaptation is importance weighting [260, 277, 322]. The key idea behind importance weighting is to adjust the weights of data points, assigning lower weights to the data generated from the training distribution and higher weights to the data generated from the test distribution, to perform parameter estimation effectively. The most typical weighting is based on the density ratio between the training distribution and the test distribution. Moreover, for reasons such as stability and model constraints, several variants exist [82, 179, 192, 322].

In Chapter 3, we focus on the observation that a certain group of these covariate shift adaptation methods can be identified with a collection of geometric objects on the statistical manifold formed by sets of probability distributions. We demonstrate that the selection of covariate shift adaptation methods can be identified with the choice of curves that connects the training and test distributions. Through this identification, we introduce the pair of parameters that determine the shape of the curve and its position on the curve. This enables us to generalize multiple covariate shift adaptation methods. Furthermore, we experimentally demonstrate that the performance of the trained model

changes relatively smoothly with variations in these parameters. On the basis of these observations, we show that through appropriate parameter optimization, our generalized covariate shift adaptation outperforms existing methods in terms of performance.

1.1.3 Generalization of Skew Divergence

In Chapter 4, we introduce that the equivalence between weighting and curves on the manifold leads to a generalization of divergence. Here, in statistics, divergence refers to a type of pseudo-distance used to evaluate the differences between probability distributions. The term pseudo-distance is used because divergence does not satisfy symmetry or triangle inequality. In particular, one of the most important divergences in statistics and machine learning is the KL-divergence [160]. The KL-divergence plays a central role in many contexts. For instance, cross-entropy [67, 243], which commonly appears as an objective function in machine learning, is closely related to the KL-divergence. As another example, in more modern machine learning approaches, it has been reported that the regularization term based on the KL-divergence contributes to improving the performance of models [95, 229, 333]. Moreover, in the framework of PAC-Bayes learning, which is used for the analysis of generalization bounds for machine learning models, the KL-divergence plays a pivotal role [7, 100, 195].

Mathematically, the KL-divergence is given as the expectation of the logarithm of the density ratio of two probability distributions with respect to one of the distributions (see Chapter 2 for the definition of the KL-divergence). Owing to the inclusion of the density ratio in its definition, the KL-divergence has the potential to diverge to infinity if there is a discrepancy in the supports of the two probability distributions. Furthermore, there are cases in which the asymmetry of the KL-divergence becomes problematic. To address these limitations, there are various variants and symmetrizations of the KL-divergence [23, 224, 303].

The skew divergence [49, 59, 94, 168, 247] is one of the variants of the KL-divergence and has numerous applications. The skew divergence addresses the issue of the KL-divergence by using a weighted average of the densities of the two distributions as the denominator, which alleviates the problem of divergence to infinity that was present in the KL-divergence. Here, we introduce how the generalization of the skew divergence can also be derived through the equivalence between averaging operations and the selection of curves on the manifold. This generalization exhibits the properties that the divergence should satisfy.

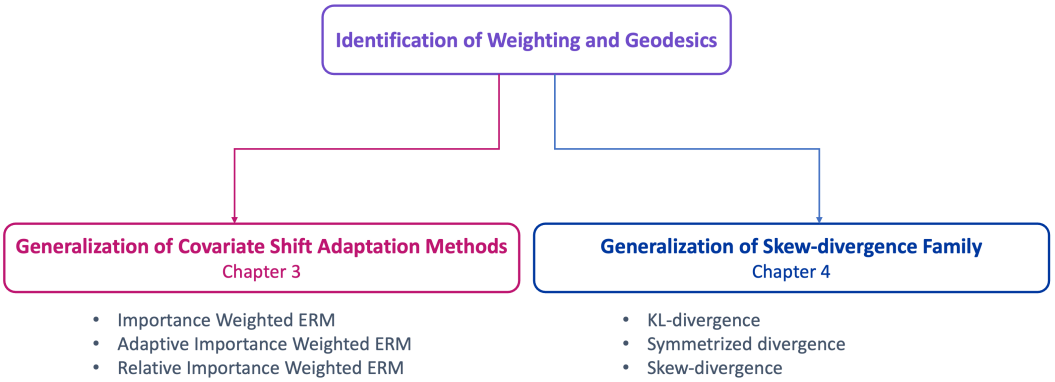

FIGURE 1.1: Identification of weighting and geodesics leads to two generalizations.

1.2 Contributions

Our contributions are summarized as follows:

- We show that the selection of covariate shift adaptation methods can be identified with the choice of curves on the statistical manifold that connect the training and test distributions.
- We suggest that this identification leads to the introduction of a pair of parameters that determine the shape of the curve and its position on the curve. This enables us to generalize covariate shift adaptation methods from the viewpoint of information geometry.
- Furthermore, we consider that the equivalence between importance weighting and curve selection leads to a new class of generalizations for the KL-divergence (see Fig. 1.1).

Our generalization provides a deeper understanding of algorithms or divergences and, increases the existing algorithms through better selection of parameters to be introduced. Furthermore, we can expect to derive generalizations of various other existing statistical procedures in a framework similar to these generalizations.

The organization of this thesis is as follows. In Chapter 2, we review the related literature. In addition, Appendix B provides additional related work. In Chapter 3, we introduce the generalization of covariate shift adaptation methods from the viewpoint of information geometry. In Chapter 4, we generalize the skew-divergence family on the basis of the identified weighting and geodesics. Finally, in Chapter 5, we provide the conclusion and possible future work. We would also like to mention that the content of this work is based on previously published journal articles [145, 146].

Chapter 2

Preliminaries

This chapter introduces the notation required throughout this study, as well as related research. First, we define the notation for basic random variables and probability distributions. Second, we summarize prerequisites for the framework of Empirical Risk Minimization and the assumptions of distribution shift, which are the focus of Chapter 3. Third, we define a notation for divergence, which is the main topic of discussion in Chapter 4. Finally, we introduce a preparation on the information geometry commonly used in these chapters. We also introduce known examples of the identification of statistical concepts with geometric objects and define the tools we focus on in particular.

2.1 Statistical Models

First, let us review the definitions of statistical models. For more details, please refer to textbooks on probability and statistics [29, 36, 230].

2.1.1 Discrete and Continuous Random Variables

Let $\mathcal{F}$ be a set of measures. A discrete random variable X takes finite or at most countably infinite values. Its probability is described by a probability mass function $p : \mathcal{X} \to [0, 1]$ as

$$p(\boldsymbol{x}) = P(X = \boldsymbol{x}), \tag{2.1}$$

satisfying $\sum_{\boldsymbol{x}\in X} p(\boldsymbol{x}) = 1$, where $\mathcal{X}$ is the support of X and $p \in \mathcal{F}$.

A continuous random variable X is a random variable taking a continuous range of values. Its probability distribution has a probability density function $p : \mathcal{X} \to \mathbb{R}_+$ satisfying $p(\boldsymbol{x}) \geq 0$, $(\forall \boldsymbol{x} \in \mathcal{X})$ and $\int_{\mathcal{X}} p(\boldsymbol{x})dx = 1$. Note that for two probability densities $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$, the convex combination

$$p(\boldsymbol{x}) = \lambda p_1(\boldsymbol{x}) + (1-\lambda)p_2(\boldsymbol{x}), \quad \lambda \in [0,1] \tag{2.2}$$

is also a probability density. Note that for some $S \subseteq \mathcal{X}$, $\int_{\mathcal{S}} p(\boldsymbol{x})d\boldsymbol{x} = \mathbb{P}(\boldsymbol{x} \in \mathcal{S})$.

It is worth noting that any discrete distribution can be seen as having a probability density function $\tilde{p}(\tilde{\boldsymbol{x}})$ as

$$\tilde{p}(\tilde{\boldsymbol{x}}) = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\delta(\tilde{\boldsymbol{x}} - \boldsymbol{x}), \tag{2.3}$$

where $p(\boldsymbol{x})$ is a probability mass function and δ is the Dirac delta function.

2.1.2 Parametric Models

Let $\Theta \subset \mathbb{R}^n$ be the parameters space and $p(\cdot;\boldsymbol{\theta}) \in \mathcal{F}$ be a probability distribution parameterized by $\boldsymbol{\theta} \in \Theta$. We consider a family of probability distributions on $\mathcal{X}$ as

$$\mathcal{S} = \{p(\cdot;\boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \Theta\} \tag{2.4}$$

where $n \in \mathbb{N}$ is the dimension of Θ. The set $\mathcal{S}$ is called a statistical model or a parametric model of dimension n.

2.1.3 Exponential and Mixture Family

Here, we introduce two distinguish families of probability distributions, called the exponential family and the mixture family. For $i = 1, \ldots, n$, consider $n+1$ smooth functions $k(\boldsymbol{x})$, $T_i(\boldsymbol{x})$ on $\mathcal{X}$ such that $T_i(\boldsymbol{x})$ is linearly independent. Consider the following probability density

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left\{k(\boldsymbol{x}) + \theta^i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{2.5}$$

where $\boldsymbol{x} \in \mathcal{X}$ and

$$\psi(\boldsymbol{\theta}) = \log\left(\int_{\mathcal{X}} e^{k(\boldsymbol{x})+\theta^i T_i(\boldsymbol{x})} d\boldsymbol{x}\right) \tag{2.6}$$

is the normalization function. In Eq. (2.5), and hereafter the Einstein summation convention will be assumed, so that summation will be automatically taken over indices repeated twice in superscript and subscript of the term, e.g., $a^i b_i = \sum_i a^i b_i$. The statistical model $\mathcal{S} = \{p(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$, with $p(\boldsymbol{x};\boldsymbol{\theta})$ is given by Eq. (2.5), is called an exponential family and $\boldsymbol{\theta}$ is its natural parameter. Note that Eq. (2.5) can be written equivalently as

$$p(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\theta^i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{2.7}$$

with $h(\boldsymbol{x}) = e^{k(\boldsymbol{x})}$.

Example 2.1. *For $\mathcal{X} = [0,\infty)$, $\Theta = (0,\infty)$, $k(x) = 0$, $T_1(x) = -x$, $\theta^1 = \theta$ and $\psi(\theta) = -\log\theta$, $p(x;\theta)$ is the exponential distribution with parameter θ as*

$$p(x;\theta) = \theta e^{-\theta x}. \tag{2.8}$$

Example 2.2. *For $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}\times(0,\infty)$, $k(x) = 0$, $T_1(x) = x$, $T_2(x) = x^2$, $\theta^1 = \frac{\mu}{\sigma^2}$, $\theta^2 = -\frac{1}{2\sigma^2}$ and*

$$\psi(\boldsymbol{\theta}) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta^2}\right),$$

$p(x;\boldsymbol{\theta})$ is the Gaussian distribution with parameter $\boldsymbol{\theta}$ and $(\mu,\sigma) \in \mathbb{R}\times(0,\infty)$.

Let $F_i : \mathcal{X} \to \mathbb{R}$ be n smooth functions, linearly independent on $\mathcal{X}$, satisfying

$$\int_{\mathcal{X}} F_i(\boldsymbol{x}) d\boldsymbol{x} = 0, \quad 1 \le i \le n. \tag{2.9}$$

Let $C(\boldsymbol{x})$ with $\int_{\mathcal{X}} C(\boldsymbol{x}) = 1$, and consider the following family of probability densities.

$$p(\boldsymbol{x};\boldsymbol{\theta}) = C(\boldsymbol{x}) + \theta^i F_i(\boldsymbol{x}). \tag{2.10}$$

The statistical manifold $\mathcal{S} = \{p(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$, with $p(\boldsymbol{x};\boldsymbol{\theta})$ given by Eq. (2.10), is called a mixture family and $\boldsymbol{\theta}$ is its mixture parameter.

Example 2.3. *Consider $n+1$ probability densities $p_1, p_2, \ldots, p_n, p_{n+1}$, which are linearly independent on $\mathcal{X}$. Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^n \mid \sum_{i=1}^n \theta^i < 1, \theta^i > 0\}$. The following weighted average is also a probability density*

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \theta^i p_i(\boldsymbol{x}) + \left(1 - \sum_{i=1}^n \theta^i\right) p_{n+1}(\boldsymbol{x}) \tag{2.11}$$

$$= p_{n+1}(\boldsymbol{x}) + \sum_{i=1}^n \theta^i (p_i(\boldsymbol{x}) - p_{n+1}(\boldsymbol{x})), \tag{2.12}$$

which becomes a mixture family with $C(\boldsymbol{x}) = p_{n+1}(\boldsymbol{x})$ *and* $F_i(\boldsymbol{x}) = p_i(\boldsymbol{x}) - p_{n+1}(\boldsymbol{x})$.

The operation of weighted averaging of probability distributions in Example 2.3 becomes crucial in the subsequent chapters when generalizing algorithms and divergences.

Definition 2.1. The set of the strictly positive probability measure $\mathcal{P}$ is defined as

$$\mathcal{P} \coloneqq \left\{ p \in \mathcal{F} \mid p(\boldsymbol{x}) > 0 \ (\forall \boldsymbol{x} \in \mathcal{X}), \int_{\mathcal{X}} p(\boldsymbol{x}) d\boldsymbol{x} = 1 \right\}, \tag{2.13}$$

and the set of nonnegative probability measure $\mathcal{P}_+$ is defined as

$$\mathcal{P}_+ \coloneqq \left\{ p \in \mathcal{F} \mid p(\boldsymbol{x}) \geq 0 \ (\forall \boldsymbol{x} \in \mathcal{X}), \int_{\mathcal{X}} p(\boldsymbol{x}) d\boldsymbol{x} = 1 \right\}, \tag{2.14}$$

2.2 Empirical Risk Minimization and Covariate Shift

First, we formulate the problem of supervised learning. We denote by $\mathcal{X} \subset \mathbb{R}^d$ the d-dimensional input space, where $d \in \mathbb{N}$. The output space is denoted by $\mathcal{Y} \subset \mathcal{R}$ (regression) or $\mathcal{Y} \subset \{1, \ldots, K\}$ (K-class classification). We assume that training examples $\{(\boldsymbol{x}_i^{tr}, y_i^{tr})\}_{i=1}^{n_{tr}}$ are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution $p_{tr}(\boldsymbol{x}, y)$, which can be decomposed into the marginal distribution and the conditional probability distribution, i.e., $p_{tr}(\boldsymbol{x}, y) = p_{tr}(\boldsymbol{x}) p_{tr}(y|\boldsymbol{x})$. We also denote the test examples by $\{(\boldsymbol{x}_i^{te}, y_i^{te})\}_{i=1}^{n_{te}}$ drawn from a test distribution $p_{te}(\boldsymbol{x}, y) = p_{te}(\boldsymbol{x}) p_{te}(y|\boldsymbol{x})$. Here $n_{tr} \in \mathbb{N}$ and $n_{te} \in \mathbb{N}$ are training and test sample size, respectively.

Let $\mathcal{H}$ be a hypothesis class. The goal of supervised learning is to obtain a hypothesis $h : \mathcal{X} \to \mathbb{R}$ ($h \in \mathcal{H}$) with the training examples that minimizes the expected loss over the test distribution.

Definition 2.2 (Expected error)**.** Given a hypothesis $h \in \mathcal{H}$ and probability distribution $p(\boldsymbol{x}, y)$ the expected error $\mathcal{R}(h; p)$ is defined as

$$\mathcal{R}(h; p) \coloneqq \mathbb{E}_{(\boldsymbol{x}, y) \sim p(\boldsymbol{x}, y)} \Big[\ell(h(\boldsymbol{x}), y) \Big], \tag{2.15}$$

where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is the loss function that measures the discrepancy between the true output value y and the predicted value $\hat{y} \coloneqq h(\boldsymbol{x})$.

The expected error of a hypothesis h is not directly accessible since the underlying distribution $p(\boldsymbol{x}, y)$ is unknown. Then, we have to measure the empirical error of hypothesis h on the observable labeled sample.

Definition 2.3 (Empirical error)**.** Given a hypothesis $h \in \mathcal{H}$ and labeled sample $S_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the empirical error $\hat{\mathcal{R}}(h; S_n)$ is defined as

$$\hat{\mathcal{R}}(h; S_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell(h(\boldsymbol{x}_i), y_i), \tag{2.16}$$

where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is the loss function.

This procedure, known as Empirical Risk Minimization (ERM), is known to be useful in learning with i.i.d. settings [37, 111, 128, 276, 298, 299]. For example, we assume that the labeled sample S is a set of i.i.d. examples according to distribution $p(\boldsymbol{x}, y)$. Then, the empirical error $\hat{\mathcal{R}}(h; S)$ is an unbiased estimator of the expected error $\mathcal{R}(h; p)$ as follows.

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{x},y)} \left[\hat{\mathcal{R}}(h; S) \right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{x},y)} \left[\ell(h(\boldsymbol{x}), y) \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{R}(h; p) \\ &= \mathcal{R}(h; p). \end{aligned} \tag{2.17}$$

Moreover, ERM has the following property.

Proposition 2.4. *We assume that the labeled sample S_n is a set of i.i.d. samples according to distribution $p(\boldsymbol{x}, y)$. We also assume that the empirical error $\hat{\mathcal{R}}(h; S_n)$ convergence uniformly to the expected error $\mathcal{R}(h; p)$. Then, the empirical error $\hat{\mathcal{R}}(h; S_n)$ is a consistent estimator of the expected error $\mathcal{R}(h; p)$.*

Proof. Let $h_{S_n} = \arg\min_{h \in \mathcal{H}} \hat{R}(h; S_n)$ and $h^* = \arg\min_{h \in \mathcal{H}} R(h; p)$. For all $h \in \mathcal{H}$, we have

$$R(h; p) - R(h^*; p) \geq 0, \tag{2.18}$$

$$\hat{R}(h; S_n) - \hat{R}(h_{S_n}; S_n) \geq 0. \tag{2.19}$$

Since h in Eq. (2.18) and (2.19) can be taken to be h_{S_n} and h^*, respectively, we obtain the following.

$$R(h_{S_n}; p) - R(h^*; p) \geq 0, \tag{2.20}$$

$$\hat{R}(h^*; S_n) - \hat{R}(h_{S_n}; S_n) \geq 0. \tag{2.21}$$

From Eq. (2.20) and (2.21),

$$\begin{aligned} 0 &\leq R(h_{S_n}; p) - R(h^*; p) + \hat{R}(h^*; S_n) - \hat{R}(h_{S_n}; S_n) \\ &= R(h_{S_n}; p) - \hat{R}(h_{S_n}; S_n) + \hat{R}(h^*; S_n) - R(h^*; p) \\ &\leq \sup_{h \in \mathcal{H}} \Big(R(h; p) - \hat{R}(h; S_n) \Big) + \hat{R}(h^*; S_n) - R(h^*; p). \end{aligned} \tag{2.22}$$

Due to the law of large numbers, the second half of the right hand side converges in probability as

$$\left| \hat{R}(h^*; S_n) - R(h^*; p) \right| \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{2.23}$$

and

$$\sup_{h \in \mathcal{H}} \left| R(h; p) - \hat{R}(h; S_n) \right| \xrightarrow{P} 0 \text{ as } n \to \infty, \tag{2.24}$$

Then, we have

$$R(h_{S_n}; p) - R(h^*; p) \xrightarrow{P} 0 \text{ as } n \to \infty, \tag{2.25}$$

$$\hat{R}(h^*; S_n) - \hat{R}(h_{S_n}; S_n) \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{2.26}$$

From Eq. (2.24) to (2.26), we have the proof. □

As discussed above, ERM has desirable properties when training and test data are generated from the identical distribution. However, in real-world scenarios, this assumption is frequently violated. This is because the data used to train a model is often collected under different conditions or at a different time than the data that the model is expected to predict. For example, a model trained on images of faces captured in a controlled environment may not perform well when applied to images of faces captured in the wild, due to differences in lighting, pose, and background. Similarly, a model trained on financial data collected during a period of economic stability may not generalize well to a period of economic uncertainty, where the underlying distribution of the data has shifted. This mismatch between the training and test data distributions is known as distribution shift, and it is a common problem in real-world machine learning applications. Failing to account for distribution shifts can lead to models that perform poorly on new data, reducing their effectiveness in practice. This assumption, where the training distribution and the test distribution differ, is referred to as distribution shift. Specifically, the following definitions of distribution shift are known.

Definition 2.5 (Covariate shift [260])**.** We consider that the two distributions $p_{tr}(\boldsymbol{x}, y)$ and $p_{te}(\boldsymbol{x}, y)$ satisfy the covariate shift assumption if the following conditions hold:

$$p_{tr}(\boldsymbol{x}) \neq p_{te}(\boldsymbol{x}),$$
$$p(y|\boldsymbol{x}) = p_{tr}(y|\boldsymbol{x}) = p_{te}(y|\boldsymbol{x}).$$

Covariate shift refers to a situation where the distribution of the input features (also known as covariates) changes between the training and testing phases of a machine learning problem. This means that the statistical properties of the input features used to train a model are different from those of the features used to test the model.

Definition 2.6 (Target shift [337])**.** We consider that the two distributions $p_{tr}(\boldsymbol{x}, y)$ and $p_{te}(\boldsymbol{x}, y)$ satisfy the target shift assumption if the following conditions hold:

$$p_{tr}(y) \neq p_{te}(y),$$
$$p(\boldsymbol{x}|y) = p_{tr}(\boldsymbol{x}|y) = p_{te}(\boldsymbol{x}|y).$$

Target shift, also known as label shift, occurs when the distribution of the target variable changes between the training and testing phases of a machine learning problem.

Definition 2.7 (Concept drift [309])**.** We consider that the two distributions $p_{tr}(\boldsymbol{x}, y)$ and $p_{te}(\boldsymbol{x}, y)$ satisfy the concept drift assumption if the following conditions hold:

$$p_{tr}(y|\boldsymbol{x}) \neq p_{te}(y|\boldsymbol{x}),$$
$$p_{tr}(\boldsymbol{x}|y) \neq p_{te}(\boldsymbol{x}|y).$$

Concept drift refers to a situation where the underlying relationship between the input features and the target variable changes over time. This means that the relationship that the model learned during training is no longer valid or accurate in the testing phase. All three phenomena can lead to a mismatch between the training and testing phases of a machine learning problem, which can result in poor model performance.

Definition 2.8 (General distribution shift [274])**.** Let $\mathcal{Z} \subset \{\mathcal{X}, \mathcal{Y}\}$ be a set of immutable variables whose marginal distribution should remain fixed, $\mathcal{W} \subset \{\mathcal{X}, \mathcal{Y}\} \setminus \mathcal{Z}$ be a set of mutable variables whose distribution can be shifted, and $\mathcal{V} = \{\mathcal{X}, \mathcal{Y}\} \setminus (\mathcal{W} \cup \mathcal{Z})$ be the remaining dependent variables. This partition of the variables defines a factorization of p_{tr} into

$$p_{tr}(\boldsymbol{v}|\boldsymbol{w}, \boldsymbol{z}) p_{tr}(\boldsymbol{w}|\boldsymbol{z}) p_{tr}(\boldsymbol{z}), \tag{2.27}$$

where $\boldsymbol{z} \in \mathcal{Z}$, $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{v} \in \mathcal{V}$. We consider that the two distributions $p_{tr}(\boldsymbol{x}, y)$ and $p_{te}(\boldsymbol{x}, y)$ satisfy the general distribution shift assumption if the following hold:

$$p_{tr}(\boldsymbol{w}|\boldsymbol{z}) \neq p_{te}(\boldsymbol{w}|\boldsymbol{z}). \tag{2.28}$$

Note that this formulation generalizes other distribution shifts. For example, if we let $\mathcal{Z} = \emptyset$ and $\mathcal{W} = \mathcal{X}$, then this corresponds to the covariate shift assumption.

In this study, we particularly focus on the covariate shift assumption. The covariate shift assumption has several known aliases and variants and has been studied extensively. For example, while closely related to the covariate shift assumption, many studies also treat similar problem settings as sample selection bias [28, 291, 300, 311]. Several studies [292, 334] introduced a random variable s to model sample selection bias, and considered the construction of the test distribution as follows:

$$p_{te}(x, y) = p(x, y) = \sum_{s} p(x, y, s), \tag{2.29}$$

where $s = 1$ means the example is selected, and $s = 0$ means the example is not selected. As another variant, the subpopulation shift [255, 326] in machine learning refers to a situation where the statistical properties of a particular subset or subpopulation of data change between the training and testing phases. This means that the distribution of data within a specific subgroup, such as a certain demographic, location, or any other defined subset, is different in the training and testing datasets. For more specific applications in the advertising industry, predicting how many clicks are associated with purchases is crucial. However, in the real world, it is known that there is a time lag between clicks and actual purchases. This problem is referred to as a feedback shift or delayed feedback problem. Delayed feedback was initially researched by Chapelle [52] and has since led to the proposal of numerous techniques to address this problem [157, 250, 283, 330]. Furthermore, several studies have explored the concept of delayed feedback in the context of bandit problems [135, 237, 301].

As noted above, there are many assumptions associated with covariate shift. In the following, we provide a concise review of related research on the covariate shift assumption.

2.2.1 Covariate Shift Adaptation

Covariate shift is one of the assumptions originally proposed by Shimodaira [260]. Covariate shift adaptation is a research area focused on methods to acquire models that achieve good performance in datasets where covariate shift is assumed. Numerous approaches for covariate shift adaptation have been proposed to ensure the performance of

machine learning models under the assumption of covariate shift. From the definition of covariate shift (Definition 2.5), it is clear that ordinal ERM fails under covariate shift. Specifically, in the ordinal ERM approach, the expected value of empirical risk under the training distribution does not match the expected risk under the test distribution.

Importance weighting has been shown to be effective in mitigating the effect of covariate shift [260, 277, 302]:

$$\min_{h\in\mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} w(\boldsymbol{x}_i^{tr}) \ell(h(\boldsymbol{x}_i^{tr}), y_i^{tr}), \tag{2.30}$$

where $w : \mathcal{X} \to (0, \infty)$ is a certain weighting function.

Definition 2.9 (Importance-Weighted ERM (IWERM) [260])**.** The Importance-Weighted ERM (IWERM) is defined as a modified ERM by using the density ratio $p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$ as the weighting function:

$$\min_{h\in\mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \frac{p_{te}(\boldsymbol{x}_i^{tr})}{p_{tr}(\boldsymbol{x}_i^{tr})} \ell(h(\boldsymbol{x}_i^{tr}), y_i^{tr}). \tag{2.31}$$

IWERM is known to have unbiasedness and consistency under the covariate shift assumption. For example, the unbiasedness of IWERM can be proved as follows.

$$\begin{aligned}
\mathbb{E}_{p_{tr}(\boldsymbol{x},y)} \left[\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} \ell(h(\boldsymbol{x}), y) \right] &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} \ell(h(\boldsymbol{x}), y) \cdot p_{tr}(\boldsymbol{x}, y) d\boldsymbol{x} dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} \ell(h(\boldsymbol{x}), y) \cdot p_{tr}(\boldsymbol{x}) p_{tr}(y|\boldsymbol{x}) d\boldsymbol{x} dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} p_{te}(\boldsymbol{x}) \ell(h(\boldsymbol{x}), y) \cdot p_{tr}(y|\boldsymbol{x}) d\boldsymbol{x} dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} p_{te}(\boldsymbol{x}) \ell(h(\boldsymbol{x}), y) \cdot p_{te}(y|\boldsymbol{x}) d\boldsymbol{x} dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}), y) \cdot p_{te}(\boldsymbol{x}, y) d\boldsymbol{x} dy \\
&= \mathbb{E}_{p_{te}(\boldsymbol{x},y)} \left[\ell(h(\boldsymbol{x}), y) \right].
\end{aligned} \tag{2.32}$$

However, IWERM tends to produce an estimator with high variance. We can reduce the variance by flattening the importance weights, which is called Adaptive Importance-Weighted ERM (AIWERM).

Definition 2.10 (Adaptive Importance-Weighted ERM (AIWERM) [260])**.** Let $\lambda \in [0, 1]$. The Adaptive Importance-Weighted ERM (AIWERM) is defined as a modified IWERM by using $(p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x}))^{\lambda}$ as the weighting function:

$$\min_{h\in\mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\frac{p_{te}(\boldsymbol{x}_i^{tr})}{p_{tr}(\boldsymbol{x}_i^{tr})} \right)^{\lambda} \ell(h(\boldsymbol{x}_i^{tr}), y_i^{tr}). \tag{2.33}$$

Relative Importance-Weighted ERM (RIWERM), a stable version of AIWERM, has also been proposed.

Definition 2.11 (Relative Importance-Weighted ERM (RIWERM) [322])**.** Let $\lambda \in [0, 1]$. The Relative Importance-Weighted ERM (RIWERM) is defined as a modified IWERM by using $p_{te}(\boldsymbol{x})/\lambda p_{tr}(\boldsymbol{x}) + (1-\lambda)p_{te}(\boldsymbol{x})$ as the weighting function:

$$\min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \frac{p_{te}(\boldsymbol{x}_i^{tr})}{\lambda p_{tr}(\boldsymbol{x}_i^{tr}) + (1-\lambda)p_{te}(\boldsymbol{x}_i^{tr})} \ell(h(\boldsymbol{x}_i^{tr}), y_i^{tr}). \tag{2.34}$$

All of the above methods are considered as different weighting methods for each point of the training data. More generally, the method of covariate shift adaptation can be essentially rephrased as a weighting strategy for training distribution. In Chapter 3, we demonstrate that the weighting of probability distribution geometrically corresponds to the selection of curves connecting p_{tr} and p_{te}, and we geometrically generalize these covariate shift adaptation methods.

2.3 Divergences between Probability Distributions

In statistics and information theory, divergences are a powerful tool for quantifying the difference between two probability distributions [22, 71]. Divergence is a mathematical function that takes two probability distributions as input and outputs a scalar value.

Definition 2.12 (Divergence)**.** The divergence $D : \mathcal{P}_+ \times \mathcal{P} \to [0, \infty]$ is defined between two probability distributions p and q, and satisfies the following properties:

(1) $D[p\|q] \geq 0$;

(2) $D[p\|q] = 0$, if and only if $p = q$;

(3) At every p, $D[p\|p + dp]$ is a positive-definite quadratic form for infinitesimal displacements dp from p.

There are several different types of divergences, each with its own unique properties and applications.

Definition 2.13 (Kullback–Leibler divergence [160])**.** The Kullback–Leibler divergence or KL-divergence $D_{\mathrm{KL}} : \mathcal{P}_+ \times \mathcal{P} \to [0, \infty]$ is defined between two probability distributions p and q by

$$D_{\mathrm{KL}}[p\|q] \coloneqq \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}. \tag{2.35}$$

KL-divergence is a measure of the difference between two probability distributions in statistics and information theory [92, 253, 266, 333]. This is also called the relative entropy, and is known not to satisfy the axiom of distance.

Since the KL-divergence is asymmetric, several symmetrizations have been proposed in the literature [175, 199, 220].

Definition 2.14 (Jensen–Shannon divergence [175])**.** The Jensen–Shannon divergence or JS-divergence $D_{JS} : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ is defined between two probability distributions p and q by

$$\begin{aligned} D_{JS}[p\|q] &:= \frac{1}{2}\left(D_{\mathrm{KL}}\left[p\middle\|\frac{p+q}{2}\right] + D_{\mathrm{KL}}\left[q\middle\|\frac{p+q}{2}\right]\right) \\ &= \frac{1}{2}\int_{\mathcal{X}}\left(p(\boldsymbol{x})\log\frac{2p(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})} + q(\boldsymbol{x})\log\frac{2q(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})}\right)d\boldsymbol{x} \\ &= D_{JS}[q\|p]. \end{aligned} \tag{2.36}$$

The JS-divergence is a symmetrized and smoothed version of the KL-divergence, and it is bounded as

$$0 \leq D_{JS}[p\|q] \leq \log 2. \tag{2.37}$$

This property contrasts with the fact that KL-divergence is unbounded.

Definition 2.15 (Jeffreys divergence [129])**.** The Jeffreys divergence $D_J[p\|q] : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined between two probability distributions p and q by

$$D_J[p\|q] := D_{\mathrm{KL}}[p\|q] + D_{\mathrm{KL}}[q\|p]. \tag{2.38}$$

Such symmetrized KL-divergences have appeared in various literatures [23, 35, 54, 219, 222, 224, 303].

Although the KL-divergence is a core concept in statistics and information theory, it is known to have computational difficulty. To be more specific, if $q(\boldsymbol{x})$ takes a small value relative to $p(\boldsymbol{x})$, the value of $D_{\mathrm{KL}}[p(\boldsymbol{x})\|q(\boldsymbol{x})]$ may diverge to infinity. The simplest idea to avoid this is to use very small $\epsilon > 0$ and modify $D_{\mathrm{KL}}[p(\boldsymbol{x})\|q(\boldsymbol{x})]$ as follows:

$$D^{+}_{\mathrm{KL}}[p\|q] := \int_{\mathcal{X}} p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})+\epsilon}d\boldsymbol{x}.$$

However, such an extension is unnatural in the sense that $q(\boldsymbol{x}) + \epsilon$ no longer satisfies the condition for a probability measure: $\int_{\mathcal{X}} \epsilon + q(x)d\boldsymbol{x} \neq 1$.

As a more natural way to stabilize KL-divergence, the following skew divergence has been proposed:

Definition 2.16 (Skew divergence [167, 175])**.** The skew divergence $D_S^{(\lambda)}[p(\boldsymbol{x})\|q(\boldsymbol{x})] : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined between two probability distributions p and q by

$$\begin{aligned} D_S^{(\lambda)}[p\|q] &\coloneqq D_{\mathrm{KL}}[p\|(1-\lambda)p + \lambda q] \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{(1-\lambda)p(\boldsymbol{x}) + \lambda q(\boldsymbol{x})} d\boldsymbol{x}, \end{aligned} \tag{2.39}$$

where $\lambda \in [0, 1]$.

Skew divergences have been experimentally shown to perform better in applications such as natural language processing [168, 315], image recognition [49, 247] and graph analysis [5, 125]. In addition, there is research on the quantum generalization of skew divergence [17].

Skew divergence can be viewed as calculating a value by weighting one probability distribution towards another and approaching it. In this study, we introduce the generalization of skew divergence, similar to the generalization of covariate shift adaptation methods, through the identification of statistical concepts and geometric procedures.

2.4 Identification of Statistical Concepts and Geometric Objects

In this section, we provide background knowledge on the identification of statistical concepts and geometric objects. The flow of this section is as follows.

- First, we introduce the construction of a Riemannian manifold by a set of probability distributions. This is a core concept in the framework of information geometry.
- Next, we explain how weighting probability distributions and selecting curves on a statistical manifold can be identified. For this identification, we introduce the generalized means, encompassing various averaging operations, and the concept of α-geodesics, which are generalized version of straight curves on the statistical manifold. In Chapters 3 and 4, we discuss generalized algorithms and divergences using this identification.
- Finally, we briefly introduce known examples where the identification of statistical concepts and geometric objects is useful.

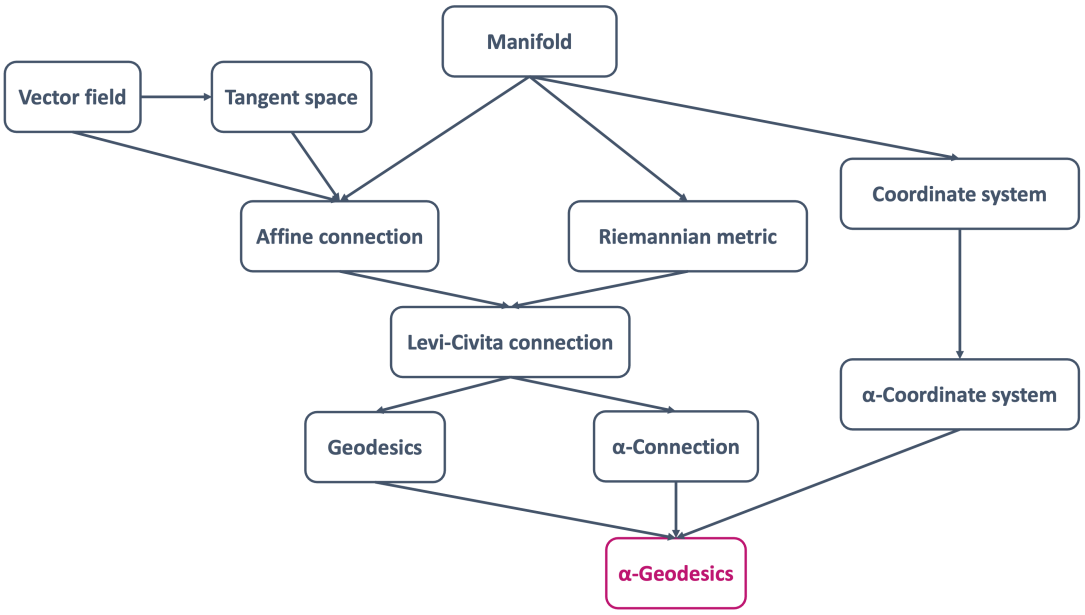

FIGURE 2.1: Definitions of geometric concepts needed in this study.

2.4.1 Statistical Manifolds

Here, we first discuss the motivation for considering statistical manifolds. Next, we introduce the definitions of the geometric concepts needed for this study.

We consider the parametric model $\mathcal{S}$ defined in Eq. (2.4). Since the probability distribution $p \in \mathcal{S}$ is uniquely determined when the parameter $\boldsymbol{\theta} \in \Theta$ is determined, the parameter space can be regarded as a coordinate system, and it is called the $\boldsymbol{\theta}$-coordinate system. A simple example is shown below.

Example 2.4. *Consider a normal distribution on $\mathbb{R}^n$ with parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$:*

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

Let $\mathcal{S}$ be a set of these distributions. The KL-divergence between two points $p_0 = p(\boldsymbol{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $p_1 = p(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ within $\mathcal{S}$ is given by

$$D_{\mathrm{KL}}[p_0 \| p_1] = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1 - \boldsymbol{I}\right) + \frac{1}{2}\log\det\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1\right).$$

In particular, when the variance is fixed as $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \sigma^2 \boldsymbol{I}$ with some fixed σ^2, we can write:

$$D_{\mathrm{KL}}[p_0 \| p_1] = \frac{1}{2\sigma^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

From the fact that KL-divergence can be written as

$$D_{KL}[p(\cdot; \boldsymbol{\theta} + d\boldsymbol{\theta}) \| p(\cdot; \boldsymbol{\theta})] = \frac{1}{2} g_{ij} d\theta^i d\theta^j + O(\|d\boldsymbol{\theta}\|^3), \tag{2.40}$$

for the Fisher metric (see Section 2.4.1.5), the $\boldsymbol{\mu}$-coordinate system can be regarded as a Cartesian coordinate system in n-dimensional Euclidean space when the variance is fixed.

From Example 2.4, we can see that the manifold constructed by a set of normal distributions with fixed variance can be regarded as a Euclidean space, but this is not necessarily true for the manifold constructed by a general probability distribution. The definition of a curve in such a non-Euclidean space depends on the coordinate system that the manifold has. In particular, a straight curve on a manifold is called a geodesic. This study considers the identification of weighting probability distributions with the selection of geodesics. Therefore, in the following we introduce the background knowledge necessary to define a generalized concept of geodesics, called α-geodesics. Figure 2.1 provides an overview of these geometric concepts.

2.4.1.1 Manifolds

A manifold is a multidimensional space that can be locally regarded as a Euclidean space. A point within a manifold can be identified by multiple parameter sets, where these parameters can be interpreted as local coordinate systems.

Let $(\mathcal{M}, d)$ be a metric space, where $\mathcal{M}$ is a set of points and d is a metric.

Definition 2.17. Let $U \subset \mathcal{M}$ be an open set. Then the pair (U, ϕ) is called a coordinate system on $\mathcal{M}$, if $\phi : U \to \phi(U) \subset \mathbb{R}^n$ is a homeomorphism of the open set U in $\mathcal{M}$ onto an open set $\phi(U)$ of $\mathbb{R}^n$. The coordinate functions f on U are defined as $x_j : U \to \mathbb{R}$, and $\phi(p) = (x_1(p), \ldots, x_n(p))$, namely $x_j = u_j \circ \phi$, where $u_j : \mathbb{R}^n \to \mathbb{R}$, $u_j(a_1, \ldots, a_n) = a_j$ is the j-th projection.

The integer n is called the dimension of the coordinate system.

Definition 2.18. An atlas $\mathcal{A}$ of dimension n associated with the metric space $(\mathcal{M}, d)$ is a collection of coordinate systems $\{(U_\alpha, \phi_\alpha)\}_\alpha$ such that

i) $U_\alpha \subset \mathcal{M}, \forall\alpha, \bigcup_\alpha U_\alpha = \mathcal{M}$. That is, U_α covers $\mathcal{M}$.

ii) If $U_\alpha \cap U_\beta \neq \emptyset$, the restriction to $\phi_\alpha(U_\alpha \cap U_\beta)$ of the map

$$F_{\alpha\beta} = \phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta) \tag{2.41}$$

is differentiable from $\mathbb{R}^n$ to $\mathbb{R}^n$.

There exist several atlases on a given metric space $\mathcal{M}$ in general. Two atlases $\mathcal{A}$ and $\mathcal{A}'$ are called compatible if their union is an atlas on $\mathcal{M}$. The set of compatible atlases with a given atlas $\mathcal{A}$ can be partially ordered by inclusion, and its maximal element is called the complete atlas $\bar{\mathcal{A}}$.

Definition 2.19. A differentiable manifold $\mathcal{M}$ is a metric space endowed with a complete atlas. The dimension n of the atlas is called the dimension of the manifold.

2.4.1.2 Tangent Space and Vector Field

Let $\mathcal{F}(\mathcal{M})$ be the set of all differentiable functions on the manifold $\mathcal{M}$. The tangent vectors on $\mathcal{M}$ is defined as follows. A tangent vector of $\mathcal{M}$ at a point $p \in \mathcal{M}$ is a function $X_p : \mathcal{F}(\mathcal{M}) \to \mathbb{R}$ such that

i) X_p is $\mathbb{R}$-linear

$$X_p(af + bg) = aX_p(f) + bX_p(g),\ \forall a, b \in \mathbb{R},\ \forall f, g \in \mathcal{F}(\mathcal{M}). \tag{2.42}$$

ii) The Leibniz rule is satisfied

$$X_p(fg) = X_p(f)g(p) + f(p)X_p(g),\ \forall f, g \in \mathcal{F}(\mathcal{M}). \tag{2.43}$$

Definition 2.20. For a differentiable curve $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ on the manifold $\mathcal{M}$, with $\gamma(0) = p$, the tangent vector

$$X_p(f) = \frac{d(f \circ \gamma)}{dt}(0), \quad \forall f \in \mathcal{F}(\mathcal{M}) \tag{2.44}$$

is called the tangent vector to $\gamma(-\epsilon, \epsilon)$ at $p = \gamma(0)$ and is denoted by $\dot{\gamma}(0)$.

Here, consider the i-th coordinate curve γ. That is, there exists a coordinate system (U, ϕ) around $p = \gamma(0)$ in which $\phi(\gamma(t)) = (x_1^{(0)}, \ldots, x_i, \ldots, x_n^{(0)})$, where $\phi(p) = (x_1^{(0)}, \ldots, x_i^{(0)}, \ldots, x_n^{(0)})$. Then, the tangent vector to γ

$$\dot{\gamma}(0) = \left.\frac{\partial}{\partial x_i}\right|_p \tag{2.45}$$

is called a coordinate tangent vector at p. This can be defined equivalently as a derivation

$$\left.\frac{\partial}{\partial x_i}\right|_p (f) = \frac{\partial(f \circ \phi^{-1})}{\partial u_i}(\phi(p)), \quad \forall f \in \mathcal{F}(\mathcal{M}), \tag{2.46}$$

where $\phi = (x_1, \ldots, x_n)$ is a coordinate system around p and $u_1, \ldots, u_n$ are the coordinate functions on $\mathbb{R}^n$.

Definition 2.21. The set of all tangent vectors at p to $\mathcal{M}$ is called the tangent space of $\mathcal{M}$ at p, and is denoted by $T_p\mathcal{M}$.

The tangent vector X_p acts on differentiable functions f on $\mathcal{M}$ as

$$X_p f = \sum_{i=1}^{n} X_i(p) \left.\frac{\partial f}{\partial x_i}\right|_p . \tag{2.47}$$

Definition 2.22. A vector field X on $\mathcal{M}$ is a smooth map X that assigns to each point $p \in \mathcal{M}$ a vector X_p in $T_p\mathcal{M}$. For any functions $f \in \mathcal{F}(\mathcal{M})$ we define the real-valued function $(Xf)_p = X_p f$.

The set of all vector fields on $\mathcal{M}$ is denoted by $\mathcal{X}(\mathcal{M})$. In a local coordinate system, a vector field is given by $X = \sum X_i \frac{\partial}{\partial x_i}$, where the components $X_i \in \mathcal{F}(\mathcal{M})$ because they are given by $X_i = X(x_i)$, where x_i is the i-th coordinate function.

2.4.1.3 Riemannian Manifolds

Let $T_p\mathcal{M}$ be the tangent space of $\mathcal{M}$ at p.

Definition 2.23. A Riemannian metric g on a differentiable manifold $\mathcal{M}$ is a symmetric, positive definite 2-covariant tensor field. A Riemannian manifold is a differentiable manifold $\mathcal{M}$ endowed with a Riemannian metric g.

A Riemannian manifold is denoted by the pair $(\mathcal{M}, g)$. The Riemannian metric g can be considered as a positive definite scalar product $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ that depends differentially on the point $p \in \mathcal{M}$. In local coordinate systems, the Riemannian metric g is written as

$$g = g_{ij} dx^i dx^j, \tag{2.48}$$

with $g_{ij} = g(\partial_i, \partial_j)$, where $\partial_i = \frac{\partial}{\partial x_i}$.

The Fisher information matrix is defined as

$$g_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial \theta^i} \log p(\boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \log p(\boldsymbol{x}; \boldsymbol{\theta}) \right] \tag{2.49}$$

$$= \int_{\mathcal{X}} \frac{\partial}{\partial \theta^i} \log p(\boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \log p(\boldsymbol{x}; \boldsymbol{\theta}) \cdot p(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x}. \tag{2.50}$$

Here, it is known that the Fisher information matrix on any statistical model is symmetric, positive definite, and non-degenerate. That is, the Fisher information matrix is a Riemannian metric.

2.4.1.4 Linear Connections and Levi-Civita Connection

A linear connection allows the differentiation of a function, a vector field, or a tensor with respect to a given field.

Definition 2.24. A linear connection ∇ on a differentiable manifold $\mathcal{M}$ is a map $\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ with the following properties:

i) $\nabla_X Y$ is $\mathcal{F}(\mathcal{M})$-linear in X;

ii) $\nabla_X Y$ is $\mathbb{R}$-linear in Y;

iii) it satisfies the Leibniz rule:

$$\nabla_X(fY) = (Xf)Y + f\nabla_X Y, \quad \forall f \in \mathcal{F}(\mathcal{M}). \tag{2.51}$$

For fixed vector fields X and Y, $\nabla_X Y$ is also a vector field on $\mathcal{M}$. In a local coordinate system $(x^1, \ldots, x^n)$ we can write

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k, \tag{2.52}$$

where Γ_{ij}^k are the coordinates of the connection with respect to the local base $\{\partial_i\}$, where $\partial_i = \frac{\partial}{\partial x^i}$. For $X = X^i \partial_i$ and $Y = Y^j \partial_j$, we have

$$\nabla_X Y = (\nabla_X Y)^k \partial_k, \tag{2.53}$$

where $(\nabla_X Y)^k = X^i(\partial_i Y^k + Y^j \Gamma_{ij}^k)$.

The differentiation of r-covariant tensor field $\mathcal{T}$ along a vector field X with respect to the linear connection ∇ as

$$(\nabla_X \mathcal{T})(Y_1, \ldots, Y_r) = X\mathcal{T}(Y_1, \ldots, Y_r) - \sum_{i=1}^{n} \mathcal{T}(Y_1, \ldots, \nabla_X Y_i, \ldots, Y_r). \tag{2.54}$$

Definition 2.25. Let g be the Riemannian metric tensor. A linear connection ∇ is called metric connection if g is parallel with respect to ∇,

$$\nabla_X g = 0, \quad \forall X \in \mathcal{X}(\mathcal{M}). \tag{2.55}$$

Equivalently,

$$Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y), \quad \forall X, Y, Z \in \mathcal{X}(\mathcal{M}). \tag{2.56}$$

A linear connection is described by two other tensors, torsion and curvature.

Definition 2.26. For a linear connection ∇, the torsion is defined as

$$\begin{aligned} &T : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M}) \\ &T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]. \end{aligned} \tag{2.57}$$

The torsion measures the noncommutativity of the derivation with respect to two vector fields. From the fact that $T(X,Y) = -T(Y,X)$, T is a 2-covariant skew-symmetric tensor. In the local coordinate system, we have

$$\begin{aligned} T_{ij} &= T(\partial_i, \partial_j) \\ &= \nabla_{\nabla_i} \partial_j - \nabla_{\partial_j} \partial_i - [\partial_i, \partial_j] \\ &= \left(\Gamma_{ij}^k - \Gamma_{ji}^k\right) \partial_k, \end{aligned}$$

and the torsion coordinates are given by $T_{ij}^k = \Gamma_{ij}^k - \Gamma_{ji}^k$. A connection ∇ is called torsion-free if $T = 0$, or $\Gamma_{ij}^k = \Gamma_{ji}^k$.

Definition 2.27. The curvature of the linear connection ∇ is given by

$$\begin{aligned} &R : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M}) \\ &R(X,Y,Z) = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \end{aligned} \tag{2.58}$$

Definition 2.28. The Ricci curvature associated with the linear connection ∇ is given by

$$\mathrm{Ric} : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{F}(\mathcal{M}), \tag{2.59}$$

$$\mathrm{Ric}(Y,Z) = \mathrm{Trace}\left(X \to R(X,Y)Z\right). \tag{2.60}$$

It is known that there exists a unique metric connection with zero torsion on the Riemannian manifold. This metric connection $\nabla^{(0)}$ is called the Levi-Civita connection of the Riemannian manifold $(\mathcal{M}, g)$. The coordinates Γ_{ij}^l of the Levi-Civita connection are called the Christoffel symbols of the second kind. The Christoffel symbols of the first kind are obtained as

$$\Gamma_{ij,k} = \Gamma_{ij}^l g_{lk}. \tag{2.61}$$

2.4.1.5 Dual Connections and Dualistic Structure

Here, we introduce the concept of dual connections. Dual connections were first introduced by A.P. Norden in affine differential geometry literature under the name of

conjugate connections [261]. They had also been independently introduced by Nagaoka and Amari in the context of information geometry [9].

Definition 2.29. Let $(\mathcal{M}, g)$ be a Riemannian manifold. Two linear connections ∇ and ∇^* are called dual, with respect to the metric g, if

$$Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y), \quad \forall X, Y, Z \in \mathcal{X}(\mathcal{M}). \tag{2.62}$$

In local coordinate system $(x^1, \ldots, x^n)$, the duality condition can be expressed as

$$\partial_{x^k} g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}, \tag{2.63}$$

where $\Gamma_{ki,h} = g_{jm}\Gamma^m_{ki}$ and $\Gamma^*_{kj,i} = g_{im}\Gamma^{*m}_{kj}$ are the coordinate components of connections ∇ and ∇^*. The triple (g, ∇, ∇^*) is called a dualistic structure on $\mathcal{M}$. A statistical manifold is a Riemannian manifold endowed with a dualistic structure, and it is a quadruple $(\mathcal{M}, g, \nabla, \nabla^*)$.

Consider the one-parameter family of connections given by the convex combination of two dual torsion-free connections ∇ and ∇^* with respect to the metric g,

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^* + \frac{1-\alpha}{2}\nabla, \quad \alpha \in \mathbb{R}. \tag{2.64}$$

The connection $\nabla^{(\alpha)}$ is called the α-connection.

2.4.1.6 Geodesics

Definition 2.30. An equation

$$\nabla^{\gamma}_{\frac{d}{dt}} \frac{d\gamma}{dt} = \mathbf{0} \tag{2.65}$$

is called the geodesic equation, and the curve satisfying this equation is called a geodesic.

Note that if $\Gamma_{ij,k} = 0 \ \forall i, j, k$, the geodesic equation is of the form $\frac{d^2\gamma_k}{dt^2} = 0$, hence the geodesic is a straight line. Here, a connection ∇ is called flat in a given coordinates if its components vanish, i.e., $\Gamma^k_{ij} = 0$.

Consider an exponential family with α connection $\nabla^{(\alpha)}$. The Christoffel symbols are

$$\Gamma^{(\alpha)}_{ij,k} = \mathbb{E}_{\boldsymbol{\theta}}\left[\left\{\partial_i\partial_j l_{\boldsymbol{\theta}} + \frac{1-\alpha}{2}(\partial_i l_{\boldsymbol{\theta}})(\partial_j l_{\boldsymbol{\theta}})\right\}(\partial_k l_{\boldsymbol{\theta}})\right], \tag{2.66}$$

and

$$\partial_i l_{\boldsymbol{\theta}} = F_i(x) - (\partial_i\psi)(\boldsymbol{\theta}), \quad (\partial_i\partial_j\psi)(\boldsymbol{\theta}). \tag{2.67}$$

So, when $\alpha = 1$, we have

$$\Gamma^{(1)}_{ij,k} = \mathbb{E}_{\boldsymbol{\theta}}[(-(\partial_i \partial_j \psi)(\boldsymbol{\theta}))(\partial_k l_{\boldsymbol{\theta}})] = 0, \tag{2.68}$$

namely, the exponential family is flat with the Fisher metric and $\alpha = 1$ connection. This implies that in exponential family, for the 1-connection $\nabla^{(1)}$ associated with the Fisher metric, the geodesic between two points correspond to natural parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is of the form $t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2$ with $t \in [0, 1]$.

2.4.2 Identification of Statistical Concepts and Geometric Objects: Known Examples

Below are some known examples of the identification of statistical concepts with geometric objects.

Example 2.5 (EM Algorithm)**.** *Let $\mathcal{N}$ be a statistical manifold, and $\mathcal{M}$ be a submanifold of $\mathcal{N}$. Given $p \in \mathcal{N}$, the point in $\mathcal{M}$ that minimizes the α-divergence (Definition 2.32) from p to $\mathcal{M}$ is called the α-projection of p to $\mathcal{M}$ [13]. Here, we consider the identification of EM algorithm (Expectation-Maximization algorithm) procedure and geometric concepts. Let $\mathcal{M}$ and $\mathcal{D}$ be a model manifold and a data manifold consisting of the distributions of candidate models and candidate data, respectively. In the E-step of the EM algorithm, we calculate the conditional expectation of log likelihood to evaluate the likelihood of a new candidate of model parameter. It is known that this procedure is equal to the α-projection with $\alpha = 1$, and it is especially called exponential projection or simply e-projection. In the M-step of the EM algorithm, we maximize the conditional expectation of log likelihood to obtain a new candidate of model parameter. It is also known that this procedure is equal to the α-projection with $\alpha = -1$, and it is especially called mixture projection or simply m-projection. Thus, the EM algorithm can be identified with the iterative projection of the two types between the model manifold and the data manifold.*

Example 2.6 (Natural Gradient)**.** *In the Euclidean space, the direction of the steepest change of a function $L(\boldsymbol{\theta})$ is given by the gradient*

$$\nabla L(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

However, in a Riemannian manifold, it is known that we need to modify that gradient as $G^{-1}(\boldsymbol{\theta})\nabla L(\boldsymbol{\theta})$, where $G(\boldsymbol{\theta}) = (g_{ij})$ is a Riemannian metric tensor. This modified gradient is called the natural gradient [11]. Since a Riemannian metric tensor is equal to the Fisher information matrix in a Riemannian manifold, we can see that the natural

gradient characterizes the steepest direction on the statistical manifold by the inverse of the Fisher information matrix.

Example 2.7 (α-Parallel Prior)**.** *In Bayesian statistics, the situation in which there is no prior information about the parameter $\boldsymbol{\theta}$ can be expressed by considering an uninformative prior distribution $\pi(\boldsymbol{\theta})$. In particular, the uniform distribution is one of the commonly used uninformed prior distributions. For example, when $\boldsymbol{x} = (x_1, \dots, x_n)^\top \sim \mathrm{Ber}(\boldsymbol{\theta})$, then $\pi(\boldsymbol{\theta}) = U(0,1)$ is one of the candidates. However, if we consider a parameter transformation such as $\boldsymbol{\eta} = \boldsymbol{\theta}^2$, then $\pi(\boldsymbol{\eta}) = \frac{1}{2\sqrt{\boldsymbol{\eta}}} = \mathrm{Be}(0.5, 1)$. This means that we know nothing about $\boldsymbol{\theta}$ but we know something about $\boldsymbol{\eta}$. One such candidate that addresses this is Jeffreys' uninformative prior, which is known as a prior invariant to parameter transformations.*

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det G(\boldsymbol{\theta})},$$

where $G(\boldsymbol{\theta})$ is the Fisher information matrix. Here, it is known that the determinant of the Fisher information matrix corresponds to a volume element parallel to the Levi-Civita connection in a Riemannian manifold. Takeuchi and Amari [282] and Matsuzoe et al. [194] have proposed a generalized α-parallel prior, which corresponds to Jeffreys' uninformative prior distribution with $\alpha = 0$.

As in the examples above, the framework of information geometry may be used to identify statistical concepts with geometric objects. In this study, we follow this framework and consider generalizations of the algorithm and divergence by identifying the weighting of the probability distribution with the selection of the curve.

2.4.3 Identification of Weighting Probability Distributions and Selecting Curves

For the generalization of Chapters 3 and 4, we prepare the following generalized mean.

Definition 2.31 (f-interpolation [145])**.** For any $a, b, \in \mathbb{R}$, some $\lambda \in [0,1]$ and some $\alpha \in \mathbb{R}$, we define f-interpolation as

$$m_f^{(\lambda,\alpha)}(a,b) = f_\alpha^{-1}\Big\{(1-\lambda) f_\alpha(a) + \lambda f_\alpha(b)\Big\}, \tag{2.69}$$

where

$$f_\alpha(a) = \begin{cases} a^{\frac{1-\alpha}{2}} & (\alpha \neq 1) \\ \log a & (\alpha = 1) \end{cases} \tag{2.70}$$

is the function that defines the f-mean [108].

We can easily see that this family includes various known weighted means including the e-mixture and m-mixture for $\alpha = \pm 1$ in the literature of information geometry [13]:

$$\begin{aligned} m_f^{(\lambda,1)}(a,b) &= \exp\{(1-\lambda)\log a + \lambda \log b\}, \\ m_f^{(\lambda,-1)}(a,b) &= (1-\lambda)a + \lambda b, \\ m_f^{(\lambda,0)}(a,b) &= \left((1-\lambda)\sqrt{a} + \lambda\sqrt{b}\right)^2, \\ m_f^{(\lambda,3)}(a,b) &= \frac{1}{(1-\lambda)\frac{1}{a} + \lambda\frac{1}{b}}. \end{aligned}$$

Also, for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ $(d > 0)$, we write

$$\boldsymbol{m} = m_f^{(\lambda,\alpha)}(\boldsymbol{u}, \boldsymbol{v}), \text{where} \quad \boldsymbol{m}_i = m_f^{(\lambda,\alpha)}(\boldsymbol{u}_i, \boldsymbol{v}_i).$$

We can relate the above function to the following generalization of probability distribution weighting.

Example 2.8. *For two probability distributions $p, q \in \mathcal{P}$ and $\lambda \in [0,1]$, consider the weighting function $w(\boldsymbol{x}) = (q(\boldsymbol{x})/p(\boldsymbol{x}))^\lambda$ as*

$$\begin{aligned} p(\boldsymbol{x})w(\boldsymbol{x}) &= p(\boldsymbol{x})\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right)^\lambda \\ \log\{p(\boldsymbol{x})w(\boldsymbol{x})\} &= \lambda\left(\log q(\boldsymbol{x}) - \log p(\boldsymbol{x})\right) + \log p(\boldsymbol{x}) \\ &= (1-\lambda)\log p(\boldsymbol{x}) + \lambda \log q(\boldsymbol{x}) \\ p(\boldsymbol{x})w(\boldsymbol{x}) &= \exp\{(1-\lambda)\log p(\boldsymbol{x}) + \lambda \log q(\boldsymbol{x})\} \\ &= m_f^{(\lambda,1)}(p(\boldsymbol{x}), q(\boldsymbol{x})). \end{aligned}$$

On the other hand, the divergence and the coordinate system it induces, which is often used in the framework of information geometry, are given as follows.

Definition 2.32 (α-divergence [12])**.** Let α be a real parameter. The α-divergence between two probability distributions p and q is defined as

$$D_\alpha[p\|q] = \frac{4}{1-\alpha^2}\Big(1 - \int_{\mathcal{X}} p(\boldsymbol{x})^{\frac{1-\alpha}{2}} q(\boldsymbol{x})^{\frac{1+\alpha}{2}}\Big) d\boldsymbol{x}. \tag{2.71}$$

Definition 2.33 (α-geodesic [13])**.** The α-geodesic connecting two probability distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is defined as

$$\begin{aligned} r_i(\lambda) &= c(t) f_\alpha^{-1}\Big\{(1-\lambda) f_\alpha(p(x_i)) + \lambda f_\alpha(q(x_i))\Big\}, \\ c(\lambda) &= \left(\sum_{i=1} r_i(\lambda)\right)^{-1}, \end{aligned} \tag{2.72}$$

where x_i is the i-th element of $\boldsymbol{x}$.

From Definitions 2.31 and 2.33, we see that f-interpoloation is the unnormalized version of the α-geodesic. We write $\tilde{m}_f^{(\lambda,\alpha)}$ for a suitably normalized f-interpolation. The important properties of α-geodesics are

- the α-geodesic is a geodesic in the α-coordinate system derived from α-divergence,
- the $-\alpha$-geodesic is linear in the $-\alpha$-representation.

Chapters 3 and 4 show how the identification of f-interpolation and α-geodesic generalizes covariate shift adaptation methods and skew divergences, respectively.

Chapter 3

Information Geometrically Generalized Covariate Shift Adaptation

In this chapter, we show that the well-known family of covariate shift adaptation methods is unified in the framework of information geometry. Furthermore, we show that parameter search for geometrically generalized covariate shift adaptation method can be achieved efficiently. Numerical experiments show that our generalization can achieve better performance than the existing methods it encompasses.

3.1 Recapitulate the Covariate Shift Assumption

Here, we consider the covariate shift assumption in Definition 2.5, a type of distribution shift. Recall that for training and test distributions p_{tr} and p_{te}, the covariate shift assumes

$$p_{tr}(\boldsymbol{x}) \neq p_{te}(\boldsymbol{x}),$$
$$p_{tr}(y|\boldsymbol{x}) = p_{te}(y|\boldsymbol{x}) = p(y|\boldsymbol{x}).$$

Since $p_{tr}(y|\boldsymbol{x}) = p_{te}(y|\boldsymbol{x}) = p(y|\boldsymbol{x})$ from the covariate shift assumption, what we are interested in is the model manifold $(\mathcal{M}, g(\boldsymbol{\theta}))$ to which the marginal distribution $p(\boldsymbol{x}; \boldsymbol{\theta})$

belongs:

$$\mathcal{M} = \Big\{ p(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \Big\}. \tag{3.1}$$

Here, $p_{tr}(\boldsymbol{x}; \boldsymbol{\theta}), p_{te}(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathcal{M}$. We note that elements in $\mathcal{M}$ is specified by its parameter $\boldsymbol{\theta}$ and we identify the parameter vector $\boldsymbol{\theta}$ to the density function $p(\boldsymbol{x}; \boldsymbol{\theta})$ if necessary. In this study, we assume that $\mathcal{M}$ is an exponential family (2.5). In the exponential family, the natural parameter $\boldsymbol{\theta}$ forms the affine coordinate system, i.e.,

$$\boldsymbol{\theta}(t) = (1-t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2 \quad (\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta,\ \forall t \in [0,1]) \tag{3.2}$$

is a geodesic on $\mathcal{M}$. As a dual coordinate of $\boldsymbol{\theta}$, which is a coordinate system induced by the dual connection (Definition 2.29), the expectation parameter $\boldsymbol{\eta}$ is defined by the Legendre transformation

$$\begin{aligned} \boldsymbol{\eta} =& \nabla \psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = \nabla \varphi(\boldsymbol{\eta}), \\ \text{where } \ \varphi(\boldsymbol{\eta}) =& \max_{\boldsymbol{\theta}'} \Big\{ \boldsymbol{\theta}' \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}') \Big\}. \end{aligned}$$

Existing weights for covariate shift adaptation are geometrically characterized, and then a generalized weight function is designed based on this geometric formulation.

3.2 Information Geometrically Generalized IWERM

In order to derive a generalized covariate shift adaptation method, we utilize the generalized mean in Definition 2.31. Using this function, we generalize the existing methods of covariate shift adaptation.

Lemma 3.1 (f-representation of AIWERM)**.** *The marginal positive measures generated by the weighting of AIWERM can be expressed by using the f-interpolation function as*

$$p_A^{(\lambda)}(\boldsymbol{x}) = m_f^{(\lambda,1)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})). \tag{3.3}$$

Proof. From Eq. (2.33), we consider its expectation as

$$\begin{aligned} \hat{h} &= \min_{h \in \mathcal{H}} \int_{\mathcal{X} \times \mathcal{Y}} \Big(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} \Big)^{\lambda} \ell(h(\boldsymbol{x}), y) p_{tr}(\boldsymbol{x}, y) d\boldsymbol{x} dy \\ &= \min_{h \in \mathcal{H}} \int_{\mathcal{X} \times \mathcal{Y}} \ell(h(\boldsymbol{x}), y) p_A^{(\lambda)}(\boldsymbol{x}) p_{tr}(y|\boldsymbol{x}) d\boldsymbol{x} dy. \end{aligned}$$

Here,

$$\begin{aligned} p_A^{(\lambda)}(\boldsymbol{x}) &= \Big(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\Big)^{\lambda} p_{tr}(\boldsymbol{x}) \\ \log p_A^{(\lambda)}(\boldsymbol{x}) &= \lambda(\log p_{te}(\boldsymbol{x}) - \log p_{tr}(\boldsymbol{x})) + \log p_{tr}(\boldsymbol{x}) \\ &= (1-\lambda)\log p_{tr}(\boldsymbol{x}) + \lambda \log p_{te}(\boldsymbol{x}) \\ p_A^{(\lambda)}(\boldsymbol{x}) &= \exp\{(1-\lambda)\log p_{tr}(\boldsymbol{x}) + \lambda \log p_{te}(\boldsymbol{x})\} \\ &= m_f^{(\lambda,1)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})). \end{aligned}$$

□

Lemma 3.2 (f-representation of RIWERM)**.** *The marginal positive measures generated by the weighting of RIWERM can be expressed by using the f-interpolation function as*

$$p_R^{(\lambda)}(\boldsymbol{x}) = m_f^{(\lambda,3)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})). \tag{3.4}$$

Proof. From Eq. (2.34),

$$\begin{aligned} p_R^{(\lambda)}(\boldsymbol{x}) &= \frac{p_{te}(\boldsymbol{x})p_{tr}(\boldsymbol{x})}{\lambda p_{tr}(\boldsymbol{x}) + (1-\lambda)p_{te}(\boldsymbol{x})} \\ &= \frac{1}{\lambda \frac{1}{p_{te}(\boldsymbol{x})} + (1-\lambda)\frac{1}{p_{tr}(\boldsymbol{x})}} \\ &= m_f^{(\lambda,3)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})). \end{aligned}$$

□

From the above discussion, the following generalized method of covariate shift adaptation is derived using the f-representation. For $\lambda \in [0,1]$ and $\alpha \in \mathbb{R}$, AIWERM and RIWERM is generalized as

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} w^{(\lambda,\alpha)}(\boldsymbol{x})\ell(h(\boldsymbol{x}), y)p_{tr}(\boldsymbol{x}, y)d\boldsymbol{x}dy, \tag{3.5}$$

where

$$w^{(\lambda,\alpha)}(\boldsymbol{x}) = \frac{m_f^{(\lambda,\alpha)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x}))}{p_{tr}(\boldsymbol{x})}. \tag{3.6}$$

Indeed, let h_A be a hypothesis generated by AIWERM. From Lemma 3.1, we can write

$$\begin{aligned}\hat{h}_A &= \min_{h\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}), y) p_A^{(\lambda)}(\boldsymbol{x}) p_{tr}(y|\boldsymbol{x}) d\boldsymbol{x}dy \\ &= \min_{h\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}), y) m_f^{(\lambda,1)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})) p_{tr}(y|\boldsymbol{x}) d\boldsymbol{x}dy \\ &= \min_{h\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}), y) \frac{m_f^{(\lambda,1)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x}))}{p_{tr}(\boldsymbol{x})} p_{tr}(\boldsymbol{x}, y) d\boldsymbol{x}dy. \end{aligned} \tag{3.7}$$

From Lemma 3.2, we also have

$$\hat{h}_R = \min_{h\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}), y) \frac{m_f^{(\lambda,3)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x}))}{p_{tr}(\boldsymbol{x})} p_{tr}(\boldsymbol{x}, y) d\boldsymbol{x}dy. \tag{3.8}$$

Then, we consider

$$\hat{h} = \min_{h\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} w^{(\lambda,\alpha)}(\boldsymbol{x}) \ell(h(\boldsymbol{x}), y) p_{tr}(\boldsymbol{x}, y) d\boldsymbol{x}dy, \tag{3.9}$$

where

$$w^{(\lambda,\alpha)}(\boldsymbol{x}) = \frac{m_f^{(\lambda,\alpha)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x}))}{p_{tr}(\boldsymbol{x})}. \tag{3.10}$$

Then, we can see that AIWERM is a special case when $\alpha = 1$ and RIWERM is a special case when $\alpha = 3$.

From Definition 2.31, we can confirm that

$$\begin{aligned} m_f^{(0,\alpha)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})) &= p_{tr}(\boldsymbol{x}), \\ m_f^{(1,\alpha)}(p_{tr}(\boldsymbol{x}), p_{te}(\boldsymbol{x})) &= p_{te}(\boldsymbol{x}), \end{aligned}$$

for all $\alpha \in \mathbb{R}$, and this means that we can obtain the set of all curves that connect $p_{tr}(\boldsymbol{x})$ and $p_{te}(\boldsymbol{x})$. We note that Zhang et al. [339] proposed a method based on basis expansion to estimate a flexible importance weight. It is similar to our proposal in the sense that improves the degree of freedom for designing the importance weight. However, our method considers the parametric form of weight, which enables us to achieve information geometric insight.

In many studies of covariate shift problems using the density ratio weighting including Yamada et al. [320], the direct estimation of the density ratio is often employed [279].

Our proposed weight function in (3.6) is also represented as density ratio:

$$w^{(\lambda,\alpha)}(\boldsymbol{x}) = \frac{\left[(1-\lambda)p_{tr}(\boldsymbol{x})^{\frac{1-\alpha}{2}} + \lambda p_{te}(\boldsymbol{x})^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}}}{p_{tr}(\boldsymbol{x})}$$
$$= \left[1-\lambda+\lambda\left(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\right)^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}}, \quad (\alpha \neq 1).$$

It is then also possible to apply the direct estimation of the density ratio using, for example kernel expansion. In our implementation, we simply use the given density ratio $p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$ by the construction of the training and the test datasets as explained in Section 3.4.1. In the practical application of the proposed method in which the generative processes of the covariates of training and test data are unknown, direct density estimation would be a promising approach.

3.2.1 Geometric Bias

AIWERM and RIWERM connect two distributions p_{tr} and p_{te} in different ways. Statistical bias and variance of IWERM, AIWERM, and RIWERM are discussed in the respective papers. In this subsection, we study the geometric bias of these methods to have a deeper understanding of these methods from the geometric viewpoint. The proposed generalization of IWERM is independent of a specific parametrization of density functions. In this subsection, for theoretical treatment, the exponential model manifold which contains $p_{tr}(\boldsymbol{x};\boldsymbol{\theta})$ and $p_{te}(\boldsymbol{x};\boldsymbol{\theta})$ are considered, hence geodesics can be described by a linear combination of parameters as explained in Appendix C. With this assumption, specifying λ and α is equivalent to selecting a point on the geodesic connecting p_{tr} and p_{te}.

Let γ_c be the geodesic connecting two distributions parameterized by $\boldsymbol{\theta}_{tr}$ and $\boldsymbol{\theta}_{te}$. Now, we define two types of geometric biases to characterize the dispersion of $\boldsymbol{\theta}_{tr}$ from $\boldsymbol{\theta}_{te}$ with respect to the direction along the α-geodesic and to the direction orthogonal to the α-geodesic.

Definition 3.3 (Geodesic bias and curvature bias)**.** We write the unit vector along the α-geodesic direction as e_1 and any unit vector in the orthogonal direction to e_1 as e_2. Then, we define two notion the biases relative to the test distribution due to weighting:

i) geodesic bias: $b_g = (1-\lambda)e_1$,

ii) curvature bias: $b_c = (1-\lambda)\mathrm{Trace}_g(\mathrm{Ric})e_2$,

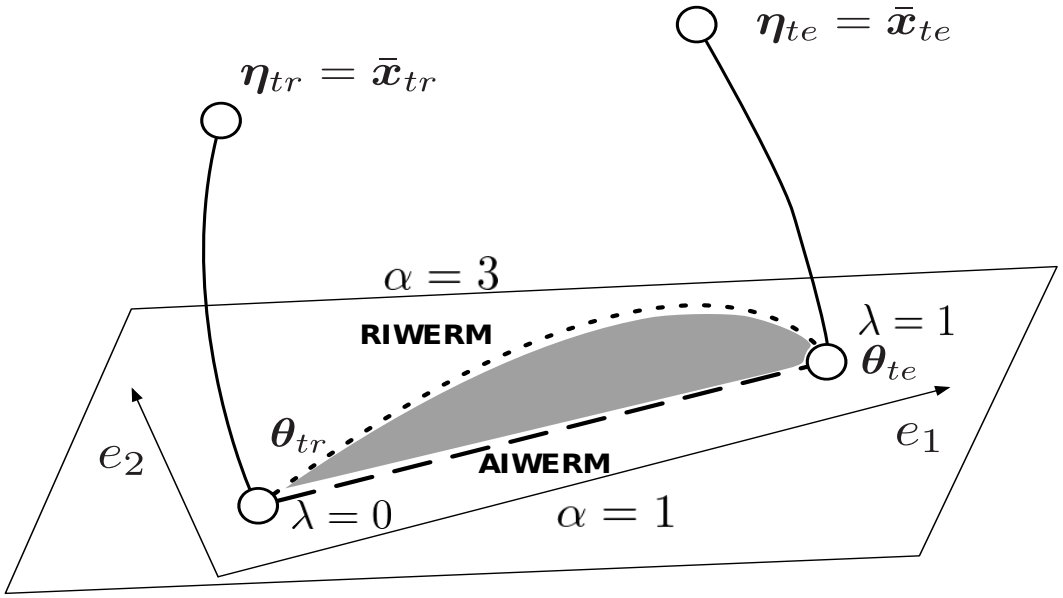

FIGURE 3.1: Geometry of covariate shift adaptation methods on the statistical manifold. In the $\boldsymbol{\theta}$-coordinate system, the dashed line corresponds to AIWERM and the dotted line corresponds to RIWERM. We write a unit vector along the α-geodesic direction as e_1 and any unit vector in the orthogonal direction to e_1 as e_2. Here, $\lambda = 0$ and $\lambda = 1$ correspond to θ_{tr} (ERM) and θ_{te} (IWERM), respectively, and $\alpha = 1$ and $\alpha = 3$ correspond to the AIWERM and RIWERM curves, respectively, in the figure.

where Trace_g is the trace operation on the metric tensor g as

$$\mathrm{Trace}_g(\boldsymbol{A}) = \sum_{ij} A^{ij} g_{ij}, \tag{3.11}$$

for some $\boldsymbol{A} = (A^{ij})$, and Ric is the Ricci curvature of the curve connecting the two points generated by the weighting:

$$\mathrm{Ric} = R_{ikj} d\boldsymbol{\theta}^i \otimes d\boldsymbol{\theta}^j. \tag{3.12}$$

Here, R_{ikj} is the Riemannian curvature tensor.

For more detail on the geometric concepts, see textbooks on Riemannian manifolds [164, 166] and Appendix C. This definition of geometric biases is consistent with the fact that IWERM, which corresponds to $\lambda = 1$, leads to an unbiased estimator of the risk in the test dataset.

Proposition 3.4. *For AIWERM, the geometric bias $b_A(\lambda)$ is given by*

$$b_A(\lambda) = (1 - \lambda) e_1. \tag{3.13}$$

For RIWERM, the geometric bias $b_R(\lambda)$ is given by

$$b_R(\lambda) = (1-\lambda)\Big\{e_1 + \mathrm{Trace}_g\Big(-4\Lambda_{ikj} d\boldsymbol{\theta}^i \otimes d\boldsymbol{\theta}^j\Big) e_2\Big\}. \tag{3.14}$$

Here, Λ is a tensor that depends on the connection.

Proof. Let

$$\boldsymbol{\theta}^{(\lambda,\alpha)} = m_f^{(\lambda,\alpha)}(\boldsymbol{\theta}_{tr}, \boldsymbol{\theta}_{te}), \tag{3.15}$$

and let $R^{(\alpha)}$ be the Riemann curvature tensor with respect to the α-connection $\nabla^{(\alpha)}$.

We define the relative curvature tensor as

$$R^{(\alpha,\beta)}(X,Y,Z) = \Big[\nabla_X^{(\alpha)}, \nabla_Y^{(\beta)}\Big] Z - \nabla_{[X,Y]}^{(\alpha)} Z \tag{3.16}$$

and the difference tensor as

$$K(X,Y) = \nabla_X^* Y - \nabla_X Y. \tag{3.17}$$

For any $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, we have

$$\begin{aligned}
\nabla_X^{(\alpha)} \nabla_Y^{(\beta)} Z &= \Big(\frac{1+\alpha}{2}\nabla_X^* + \frac{1-\alpha}{2}\nabla_X\Big)\Big(\frac{1+\beta}{2}\nabla_Y^* + \frac{1-\beta}{2}\nabla_Y\Big) Z \\
&= \frac{(1+\alpha)(1+\beta)}{4}\nabla_X^*\nabla_Y^* Z + \frac{(1+\alpha)(1-\beta)}{4}\nabla_X^*\nabla_Y Z \\
&\quad + \frac{(1-\alpha)(1+\beta)}{4}\nabla_X\nabla_Y^* Z + \frac{(1-\alpha)(1-\beta)}{4}\nabla_X\nabla_Y Z. \qquad (3.18) \\
\nabla_Y^{(\beta)} \nabla_X^{(\alpha)} Z &= \Big(\frac{1+\beta}{2}\nabla_Y^* + \frac{1-\beta}{2}\nabla_Y\Big)\Big(\frac{1+\alpha}{2}\nabla_X^* + \frac{1-\alpha}{2}\nabla_X\Big) Z \\
&= \frac{(1+\alpha)(1+\beta)}{4}\nabla_Y^*\nabla_X^* Z + \frac{(1-\alpha)(1+\beta)}{4}\nabla_Y^*\nabla_X Z \\
&\quad + \frac{(1+\alpha)(1-\beta)}{4}\nabla_Y\nabla_X^* Z + \frac{(1-\alpha)(1-\beta)}{4}\nabla_Y\nabla_X Z. \qquad (3.19) \\
\nabla_{[X,Y]}^{(\alpha)} Z &= \frac{1+\alpha}{2}\nabla_{[X,Y]}^* Z + \frac{1-\alpha}{2}\nabla_{[X,Y]} Z. \qquad (3.20)
\end{aligned}$$

Then

$$\begin{aligned}
R^{(\alpha,\beta)}(X,Y,Z) &= \nabla_X^{(\alpha)}\nabla_Y^{(\beta)} Z - \nabla_X^{(\beta)}\nabla_Y^{(\alpha)} Z - \nabla_{[X,Y]}^{(\alpha)} Z \\
&= \frac{(1+\alpha)(1+\beta)}{4}(\nabla_X^*\nabla_Y^* - \nabla_Y^*\nabla_X^*) Z \\
&\quad + \frac{(1+\alpha)(1-\beta)}{4}(\nabla_X^*\nabla_Y - \nabla_Y\nabla_X^*) Z \\
&\quad + \frac{(1-\alpha)(1+\beta)}{4}(\nabla_X\nabla_Y^* - \nabla_Y^*\nabla_X) Z \\
&\quad + \frac{(1-\alpha)(1-\beta)}{4}(\nabla_X\nabla_Y - \nabla_Y\nabla_X) Z
\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{1+\alpha}{2}\nabla^*_{[X,Y]}Z - \frac{1-\alpha}{2}\nabla_{[X,Y]}Z \\
&= \frac{(1+\alpha)(1+\beta)}{4}\Big\{R^*(X,Y,Z) + \nabla^*_{[X,Y]}Z\Big\} \\
&\quad + \frac{(1+\alpha)(1-\beta)}{4}\Big\{R^{(1,-1)}(X,Y,Z) + \nabla^*_{[X,Y]}Z\Big\} \\
&\quad + \frac{(1-\alpha)(1+\beta)}{4}\Big\{R^{(-1,1)}(X,Y,Z) + \nabla_{[X,Y]}Z\Big\} \\
&\quad + \frac{(1-\alpha)(1-\beta)}{4}\Big\{R^{(-1,-1)}(X,Y,Z) + \nabla^*_{[X,Y]}Z\Big\} \\
&\quad - \frac{1+\alpha}{2}\nabla^*_{[X,Y]}Z - \frac{1-\alpha}{2}\nabla_{[X,Y]}Z \qquad (3.21)
\end{aligned}
$$

$$
\begin{aligned}
4R^{(\alpha,\beta)} &= (1+\alpha)(1+\beta)R^* + (1-\alpha)(1-\beta)R \\
&\quad + (1+\alpha)(1-\beta)R^{(1,-1)} + (1-\alpha)(1+\beta)R^{(-1,1)}. \qquad (3.22)
\end{aligned}
$$

We also have

$$
\begin{aligned}
K^{(\alpha,\beta)}(X,Y) &= \nabla^{(\beta)}_X Y - \nabla^{(\alpha)}_X Y \\
&= \Big\{\frac{1+\beta}{2}\nabla^*_X Y + \frac{1-\beta}{2}\nabla_X Y\Big\} - \Big\{\frac{1+\alpha}{2}\nabla^*_X Y + \frac{1-\alpha}{2}\nabla_X Y\Big\} \\
&= \frac{\beta-\alpha}{2}\nabla^*_X Y - \frac{\beta-\alpha}{2}\nabla_X Y \\
&= \frac{\beta-\alpha}{2}K(X,Y), \qquad (3.23)
\end{aligned}
$$

$$
\begin{aligned}
K^{(\alpha,\beta)}\Big(X, K^{(\alpha,\beta)}(Y,Z)\Big) &= \frac{\beta-\alpha}{2}K\Big(X, K^{(\alpha,\beta)}(Y,Z)\Big) \\
&= \frac{(\beta-\alpha)^2}{4}K\Big(X, K(Y,Z)\Big). \qquad (3.24)
\end{aligned}
$$

Combining them, the following relations hold:

$$
\begin{aligned}
K^{(\beta,\alpha)}\Big(X, K^{(\beta,\alpha)}(Y,Z)\Big) &= K^{(\beta,\alpha)}\Big(X, \nabla^{(\alpha)}_Y Z - \nabla^{(\beta)}_Y Z\Big) \qquad (3.25) \\
&= K^{(\beta,\alpha)}\Big(X, \nabla^{(\alpha)}_Y Z\Big) - K^{(\beta,\alpha)}\Big(X, \nabla^{(\beta)}_Y Z\Big) \\
&= \nabla^{(\alpha)}_X\nabla^{(\alpha)}_Y Z - \nabla^{(\beta)}_X\nabla^{(\alpha)}_Y Z - \nabla^{(\alpha)}_X\nabla^{(\beta)}_Y Z + \nabla^{(\beta)}_X\nabla^{(\beta)}_Y Z, \qquad (3.26)
\end{aligned}
$$

$$
\frac{(\alpha-\beta)^2}{4}K\Big(X, K(Y,Z)\Big) = \nabla^{(\alpha)}_X\nabla^{(\alpha)}_Y Z - \nabla^{(\beta)}_X\nabla^{(\alpha)}_Y Z - \nabla^{(\alpha)}_X\nabla^{(\beta)}_Y Z + \nabla^{(\beta)}_X\nabla^{(\beta)}_Y Z. \qquad (3.27)
$$

Swapping X and Y, we have

$$
\begin{aligned}
&\frac{(\alpha-\beta)^2}{4}\Big\{K\Big(X, K(Y,Z)\Big) - K\Big(Y, K(X,Z)\Big)\Big\} \\
&= R^{(\alpha)}(X,Y,Z) + R^{(\beta)}(X,Y,Z) \\
&\quad - \Big\{\Big[\nabla^{(\alpha)}_X, \nabla^{(\beta)}_Y\Big]Z - \nabla^{(\alpha)}_{[X,Y]}Z\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&- \Big\{\Big[\nabla^{(\beta)}_X, \nabla^{(\alpha)}_Y\Big]Z - \nabla^{(\beta)}_{[X,Y]}Z\Big\} \\
&= R^{(\alpha)}(X,Y,Z) + R^{(\beta)}(X,Y,Z) - R^{(\alpha,\beta)}(X,Y,Z) - R^{(\beta,\alpha)}(X,Y,Z). \qquad (3.28)
\end{aligned}
$$

Making $\alpha = \beta$, we have

$$
\begin{aligned}
4R^{(\alpha)} &= (1+\alpha)^2 R^* + (1-\alpha)^2 R + (1-\alpha^2)R^{(1,-1)} + (1-\alpha^2)R^{(-1,1)} \\
&= (1+\alpha^2)R^* + (1-\alpha)^2 R + (1-\alpha^2)\Big(R^{(1,-1)} + R^{(-1,1)}\Big). \qquad (3.29)
\end{aligned}
$$

Making $\alpha = 1$ and $\beta = -1$, we also have

$$
\begin{aligned}
R^{(1,-1)}(X,Y,Z) + R^{(-1,1)}(X,Y,Z) &= R^*(X,Y,Z) + R(X,Y,Z) \\
&\quad - \Big\{K\Big(X, K(Y,Z)\Big) - K\Big(Y, K(X,Z)\Big)\Big\}. \qquad (3.30)
\end{aligned}
$$

From Eq. (3.29) and (3.30), we obtain

$$
\begin{aligned}
4R^{(\alpha)} &= (1+\alpha)^2 R^*(X,Y,Z) + (1-\alpha)^2 R(X,Y,Z) \\
&\quad + (1-\alpha^2)R^*(X,Y,Z) \\
&\quad + (1-\alpha^2)\Big\{K\Big(X, K(Y,Z)\Big) - K\Big(Y, K(X,Z)\Big)\Big\} \\
&= 2(1+\alpha)R^*(X,Y,Z) + 2(1-\alpha)R(X,Y,Z) \\
&\quad + (1-\alpha^2)\Big\{K\Big(Y, K(X,Z)\Big) - K\Big(X, K(Y,Z)\Big)\Big\}. \qquad (3.31)
\end{aligned}
$$

Since the exponential family is dually flat, that is $R = 0$ and $R^* = 0$, and the Riemann curvature tensor with respect to $\nabla^{(\alpha)}$ is

$$
R^{(\alpha)}(X,Y,Z) = \frac{1-\alpha^2}{4}\Lambda, \qquad (3.32)
$$

$$
\Lambda = \Big(K(Y, K(X,Z)) - K(X, K(Y,Z))\Big). \qquad (3.33)
$$

Then, the geometric bias vector of $\boldsymbol{\theta}^{(\lambda,\alpha)}$ is

$$
b(\alpha,\lambda) = (1-\lambda)\Big\{e_1 + \mathrm{Trace}_g\Big(\frac{1-\alpha^2}{2}\Lambda_{ikj} d\boldsymbol{\theta}^i \otimes d\boldsymbol{\theta}^j\Big)e_2\Big\}, \qquad (3.34)
$$

where Trace_g is the trace operation on the metric tensor g, and Λ_{ikj} is the element of Λ in Eq. (3.33). Since AIWERM and RIWERM are two special cases for $\alpha = 1$ and $\alpha = 3$, we have

$$
b(1,\lambda) = (1-\lambda)e_1, \qquad (3.35)
$$

$$b(3,\lambda) = (1-\lambda)\Big\{e_1 + \mathrm{Trace}_g\Big(- 4\Lambda_{ikj} d\boldsymbol{\theta}^i \otimes d\boldsymbol{\theta}^j\Big) e_2\Big\}. \tag{3.36}$$

□

Intuitively, the geometric bias reveals in which direction the two parameters are misaligned. IWERM, which corresponds to AIWERM and RIWERM with $\lambda = 1$, is optimal when the sample size is large enough, but in real problems with limited sample size, it is often desirable to adopt a point between $\boldsymbol{\theta}_{tr}$ and $\boldsymbol{\theta}_{te}$. AIWERM and RIWERM consider distinct curves and specify a point on them by the parameter λ. Our geometric analysis revealed that these curves are included in the set of curves represented by dual f-representation of the parameter coordinate system, and the geometric biases of these particular cases (AIWERM and RIWERM) are identified. The results presented in this subsection do not claim the superiority of a particular method and are of importance in their own right as a geometric analysis of the covariate shift method.

3.3 Hyper-parameter optimization of the generalized IWERM

The existing covariate shift adaptation methods described above can be regarded as having determined a good "weighting direction" in some sense in advance and then the "weighting magnitude" is adjusted according to the parameter λ. This approach is very convenient in terms of computational efficiency since the only optimized parameter is $\lambda \in [0, 1]$.

However, geometrically, these methods only consider certain curves on the manifold as candidate solutions, as can be seen from Figure 3.1, which means that the solution space is very small.

Our information geometrically generalized IWERM (IGIWERM) can handle all curves $\gamma_\alpha(\lambda)$ that connect $p_{tr}(\boldsymbol{x})$ and $p_{te}(\boldsymbol{x})$, by adding only one parameter. For example, by setting $\alpha \in [1, 3]$, the shaded area in Figure 3.1 can be used as the solution space. Although a good choice of this parameter pair is expected to yield a good solution, the problem of how to determine λ and α remains.

3.3.1 Information criterion

When the predictive model is of a simple parametric form, information criterion derived in [260] is available (see appendix of [260] for the proof.)

Algorithm 1 Bayesian optimization for IGIWERM

Input: range I_α of parameter α, acquisition function $a(\lambda, \alpha|D)$, target function $L(h; \lambda, \alpha)$, initial points D_{init} compose of a set of parameters $\Xi = \{(\lambda, \alpha)\}$ and corresponding values of the target function

Output: (λ^*, α^*) that minimizes $\min_{h\in\mathcal{H}} L(h; \lambda, \alpha)$

Initialize $D = D_{init}$

while Not converge **do**

$(\hat{\lambda}, \hat{\alpha}) = \underset{\lambda\in[0,1],\alpha\in I_\alpha}{\arg\min}\ a(\lambda, \alpha|D), \quad \Xi = \Xi \cup \{(\hat{\lambda}, \hat{\alpha})\}$

$\hat{e} = L(h; \hat{\lambda}, \hat{\alpha}), \quad D = D \cup \{(\hat{\lambda}, \hat{\alpha}, \hat{e})\}$

end while

$(\lambda^*, \alpha^*) = \underset{(\lambda,\alpha)\in\Xi}{\arg\min} \{\min_{h\in\mathcal{H}} L(h; \lambda, \alpha)\}$

Corollary 3.5 (Information criterion for IGIWERM)**.** *Let the information criterion for MWLE (Maximum Weighted Likelihood Estimator) of IWERM be*

$$IC_{GW} \coloneqq -2L_1(\hat{\boldsymbol{\theta}}) + 2tr(J_w H_w^{-1}), \tag{3.37}$$

where $L_1(\boldsymbol{\theta}) = \sum_{i=1}^{n_{tr}} dr(\boldsymbol{x}_i^{tr}) \log p(y_i^{tr}|\boldsymbol{x}_i^{tr}, \boldsymbol{\theta})$, $dr(\boldsymbol{x}) = \frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}$ *and*

$$J_w = -\mathbb{E}_{p_{tr}}\left[dr(\boldsymbol{x})\frac{\partial \log p(y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_w^*} \times \frac{\partial\Big(\frac{m_f^{\lambda,\alpha}(p_{tr}(x), p_{te}(\boldsymbol{x}))}{p_{tr}(\boldsymbol{x})} \log p(y|\boldsymbol{x}, \boldsymbol{\theta})\Big)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_w^*}\right],$$

$$H_w = \mathbb{E}_{p_{tr}}\left[\frac{\partial^2\Big(\frac{m_f^{\lambda,\alpha}(p_{tr}(x), p_{te}(\boldsymbol{x}))}{p_{tr}(\boldsymbol{x})} \log p(y|\boldsymbol{x}, \boldsymbol{\theta})\Big)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right].$$

The second term is the bias correction term for the weighted likelihood. Here, $\boldsymbol{\theta}_w^*$ *is the minimizer of the weighted empirical risk. The matrices* J_w *and* H_w *may be replaced by their consistent estimates. Then,* $IC_{GW}/2n$ *is an unbiased estimator of the expected loss up to* $O(n^{-1})$ *term:*

$$\mathbb{E}_{p_{tr}}\Big[IC_{GW}/2n\Big] = \mathbb{E}_{p_{tr}}\Big[\ell_1(\hat{\boldsymbol{\theta}}_w)\Big] + o(n^{-1}). \tag{3.38}$$

3.3.2 Bayesian optimization

The information criterion in Corollary 3.5 does not work for complicated nonparametric models. As a method that can be applied in general situations, we consider using Bayesian optimization [85, 265] to find optimal weighting by IGIWERM. Bayesian optimization assumes that the target function is drawn from a prior distribution over functions, typically a Gaussian process, updating a posterior as we observe the target

function value in new places. We use the validation loss as the target function:

$$L(h;\lambda,\alpha) = \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} \frac{p_{te}(\boldsymbol{x}_i^{val})}{p_{tr}(\boldsymbol{x}_i^{val})} \ell(h_{\lambda,\alpha}(\boldsymbol{x}_i^{val}), y_i^{val}), \tag{3.39}$$

where n_{val} is the validation sample size and $h_{\lambda,\alpha}$ is given by IWERM with λ and α. Note that validation sample is a hold-out subset of the training data. This validation procedure is based on the importance weighted cross-validation used in [277]. In Bayesian optimization, an acquisition function $a(\lambda, \alpha|D)$ is used for measuring the goodness of candidate point (λ, α) based on the current dataset D. As the acquisition function, we adopt the expected improvement [133, 202]. In this strategy, we choose the next query point which has the highest expected improvement over the current minimum target value (Algorithm 1). In our numerical experiments, we set the range of parameter α as $I_\alpha = [1, 3]$ in Algorithm 1. In the expected improvement strategy, the $(t+1)$st point $(\lambda_{t+1}, \alpha_{t+1})$ is selected according to the following equation.

$$(\lambda_{t+1}, \alpha_{t+1}) = \arg\min_{(\lambda,\alpha)} \mathbb{E}\Big[\max\Big(0, h_{t+1}(\lambda,\alpha) - L(\lambda^\dagger, \alpha^\dagger)\Big) \Big| D_t \Big],$$

where $L(\lambda^\dagger, \alpha^\dagger)$ is the maximum value of empirical risk that has been encountered so far, $h_{t+1}(\lambda, \alpha)$ is the posterior mean of the surrogate at the $(t+1)$st step and $D_t = \{(\lambda,\alpha)_i, L(\lambda_i, \alpha_i)\}_{i=1}^t$. This equation for Gaussian process surrogate has an analytical expression:

$$\begin{aligned} a_{EI}(\lambda,\alpha) &= (\mu_t(\lambda,\alpha) - L(\lambda^\dagger,\alpha^\dagger))\Phi(Z) + \sigma_t(\lambda,\alpha)\phi(Z), \\ Z &= \frac{\mu_t(\lambda,\alpha) - L(\lambda^\dagger,\alpha^\dagger)}{\sigma_t(\lambda,\alpha)}, \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are normal cumulative and density functions, respectively, and μ_t and σ_t are mean and standard deviation of $\{(\lambda,\alpha)_i\}_{i=1}^t$.

3.3.3 Learning guarantee

Generalization bounds of the weighted maximum likelihood estimator for the target domain are derived in [62], and our weight function is compatible with their bound. The gap between the expected (with respect to test distribution) loss $\mathcal{R}(h)$ and empirical

risk $L(h;\lambda,\alpha)$ is bounded as

$$|\mathcal{R}(h) - L(h;\lambda,\alpha)| \leq \left|\mathbb{E}_{p_{tr}}\left[\left\{\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} - w^{(\lambda,\alpha)}(\boldsymbol{x})\right\}\right]\right| \ell(h(\boldsymbol{x}, y(\boldsymbol{x})))$$
$$+2^{5/4}\max\left(\sqrt{\mathbb{E}_{p_{tr}}\left[w^{(\lambda,\alpha)}(\boldsymbol{x})^2\ell(h(\boldsymbol{x}), y(\boldsymbol{x}))^2\right]}, \sqrt{\mathbb{E}_{\hat{p}_{tr}}\left[w^{(\lambda,\alpha)}(\boldsymbol{x})^2\ell(h(\boldsymbol{x}), y(\boldsymbol{x}))^2\right]}\right)$$
$$\times\left(\frac{p\log\frac{2n_{tr}}{p} + \log\frac{4}{\delta}}{n_{tr}}\right)^{\frac{3}{8}}. \tag{3.40}$$

In the above inequality, p is the pseudo-dimension of the function class

$$\{w^{\lambda,\alpha}(\boldsymbol{x})\ell(h(\boldsymbol{x}), y(\boldsymbol{x}))|h \in \mathcal{H}\},$$

where $y(\boldsymbol{x})$ is the ground truth function of connecting $\boldsymbol{x}$ and y as $y = y(\boldsymbol{x})$. The first term of the r.h.s. of the above inequality is the bias introduced by using $w^{\lambda,\alpha}$ instead of a standard density ratio, and the second term reflects the variance. It is worth mentioning that the term $\mathbb{E}_{p_{tr}}(w^{(\lambda,\alpha)}(\boldsymbol{x}))^2\ell^2(h(\boldsymbol{x}, y(\boldsymbol{x})))$ is further bounded by

$$d_2(p_{te}||p_{tr}) = \int_{x\in\mathcal{X}} \frac{p_{te}^2(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} d\boldsymbol{x}.$$

Then, the weight defined by Eq. (3.6) is bounded when $\alpha \neq 1$ and achieves a standard rate $O(n_{tr}^{-1/2})$. When $\alpha = 1$, the weight is unbounded and its rate is $O(n_{tr}^{-3/8})$. Details are shown in Appendix A.1.

3.4 Numerical Experiments

In this section, we present experimental results of domain adaptation problems under covariate shift using both synthetic and real data[1]. Since the main purpose of the experiments is to see the effect of our generalization of the importance weighted ERM and comparison to the proposed and conventional IWERM methods, in all experiments, we assume that the density ratio p_{te}/p_{tr} is known as detailed in Section 3.4.1.

3.4.1 Induction of Covariate Shift

Since each dataset is composed of data points generated from independent and identical distributions, we need to artificially induce covariate shifts. We induce the covariate shift as follows [61]:

1. As a preprocessing step, we perform Z-score standardization on all input data.

[1]Source code to reproduce the results is available from https://github.com/nocotan/IGIWERM

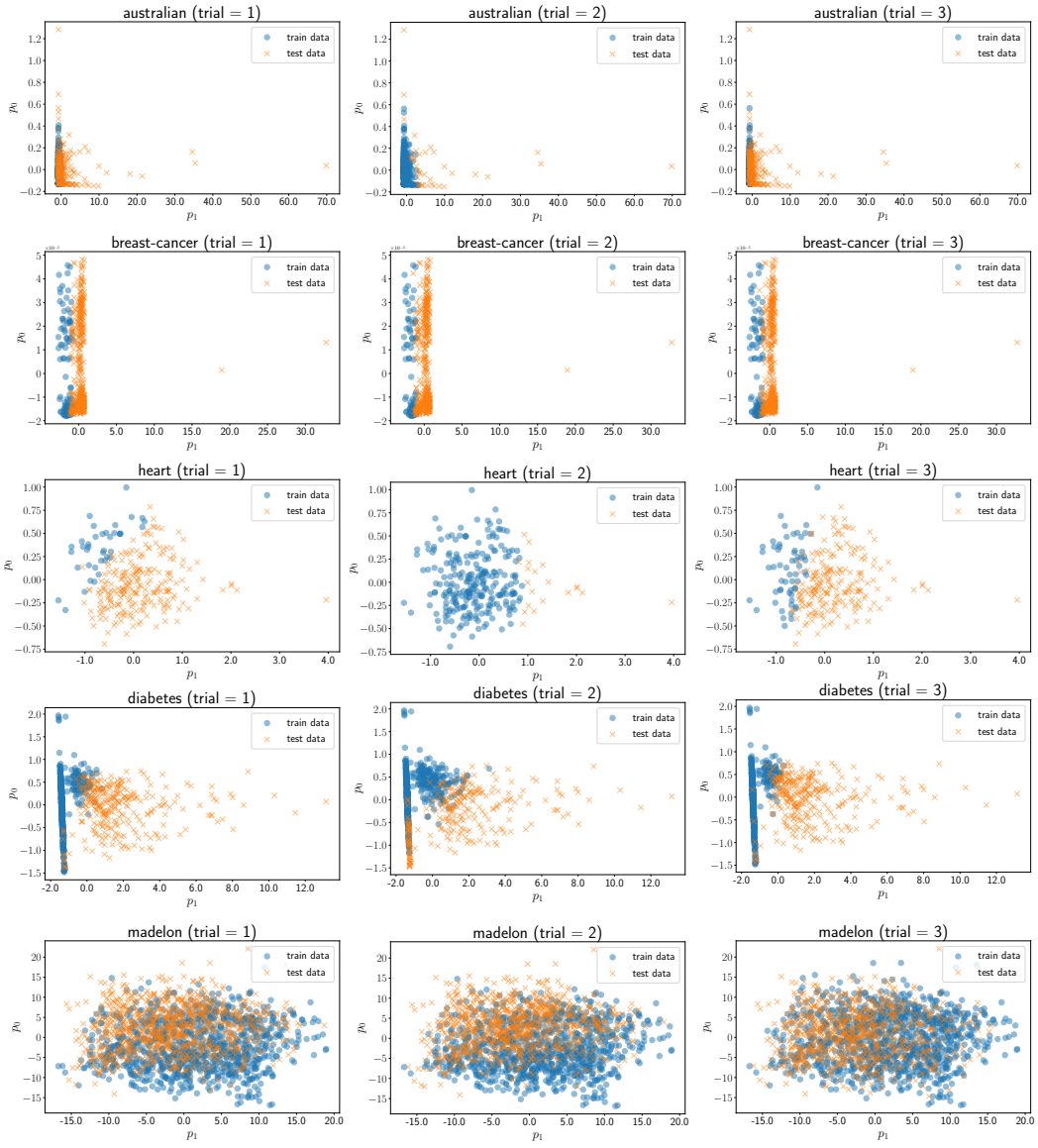

FIGURE 3.2: Plot of covariate shifts using the method of Cortes et al. [61]. Each dataset is included in LIBSVM and mapped to two dimensions by PCA.

2. Then, an example $(\boldsymbol{x}, y)$ is assigned to the training dataset with probability

$$\exp(v)/(1+\exp(v)),$$

and to the test dataset with probability

$$1/(1+\exp(v)),$$

where $v = 16\boldsymbol{w}^T\boldsymbol{x}/\sigma$ with σ being the standard deviation of $\boldsymbol{w}^T\boldsymbol{x}$ determined by using the given dataset, and $\boldsymbol{w} \in \mathbb{R}^d$ is a given projection vector. Here, the projection vector $\boldsymbol{w}$ is given randomly for each experimental process.

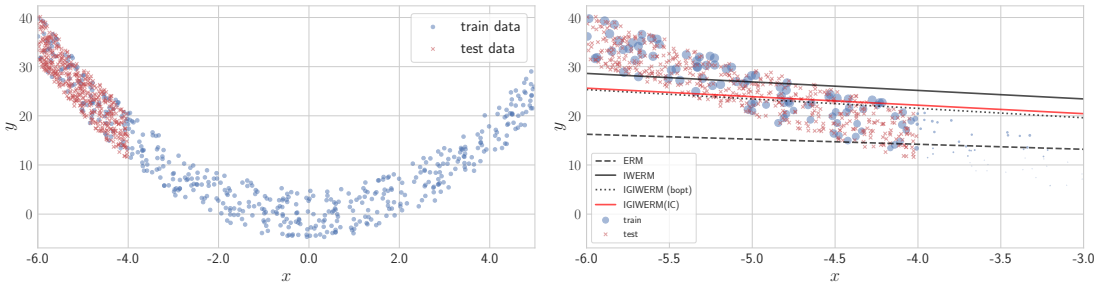

FIGURE 3.3: Left: generated data from $y = x^2 + \varepsilon$. We see that $p_{tr}(x)$ and $p_{te}(x)$ are different. Right: results of fitting by ERM, IWERM, and IGIWERM.

TABLE 3.1: Mean squared errors of covariate shift adaptation methods in regression problems over 10 trials. Here, IGIWERM (bopt) is the Bayesian optimization based, and IGIWERM (IC) is the information criterion-based strategy.

Weighting strategy	MSE
ERM	160.19($\pm$4.25)
IWERM	33.76($\pm$3.82)
AIWERM	31.14($\pm$2.97)
RIWERM	30.03($\pm$2.74)
IGIWERM (bopt)	**28.89**($\pm$**2.42**)
IGIWERM (IC)	**28.38**($\pm$**2.12**)

By this construction of the training and test datasets, the density ratio p_{te}/p_{tr} are explicitly determined as

$$\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} = \frac{1}{\exp(16\boldsymbol{w}^T\boldsymbol{x}/\sigma)},$$

where the projection vector $\boldsymbol{w} \in \mathbb{R}^d$ is given. Although density ratio estimation could be employed in our experiments, we assume that the distribution is known in order to compare the performance of the proposed method without relying on the accuracy of the density or density ratio estimation.

Figure 3.2 shows a plot by PCA of each dataset split into the training set and test set. This figure shows that we are able to induce a covariate shift by partitioning the dataset.

3.4.2 Illustrative Example in Regression

Here, we predict the response $y \in \mathbb{R}$ using ordinary linear regression: $y = \beta_0 + \beta_1 x + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(a, b)$ denotes the normal distribution with mean a and b. In the numerical example below, we assume the true $p(y|x)$ given as $y = x^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 5)$. The $p_{tr}(x)$ and $p_{te}(x)$ of the covariate x are $\mathcal{U}(-6, 6)$ and $\mathcal{U}(-6, -4)$, respectively. The training sample size is $n_{tr} = 1000$ and the test sample size is $n_{te} = 300$. The left-hand side of Fig. 3.3 shows the data to be generated. We can see that $p_{tr}(x) \neq p_{te}(x)$.

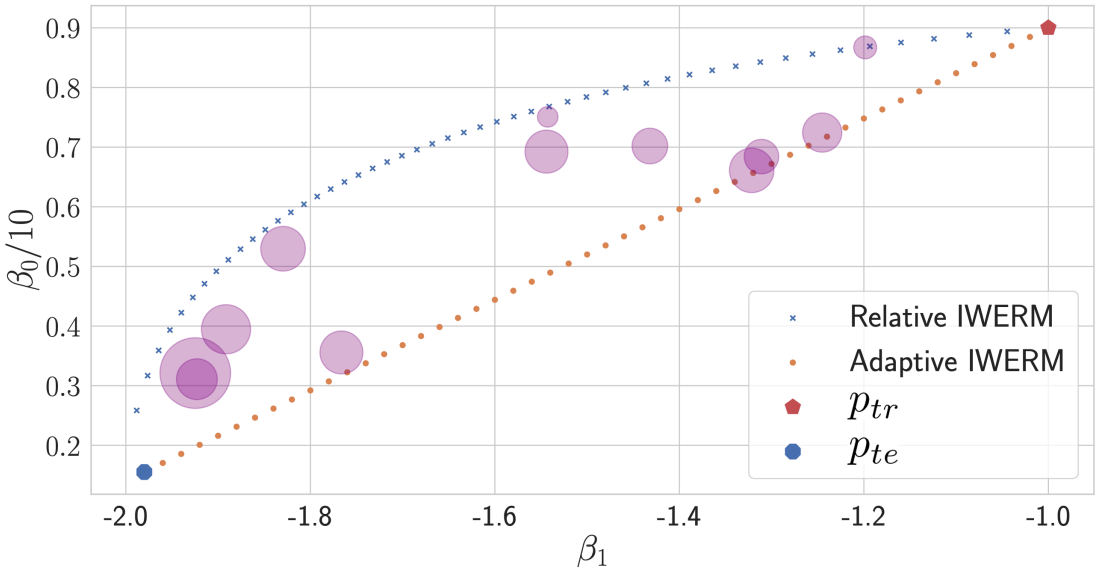

FIGURE 3.4: Bayesian optimization for IGIWERM. The coordinates of the purple circles are the parameters explored by Bayesian optimization, and the size of the purple circles indicates the goodness of the parameters (inverse of the MSE).

The right panel of Fig. 3.3 shows the results of fitting by unweighted ERM, IWERM, and IGIWERM. Here, the parameters of IGIWERM are explored by using Algorithm 1, as shown in Fig. 3.4. In this figure we plot some of the points obtained during the Bayesian optimization search process. The coordinates of the purple circles are the parameters explored by Bayesian optimization, and the radius of the purple circles is proportional to the goodness $r(\boldsymbol{\beta})$ of the parameters (inverse of the MSE):

$$r(\boldsymbol{\beta}) = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i, \boldsymbol{\beta}))^2 \right)^{-1}. \tag{3.41}$$

By choosing the size $r(\boldsymbol{\beta})$ of the plot for each point in this manner, the better-evaluated parameters can be plotted in larger circles. From this figure, it can be seen that our generalized weighting is not restricted to lying just on two curves corresponding to AIWERM and RIWERM.

For the normal linear regression, the information criterion (3.37) is calculated from

$$IC_{GW}(\lambda, \alpha) = \frac{1}{2} \sum_{i=1}^{n_{tr}} dr(\boldsymbol{x}_i^{tr}) \left\{ \frac{\hat{\epsilon}_1^2}{\hat{\sigma}^2 + \log(2\pi\hat{\sigma}^2)} \right\} + \sum_{i=1}^{n_{tr}} dr(\boldsymbol{x}_i^{tr}) \left\{ \frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2} \hat{h}_i + \frac{m_f^{\lambda,\alpha}(p_{tr}(\boldsymbol{x}^{tr}), p_{te}(\boldsymbol{x}^{tr}))}{2\hat{c}_w p_{tr}(\boldsymbol{x}^{tr})} \left(\frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2} - 1 \right)^2 \right\}.$$

TABLE 3.2: Mean misclassification rates averaged over 10 trails on LIBSVM benchmark datasets. The numbers in the brackets are the standard deviations. For the methods with (optimal), the optimal parameters for the test data are obtained by linear search.

Dataset	#features	#data	unweighted	IWERM	AIWERM (optimal)	RIWERM (optimal)	ours
australian	14	690	33.46(±23.65)	22.13(±3.37)	21.98(±3.36)	21.73(±3.82)	**18.85**(±**3.99**)
breast-cancer	10	683	38.28(±10.98)	41.23(±15.39)	36.41(±9.68)	36.13(±10.81)	**31.65**(±**8.49**)
heart	13	270	45.17(±6.98)	39.94(±8.55)	39.76(±8.49)	39.76(±8.92)	**35.37**(±**6.84**)
diabetes	8	768	33.19(±5.69)	37.22(±6.63)	33.11(±6.45)	33.38(±5.74)	**32.83**(±**5.62**)
madelon	500	2,000	47.78(±1.53)	47.28(±2.20)	47.10(±2.13)	47.12(±1.65)	**46.56**(±**2.12**)

Here,

$$\hat{c}_w = \sum_{i=1}^{n} \frac{m_f^{\lambda,\alpha}(p_{tr}(x_i^{tr}), p_{te}(x_i^{tr}))}{p_{tr}(x_i^{tr})}, \tag{3.42}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{m_f^{\lambda,\alpha}(p_{tr}(x_i^{tr}), p_{te}(x_i^{tr}))}{p_{tr}(x_i^{tr})} \hat{\epsilon}_i^2 / \hat{c}_w, \tag{3.43}$$

and $\hat{\epsilon}_i$ is the residual. Table 3.1 shows that the IGIWERM outperforms existing methods.

3.4.3 Experiments on binary classification

We show the results of our experiments on the LIBSVM dataset[2]. In the experiments, we randomly generate a mapping vector $\boldsymbol{w}$ for each trial and perform 10 trials for each dataset. We use SVM with the Radial Basis Function (RBF) kernel as the base classifier. In this experiment, the parameters λ of AIWERM and RIWERM are chosen optimally by linear search using the test data. The experimental results on benchmark datasets are summarized in Table 3.2. The table shows that the proposed IGIWERM outperforms the conventional methods even when the parameters of those methods are optimized by using the test dataset.

3.4.4 Experimental results on LIBSVM dataset

We show that for the LIBSVM dataset, IGIWERM is effective even for multiple models. Table 3.4 shows the results for each model. We use the scikit-learn [234] implementation of the models, and the hyperparameters of each model are the default values of this library. Figure 3.6 also shows the relationship between the two parameters of IGIWERM and the errors that can be achieved. For this visualization, we explore the parameter pairs by grid search and evaluate their performance at that time. From this figure, it is seen that the best performance is often achieved when $\alpha \neq 1$ and $\alpha \neq 3$, showing the sub-optimality of conventional methods.

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

TABLE 3.3: Mean misclassification rates averaged over 10 trails on scikit-learn [234] multi-class classification benchmark datasets. The numbers in the brackets are the standard deviations. For the methods with (optimal), the optimal parameters for the test data are obtained by the linear search. The lowest misclassification rates among the five methods are shown in bold.

Dataset	model	unweighted	IWERM	AIWERM (optimal)	RIWERM (optimal)	ours
digits	Logistic Regression	6.92(±2.25)	7.10(±1.76)	6.90(±1.81)	6.89(±1.80)	**6.79(±1.82)**
	SVM	4.21(±2.04)	3.94(±1.14)	4.20(±1.22)	4.02(±1.18)	**3.88(±1.11)**
	AdaBoost	67.98(±12.35)	71.77(±6.25)	71.26(±8.21)	70.47(±6.46)	**65.20(±8.20)**
	Naive Bayes	19.07(±2.45)	18.68(±2.61)	19.31(±2.64)	18.78(±2.68)	**18.58(±2.66)**
	Random Forest	6.85(±2.40)	6.30(±1.80)	6.23(±1.53)	6.27(±1.64)	**6.17(±1.90)**
iris	Logistic Regression	54.69(±20.51)	36.49(±22.39)	45.78(±18.09)	35.37(±23.19)	**28.89(±20.54)**
	SVM	55.16(±22.60)	36.65(±22.56)	33.21(±20.25)	30.46(±21.14)	**29.04(±20.02)**
	AdaBoost	27.19(±22.56)	26.00(±22.98)	26.00(±22.98)	26.00(±22.98)	**19.62(±21.36)**
	Naive Bayes	35.88(±23.24)	35.97(±26.79)	37.99(±23.55)	33.84(±25.64)	**27.52(±21.00)**
	Random Forest	26.16(±22.86)	32.17(±21.21)	26.00(±22.98)	28.47(±22.63)	**22.77(±21.74)**
covtype	Logistic Regression	45.36(±13.08)	32.04(±5.833)	30.62(±4.64)	25.20(±8.99)	**19.99(±5.56)**
	SVM	–	–	–	–	–
	AdaBoost	47.53(±14.51)	25.55(±12.27)	25.47(±14.14)	27.86(±10.37)	**18.96(±7.29)**
	Naive Bayes	41.13(±15.65)	30.66(±15.09)	28.48(±15.66)	27.20(±15.79)	**19.64(±15.62)**
	Random Forest	23.51(±3.31)	18.18(±2.01)	17.28(±2.05)	17.13(±2.26)	**16.42(±2.08)**

TABLE 3.4: Mean misclassification rates averaged over 10 trails on LIBSVM benchmark datasets. The numbers in the brackets are the standard deviations. For the methods with (optimal), the optimal parameters for the test data are obtained by the linear search. The lowest misclassification rates among the five methods are shown in bold.

Dataset	model	unweighted	IWERM	AIWERM (optimal)	RIWERM (optimal)	ours
australian	Logistic Regression	22.76(±9.53)	22.75(±7.73)	22.19(±9.35)	22.39(±7.76)	**21.47(±8.91)**
	SVM	33.46(±23.65)	22.13(±3.37)	21.98(±3.36)	21.73(±3.82)	**18.85(±3.99)**
	AdaBoost	10.20(±5.09)	16.35(±10.67)	11.23(±3.49)	15.47(±2.72)	**9.15(±3.93)**
	Naive Bayes	17.04(±6.02)	14.92(±5.96)	16.33(±6.07)	15.49(±6.07)	**14.74(±5.84)**
	Random Forest	8.89(±3.72)	8.64(±3.20)	**8.29(±3.21)**	8.38(±3.47)	8.61(±3.73)
breast-cancer	Logistic Regression	32.32(±3.08)	32.87(±3.56)	32.32(±3.08)	32.32(±3.08)	**32.26(±3.04)**
	SVM	38.28(±10.98)	41.23(±15.39)	36.41(±9.68)	36.13(±10.81)	**31.65(±8.49)**
	AdaBoost	5.09(±1.65)	5.63(±1.38)	6.01(±1.13)	5.95(±1.50)	**4.84(±1.50)**
	Naive Bayes	11.27(±4.09)	19.65(±15.14)	10.36(±5.14)	18.00(±14.24)	**10.03(±3.53)**
	Random Forest	3.32(±1.23)	3.19(±1.10)	3.19(±1.13)	3.18(±1.04)	**3.13(±1.04)**
heart	Logistic Regression	39.68(±7.90)	40.55(±9.10)	39.96(±7.40)	39.94(±6.93)	**36.56(±8.29)**
	SVM	45.17(±6.98)	39.94(±8.55)	39.76(±8.49)	39.74(±8.92)	**35.37(±6.84)**
	AdaBoost	30.87(±12.04)	29.24(±6.47)	29.37(±12.19)	31.27(±8.37)	**26.96(±13.12)**
	Naive Bayes	22.79(±6.17)	24.78(±7.99)	22.87(±6.02)	24.58(±7.98)	**21.97(±6.41)**
	Random Forest	20.87(±6.64)	20.98(±6.62	20.96(±6.67)	21.96(±6.70)	**19.95(±6.61)**
diabetes	Logistic Regression	37.62(±4.35)	40.22(±4.10)	38.38(±3.85)	40.11(±3.74)	**36.86(±4.81)**
	SVM	33.19(±5.69)	37.22(±6.63)	33.11(±6.45)	33.38(±5.74)	**32.83(±5.62)**
	AdaBoost	37.69(±4.28)	40.13(±5.28)	40.76(±4.31)	41.26(±5.16)	**33.45(±4.35)**
	Naive Bayes	39.29(±3.98)	39.21(±3.18)	39.26(±2.97)	39.35(±2.85)	**38.10(±4.02)**
	Random Forest	30.09(±3.03)	30.90(±3.52)	31.07(±3.10)	30.51(±3.67)	**29.46(±2.99)**
madelon	Logistic Regression	47.31(±1.80)	47.80(±1.57)	47.16(±1.68)	46.81(±1.56)	**46.31(±1.69)**
	SVM	47.78(±1.53)	47.28(±2.20)	47.10(±2.13)	47.12(±1.65)	**46.56(±2.12)**
	AdaBoost	42.92(±1.40)	42.91(±1.68)	43.36(±1.81)	42.90(±1.40)	**40.64(±7.32)**
	Naive Bayes	41.90(±1.05)	41.43(±8.76)	41.79(±8.05)	41.62(±7.43)	**41.03(±8.32)**
	Random Forest	35.90(±0.83)	35.42(±1.75)	35.13(±1.30)	**34.79(±1.75)**	35.56(±2.03)

3.4.5 Experimental results on regression

In this section, we present experimental results for the regression problem. All the datasets used in this experiment are available in the scikit-learn [234] dataset collection. We use SVR with the Radial Basis Function (RBF) kernel as the base regressor. Table 3.5 shows the results of this experiment. In this experiment, we use MSE as a metric,

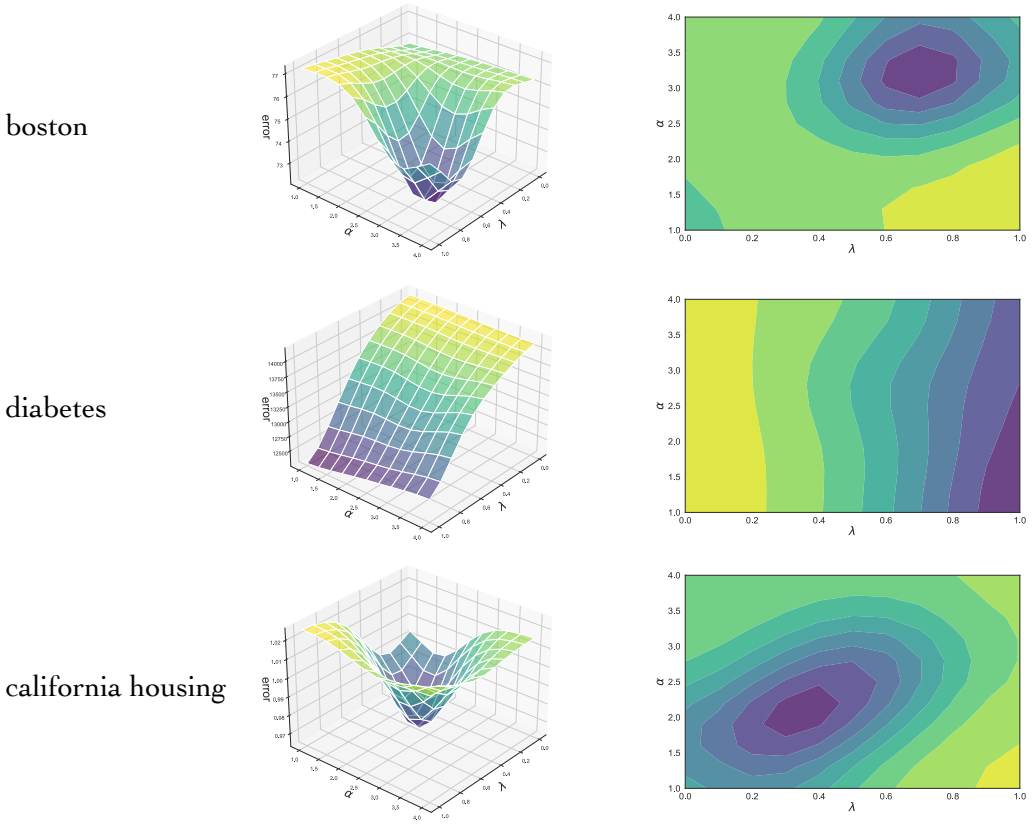

FIGURE 3.5: Visualization of grid search for α and λ on scikit-learn regression dataset. For the sake of visualization, we apply a moving average.

TABLE 3.5: Mean squared errors averaged over 10 trails on scikit-learn [234] regression benchmark datasets. The numbers in the brackets are the standard deviations. For the methods with (optimal), the optimal parameters for the test data are obtained by the linear search. The lowest mean squared errors are shown in bold.

Dataset	#features	#data	unweighted	IWERM	AIWERM (optimal)	RIWERM (optimal)	ours
boston	13	506	83.22(±5.72)	69.87(±2.31)	69.68(±1.46)	69.96(±1.84)	**68.36**(±**1.20**)
diabetes	10	442	0.049(±0.007)	0.0501(±0.009)	0.049(±0.008)	0.049(±0.009)	**0.048**(±**0.007**)
california housing	8	20,640	1.432(±0.095)	1.3214(±0.345)	1.260(±0.125)	1.261(±0.086)	**1.232**(±**0.095**)

and this table shows that IGIWERM is superior to existing covariate shift adaptation methods. Figure 3.5 shows the relationship between the two parameters of IGIWERM and the mean squared errors that can be achieved. This figure shows that, as in the case of binary classification, the optimal parameters do not necessarily match those of existing methods.

3.4.6 Experimental results on multi-class classification

We also introduce the additional experimental results for the multi-class classification problem. All the datasets used in this experiment are available in the scikit-learn [234] dataset collection. We also use the scikit-learn [234] implementation of the models, and

Table 3.6: Computational cost of ERM and IGIWERM.

Method	Computation time [sec]
ERM	1.130(±0.238)
IGIWERM	9.887(±0.845)

the hyperparameters of each model are the default values of this library. We note that the number of training sample for `covtype` is so large hence the results with SVM for this dataset are omitted. Table 3.3 shows the experimental results, and we see that our proposed generalization outperforms existing methods. Figure 3.7 shows the relationship between the two parameters of IGIWERM and the errors that can be achieved. This figure also shows the sub-optimality of conventional methods.

3.4.7 Computational Cost

Here, we investigate the computational cost of our IGIWERM. The experimental setup is the same as in Section 3.4.2. The mean and standard deviation of the computation time obtained in the 10 trials are shown in Table 3.6. From this table, we can see that our IGIWERM takes constant times longer to compute than the vanilla ERM.

3.5 Role of Parameters

In this section, we discuss the role of the parameter pair (α, λ) introduced by our generalization. Geometrically, the parameter α determines the shape of the curve connecting the training and test distributions, and the parameter λ determines the position on the curve.

Recall that, from Eq. (3.40), the gap between the expected error and empirical error is upper bounded by two terms of the form

$$C_1 \left(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} - w^{(\lambda,\alpha)}(\boldsymbol{x}) \right) + C_2 \left(\frac{1}{n_{tr}} \right), \tag{3.44}$$

for the IGIWERM under the covariate shift assumption. Here $C_1(\cdot)$ and $C_2(\cdot)$ are terms that depend on $\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})} - w^{(\lambda,\alpha)}(\boldsymbol{x})$ and $\frac{1}{n_{tr}}$, respectively. The first term depends on the difference between the density ratio and the generalized weighting, while the second term depends on the size of the training sample. Here, the generalized weighting $w^{(\lambda,\alpha)}(\boldsymbol{x})$ is equal to the density ratio when $\lambda = 1$. Therefore, when the training sample size is large, the effect of the second term vanishes asymptotically, and λ should be close to 1. In fact, a major cause of the instability of importance-weighted ERM is the difficulty of

computing the density ratio. For example, considering the case with small $\lambda = 1$ and 0.1 in AIWERM with $p_{tr}(\boldsymbol{x}) = 0.001$ and $p_{te}(\boldsymbol{x}) = 1$, we have $p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x}) = 1000$ and $(p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x}))^{0.1} \simeq 2$, which shows that the weighting can be smoothed with a small λ.

We have not obtained any analytical conclusions regarding the parameter α. However, since RIWERM by [322] achieves stabilization by using a large α, we can conjecture that a large α induces stability.

Finally, we consider the behavior of the parameters from a comparison of the two optimization strategies: information criterion and Bayesian optimization. Figure 3.8 shows the results of the comparison of the two optimization strategies in the experimental setup of Section 3.4.2, with the dimensionality of the input vectors set to 1, 2, and 4. From this figure, we can see the following.

- There is a correlation between parameter pairs obtained by two optimization strategies.
- The correlation with respect to the parameter λ is strong when the dimensionality is small and decreases as the dimensionality increases. Also, the information criterion is more likely to choose a larger value of λ than Bayesian optimization, and as the dimensionality increases, a smaller λ is more likely to be chosen for both optimization methods.
- The correlation with respect to the parameter α is weak when the dimensionality is small and increases as the dimensionality increases. Also, Bayesian optimization is more likely to choose a larger value of α than the information criterion, and as the dimensionality increases, a larger α is more likely to be chosen for both optimization methods.

In general, we can expect the greater the dimension of the input vector, the less confidence we have in the importance weighting. Thus, from the results in Figure 3.8, we can conjecture that λ is small and α is large when the importance weighting is unreliable.

3.6 Further Discussion

We generalized existing methods of covariate shift adaptation in the geometrical framework. By our information geometrical formulation, geometric biases of conventional methods are elucidated. Unlike the dominant approaches restricted to a specific curve on a manifold in the literature, our generalization has a much larger solution space with only two parameters. Our experiments highlighted the advantage of our method over

previous approaches, suggesting that our generalization can achieve better performance than the existing methods. A drawback of our proposed method is its relatively high computational cost for optimizing parameters α and λ. We used Bayesian optimization for efficient parameter search, and further sophisticated approaches will be explored in our future work.

As mentioned in the introduction, the importance weighting is used with deep neural network models [82], in which the importance weight in the feature representation obtained by DNN is considered. It is also worth mentioning that Sakai and Shimizu [252] used RIWERM in the study of covariate shift on the learning from positive and unlabeled data. Our generalization will be applicable to their methods to improve the performance under a small sample regime. In particular, in a standard approach for optimizing the implicit weight function $w(\boldsymbol{x})$, it is common to add a regularization term $(w(\boldsymbol{x}) - p_{tr}(\boldsymbol{x})/p_{te}(\boldsymbol{x}))^2$ to the optimization objective. The use of the derived geodesic and curvature biases to regularize the optimal weight function will be investigated in connection with the modern weight learning approach using deep neural network models. The relation between geometric bias and statistical bias should be explored. In addition, the development of a method for choosing good initial values for the parameter pairs α and λ is expected to significantly reduce the computational cost.

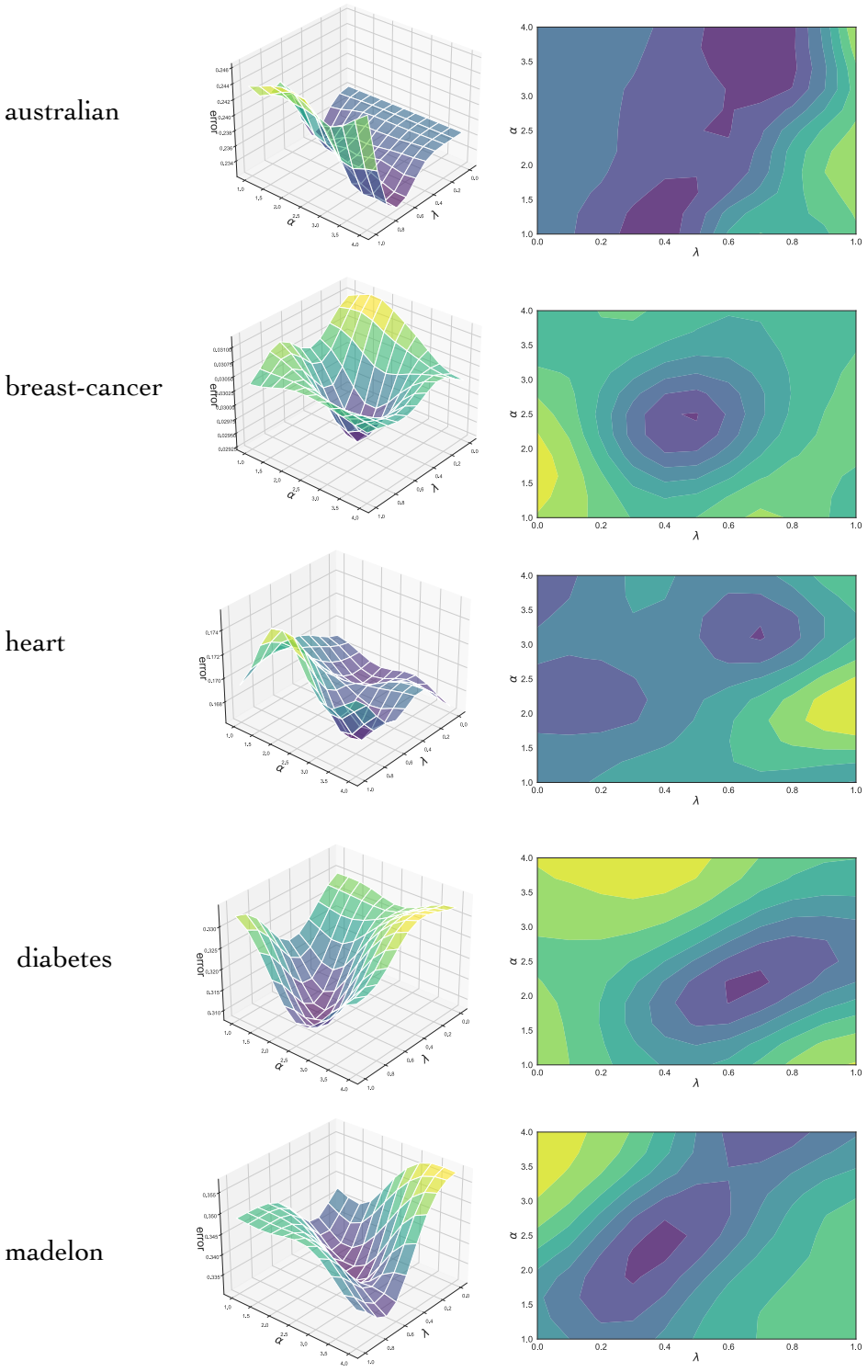

FIGURE 3.6: Visualization of grid search for α and λ on LIBSVM dataset. For the sake of clarity, we apply a moving average.

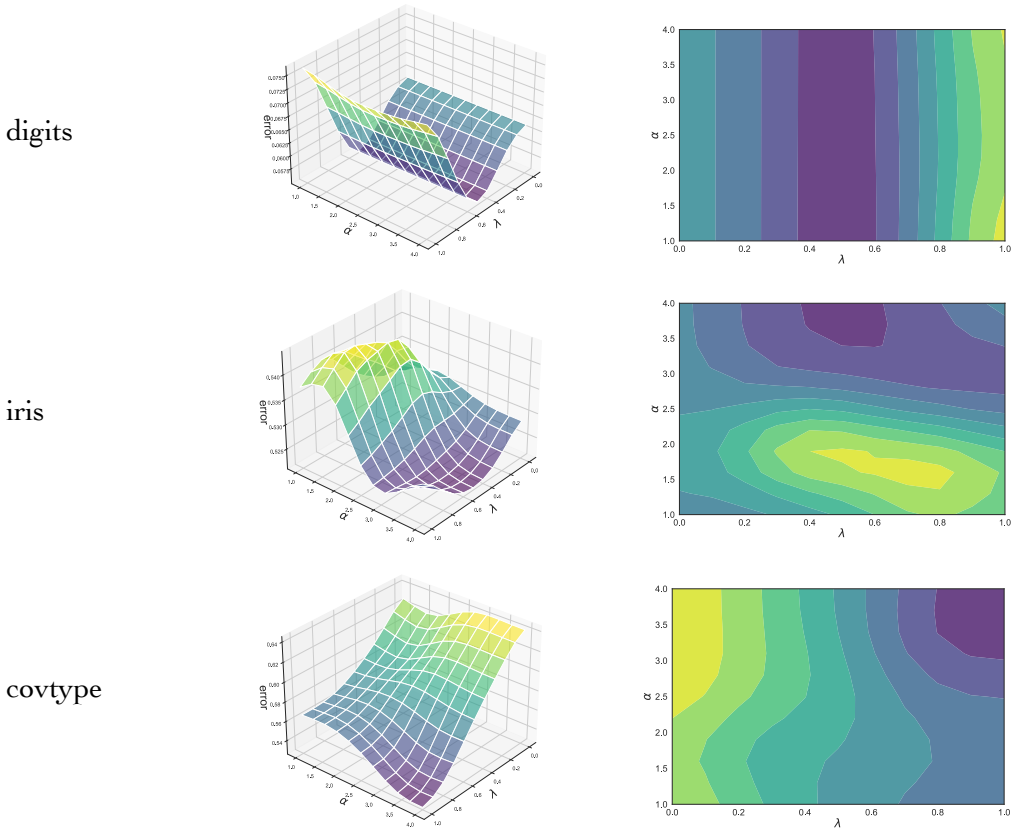

FIGURE 3.7: Visualization of grid search for α and λ on scikit-learn multi-class classification dataset. For the sake of visualization, we apply a moving average.

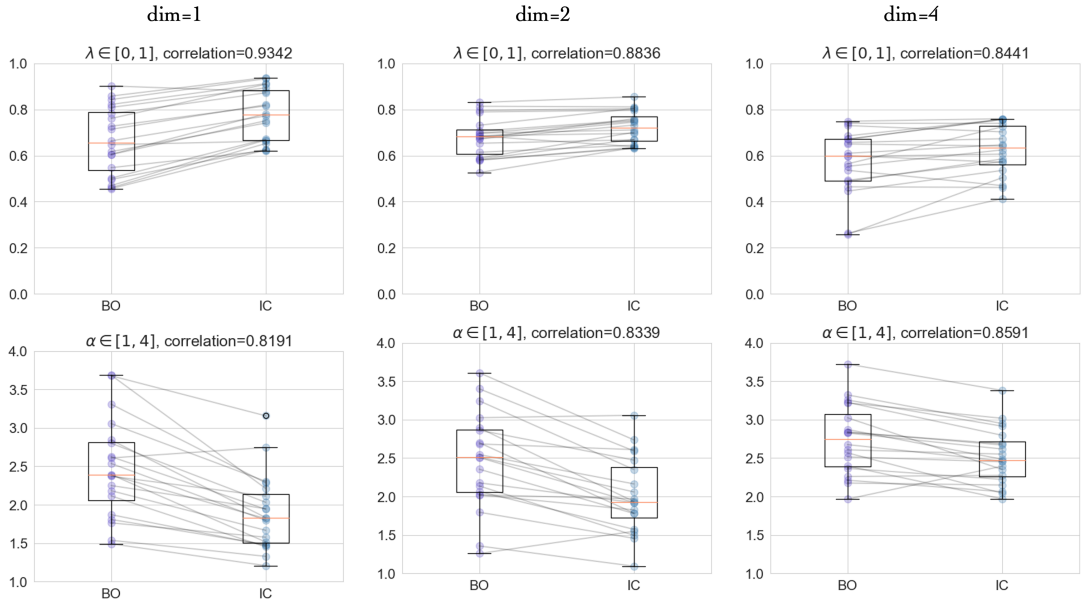

FIGURE 3.8: Comparison between Bayesian optimization and information criterion.

Chapter 4

α-Geodesical Skew Divergence

The asymmetric skew divergence smooths one of the distributions by mixing it, to a degree determined by the parameter λ, with the other distribution. Such divergence is an approximation of the KL divergence that does not require the target distribution to be absolutely continuous with respect to the source distribution. In this chapter, an information geometric generalization of the skew divergence called the α-geodesical skew divergence is proposed, and its properties are studied.

4.1 Definition of α-geodesical skew divergence

Consider the skew divergence of Definition 2.16, which is restated below:

$$\begin{aligned} D_S^{(\lambda)}[p\|q] &\coloneqq D_{\mathrm{KL}}[p\|(1-\lambda)p+\lambda q] \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{(1-\lambda)p(\boldsymbol{x})+\lambda q(\boldsymbol{x})} d\boldsymbol{x}, \end{aligned} \tag{4.1}$$

where $p, q \in \mathcal{P}$ and $\lambda \in [0,1]$. The skew divergence is a modified version of the KL-divergence that solves numerical difficulties and has many known applications.

The skew divergence is generalized based on the f-interpolation (Definition 2.31). From Definition 2.33, the f-interpoloation is considered as an unnormalized version of the α-geodesic. Using the notion of geodesics, skew divergence is generalized in terms of information geometry as follows.

Definition 4.1 (α-Geodesical Skew Divergence)**.** The α-geodesical skew divergence $D_{GS}^{(\alpha,\lambda)} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined between two probability distributions p and q by:

$$\begin{aligned} D_{GS}^{(\alpha,\lambda)}\Big[p\|q\Big] &\coloneqq D_{KL}\Big[p\|m_f^{(\lambda,\alpha)}(p,q)\Big] \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{m_f^{(\lambda,\alpha)}(p(\boldsymbol{x}), q(\boldsymbol{x}))} d\boldsymbol{x}, \end{aligned} \tag{4.2}$$

where $\alpha \in \mathbb{R}$ and $\lambda \in [0, 1]$.

Some special cases of α-geodesical skew divergence are listed below:

$$\begin{aligned}
(\forall \alpha \in \mathbb{R}, \lambda = 1)\ & D_{GS}^{(\alpha,1)}[p\|q] = D_{KL}[p\|q] \\
(\forall \alpha \in \mathbb{R}, \lambda = 0)\ & D_{GS}^{(\alpha,0)}[p\|q] = D_{KL}[p\|q] = 0 \\
(\alpha = 1, \forall \lambda \in [0,1])\ & D_{GS}^{(1,\lambda)}[p\|q] = \lambda D_{KL}[p\|q] \quad \text{(scaled KL-divergence)} \\
(\alpha = -1, \forall \lambda \in [0,1])\ & D_{GS}^{(-1,\lambda)}[p\|q] = D_S^{(\lambda)}[p\|q] \quad \text{(skew divergence)} \\
(\alpha = 0, \forall \lambda \in [0,1])\ & D_{GS}^{(0,\lambda)}[p\|q] = \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\{(1-\lambda)\sqrt{p(\boldsymbol{x})} + \lambda\sqrt{q(\boldsymbol{x})}\}^2} d\boldsymbol{x} \\
(\alpha = 3, \forall \lambda \in [0,1])\ & D_{GS}^{(3,\lambda)}[p\|q] = D_S^{(\lambda)}[p\|q] + H(p) + H(q) \\
(\alpha = \infty, \forall \lambda \in [0,1])\ & D_{GS}^{(\infty,\lambda)}[p\|q] = \int_{\mathcal{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{\min\{p(\boldsymbol{x}), q(\boldsymbol{x})\}} d\boldsymbol{x} \\
(\alpha = -\infty, \forall \lambda \in [0,1])\ & D_{GS}^{(-\infty,\lambda)}[p\|q] = \int_{\mathcal{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{\max\{p(\boldsymbol{x}), q(\boldsymbol{x})\}} d\boldsymbol{x}
\end{aligned}$$

Here, $H(p)$ is the entropy of p. Also, α-geodesical skew divergence is a special form of the generalized skew K-divergence [220, 223], which is a family of abstract means-based divergences. In this paper, the skew K-divergence touched upon in [220] is characterized in terms of α-geodesic on positive measures, and its geometric and functional analytic properties are investigated. When the Kolmogorov-Nagumo average (i.e., when the function f^{-1} in Definition 2.31 is a strictly monotone convex function) the geodesic has been shown to be well-defined [78].

4.1.1 Symmetrization of α-Geodesical Skew Divergence

It is easy to symmetrize the α-geodesical skew divergence as follows.

Definition 4.2 (Symmetrized α-Geodesical Skew Divergence)**.** The symmetrized α-geodesical skew divergence $\bar{D}_{GS}^{(\alpha,\lambda)} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined between two Radon–Nikodym

densities p and q of μ-absolutely continuous probability measures by:

$$\bar{D}_{GS}^{(\alpha,\lambda)}[p\|q] \coloneqq \frac{1}{2}\left(D_{GS}^{(\alpha,\lambda)}[p\|q] + D_{GS}^{(\alpha,\lambda)}[q\|p]\right), \tag{4.3}$$

where $\alpha \in \mathbb{R}$ and $\lambda \in [0,1]$.

It is seen that $\bar{D}_{GS}^{(\alpha,\lambda)}[p\|q]$ includes several symmetrized divergences.

$$\begin{aligned}
\bar{D}_{GS}^{(\alpha,1)}[p\|q] &= \frac{1}{2}\left(D_{KL}[p\|q] + D_{KL}[q\|p]\right), \quad \text{(half of Jeffreys divergence)} \\
\bar{D}_{GS}^{(-1,\frac{1}{2})}[p\|q] &= \frac{1}{2}\left(D_{KL}\Big[p\|\frac{p+q}{2}\Big] + D_{KL}\Big[q\|\frac{p+q}{2}\Big]\right), \quad \text{(JS-divergence)} \\
\bar{D}_{GS}^{(-1,\lambda)}[p\|q] &= \frac{1}{2}\left(D_{KL}\Big[p\|(1-\lambda)p+\lambda q\Big] + D_{KL}\Big[q\|(1-\lambda)q+\lambda p\Big]\right).
\end{aligned}$$

The last one is the λ-JS-divergence [218], which is a generalization of the JS-divergence.

4.2 Properties of α-geodesical skew divergence

In this section, the properties of the α-geodesical skew divergence are studied.

Proposition 4.3 (Non-negativity of the α-geodesical skew divergence)**.** *For $\alpha \geq -1$ and $\lambda \in [0,1]$, the α-geodesical skew divergence $D_{GS}^{(\alpha,\lambda)}[p\|q]$ satisfies the following inequality:*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \geq 0. \tag{4.4}$$

Proof. When λ is fixed, the f-interpolation has the following inverse monotonicity with respect to α:

$$m_f^{(\lambda,\alpha)}(p,q) \geq m_f^{(\lambda,\alpha')}(p,q), \ (\alpha \leq \alpha'). \tag{4.5}$$

From Jensen's inequality,

$$\begin{aligned}
\int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} &\leq \log\left(\int_{\mathcal{X}} p(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}\right) \\
&= \log \int_{\mathcal{X}} q(\boldsymbol{x}) d\boldsymbol{x} \\
&= \log 1 = 0.
\end{aligned} \tag{4.6}$$

We also have

$$\int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} = \int_{\mathcal{X}} p(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x}. \tag{4.7}$$

From Eqs. (4.6) and (4.7),

$$\int_{\mathcal{X}} p(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \leq 0$$
$$- \int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \leq - \int p(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x}. \tag{4.8}$$

From Eqs. (4.8) and (4.5), one obtains

$$\begin{aligned} D_{GS}^{(\alpha,\lambda)}[p\|q] &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{m_f^{(\alpha,\lambda)}(p(\boldsymbol{x}), q(\boldsymbol{x}))} d\boldsymbol{x} \\ &\geq \Big(\int_{\mathcal{X}} p(\boldsymbol{x}) d\boldsymbol{x} \Big) \log \frac{p(\boldsymbol{x})}{m_f^{(\alpha,\lambda)}(p(\boldsymbol{x}), q(\boldsymbol{x}))} \\ &\geq 1 \cdot \log 1 = 0. \end{aligned}$$

□

Proposition 4.4 (Asymmetry of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence is not symmetric in general:*

$$D_{GS}^{(\alpha,\lambda)}[p|q] \neq D_{GS}^{(\alpha,\lambda)}[q\|p]. \tag{4.9}$$

Proof. For example, if $\lambda = 1$, then $\forall \alpha \in \mathbb{R}$, it holds that

$$D_{GS}^{(\alpha,1)}[p\|q] - D_{GS}^{(\alpha,1)}[q\|p] = D_{KL}[p\|q] - D_{KL}[q\|p],$$

and the asymmetry of the KL-divergence results in an asymmetry of the geodesic skew divergence. □

When a function $f(x)$ of $x \in [0, 1]$ satisfies $f(x) = f(1 - x)$, it is referred to be centrosymmetric.

Proposition 4.5 (Non-centrosymmetricicy of the α-geodesical skew divergence with respect to λ)**.** *α-Geodesical skew divergence is not centrosymmetric in general with respect to the parameter $\lambda \in [0, 1]$:*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \neq D_{GS}^{(\alpha,1-\lambda)}[p\|q]. \tag{4.10}$$

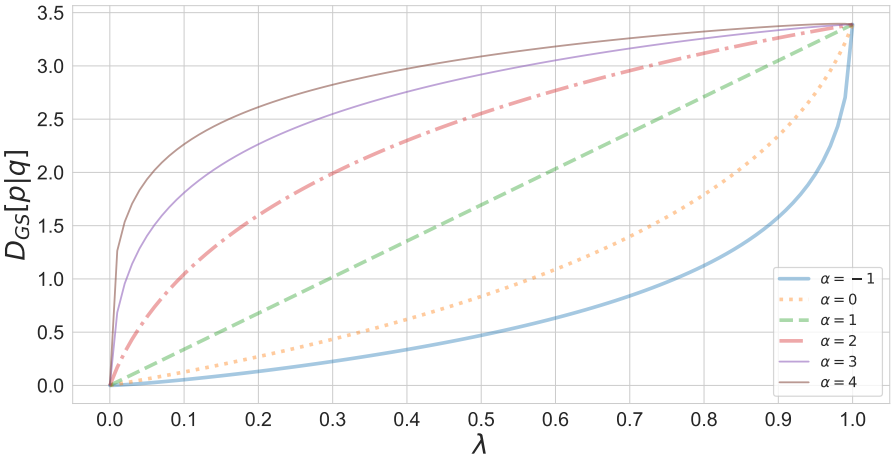

FIGURE 4.1: Monotonicity of the α-geodesical skew divergence with respect to α. The α-geodesical skew divergence between the binomial distributions $p = B(10, 0.3)$ and $q = B(10, 0.7)$ has been calculated.

Proof. For example, if $\lambda = 1$, then $\forall \alpha \in \mathbb{R}$, we have

$$\begin{aligned} D_{GS}^{(\alpha,\lambda)}[p\|q] - D_{GS}^{(\alpha,1-\lambda)}[p\|q] &= D_{GS}^{(\alpha,1)}[p\|q] - D_{GS}^{(\alpha,0)}[p\|q] \\ &= \int_{\mathcal{X}} p \ln \frac{p}{q} - \int_{\mathcal{X}} p \ln \frac{p}{p} \\ &= \int_{\mathcal{X}} p \ln \frac{p}{q} \geq 0. \end{aligned} \tag{4.11}$$

□

Proposition 4.6 (Monotonicity of the α-geodesical skew divergence with respect to α)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha \in \mathbb{R}, \lambda \in [0, 1]$.*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \geq D_{GS}^{(\alpha',\lambda)}[p\|q], \ (\alpha \geq \alpha').$$

Proof. Obvious from the inverse monotonicity of the f-interpolation (4.5) and the monotonicity of the logarithmic function. □

Figure 4.1 shows the monotonicity of the geodesic skew divergence with respect to α. In this figure, divergence is calculated between two binomial distributions.

Proposition 4.7 (Subadditivity of the α-geodesical skew divergence with respect to α)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha, \beta \in \mathbb{R}, \lambda \in [0, 1]$*

$$D_{GS}^{(\alpha+\beta,\lambda)}[p\|q] \leq D_{GS}^{(\alpha,\lambda)}[p\|q] + D_{GS}^{(\beta,\lambda)}[p\|q].$$

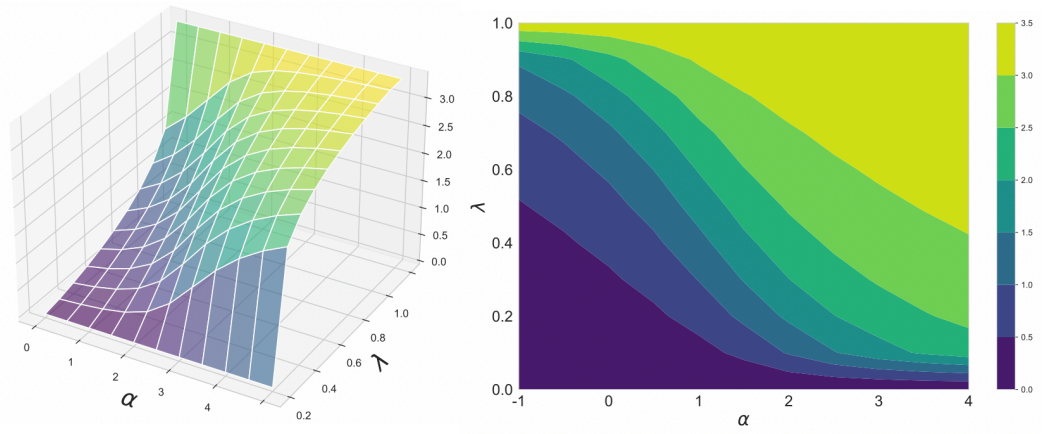

FIGURE 4.2: Continuity of the α-geodesical skew divergence with respect to α and λ. The α-geodesical skew divergence between the binomial distributions $p = B(10, 0.3)$ and $q = B(10, 0.7)$ has been calculated.

Proof. For some α and λ, $m_f^{(\lambda,\alpha)}$ takes the form of the Kolmogorov mean [151], and obvious from its continuity, monotonicity and self-distributivity. □

Proposition 4.8 (Continuity of the α-geodesical skew divergence with respect to α and λ)**.** *α-Geodesical skew divergence has the continuity property.*

Proof. We can prove from the continuity of the KL-divergence and the Kolmogorov mean. □

Figure 4.2 shows the continuity of the geodesic skew divergence with respect to α and λ. Both source and target distributions are binomial distributions. From this figure, it can be seen that the divergence changes smoothly as the parameters change.

Lemma 4.9. *For all $\lambda \in (0, 1)$, we have*

$$\lim_{\alpha\to\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] = \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\min\{p(\boldsymbol{x}), q(\boldsymbol{x})\}} d\boldsymbol{x}. \tag{4.12}$$

Proof. Let $u = \frac{1-\alpha}{2}$. Then $\lim_{\alpha\to\infty} u = -\infty$. Assuming $p_0(\boldsymbol{x}) \leq p_1(\boldsymbol{x})$ for some $\boldsymbol{x} \in \mathcal{X}$, it holds that

$$\begin{aligned}
\lim_{\alpha\to\infty} m_f^{(\lambda,\alpha)}(p_0(\boldsymbol{x}), p_1(\boldsymbol{x})) &= \lim_{u\to-\infty} \Big((1-\lambda)p_0(\boldsymbol{x})^u + \lambda p_1(\boldsymbol{x})^u\Big)^{\frac{1}{u}} \\
&= p_0(\boldsymbol{x}) \lim_{u\to-\infty} \left((1-\lambda) + \lambda\Big(\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}\Big)^u\right)^{\frac{1}{u}} \\
&= p_0(\boldsymbol{x}) = \min\{p_0(\boldsymbol{x}), p_1(\boldsymbol{x})\}.
\end{aligned}$$

Similarly,

$$\lim_{\alpha\to-\infty} m_f^{(\lambda,\alpha)}(p_0(\boldsymbol{x}),p_1(\boldsymbol{x})) = \max\{p_0(\boldsymbol{x}),p_1(\boldsymbol{x})\} \tag{4.13}$$

and

$$\min\{p_0(\boldsymbol{x}),p_1(\boldsymbol{x})\} \leq m_f^{(\lambda,\alpha)}(p_0(\boldsymbol{x}),p_1(\boldsymbol{x})) \leq \max\{p_0(\boldsymbol{x}),p_1(\boldsymbol{x})\}, \tag{4.14}$$

for all $\alpha \in \mathbb{R}$ and $p_0(\boldsymbol{x}), p_1(\boldsymbol{x}) \in (0,1]$. Let $C_m(\boldsymbol{x}) = \min\{p_0(\boldsymbol{x}),p_1(\boldsymbol{x})\}$ and $C_M(\boldsymbol{x}) = \max\{p_0(\boldsymbol{x}),p_1(\boldsymbol{x})\}$, we can write

$$\begin{aligned} &C_m(\boldsymbol{x}) \leq m_f^{(\lambda,\alpha)}(p_0(\boldsymbol{x}),p_1(\boldsymbol{x})) \leq C_M(\boldsymbol{x}) \\ &\quad 0 \leq C_m(\boldsymbol{x}) \leq m_f^{(\lambda,\alpha)}(p_0(\boldsymbol{x}),p_1(\boldsymbol{x})) \leq C_M(\boldsymbol{x}) \leq 1 \left| m_f^{(\lambda,\alpha)}(p_0(\boldsymbol{x}),p_1(\boldsymbol{x})) \right| \leq 1. \end{aligned} \tag{4.15}$$

From the Lebesgue's dominated convergence theorem, we have the following equality almost surely

$$\begin{aligned} \lim_{\alpha\to\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\lim_{\alpha\to\infty} m_f^{(\lambda,\alpha)}(p(\boldsymbol{x}),q(\boldsymbol{x}))} d\boldsymbol{x} \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\min\{p(\boldsymbol{x}),q(\boldsymbol{x})\}} d\boldsymbol{x} \end{aligned}$$

holds. □

Lemma 4.10. *For all $\lambda \in (0,1)$, we have*

$$\lim_{\alpha\to-\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] = \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\max\{p(\boldsymbol{x}),q(\boldsymbol{x})\}} d\boldsymbol{x}. \tag{4.16}$$

Proof. We obtain a proof similar to Lemma 4.9. □

Proposition 4.11 (Lower bound of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha \in \mathbb{R}, \lambda \in (0,1)$.*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \geq \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\max\{p(\boldsymbol{x}),q(\boldsymbol{x})\}} d\boldsymbol{x}. \tag{4.17}$$

Proof. It follows from the definition of the inverse monotonicity of f-interpolation (4.5) and Lemma 4.10. □

Proposition 4.12 (Upper bound of the α-geodesical skew divergence))**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha \in \mathbb{R}, \lambda \in (0,1)$.*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \leq \int_{\mathcal{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{\min\{p(\boldsymbol{x}),q(\boldsymbol{x})\}} d\boldsymbol{x}. \tag{4.18}$$

Proof. It follows from the definition of the f-interpolation (4.5) and Lemma 4.9. □

Theorem 4.13 (Strong convexity of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence $D_{GS}^{(\alpha,\lambda)}[p\|q]$ is strongly convex in p with respect to the total variation norm.*

Proof. Let $r \coloneqq m_f^{(\alpha,\lambda)}(p,q)$ and $f_j \coloneqq \frac{p_j}{r}$ $(j=0,1)$, so that $f_t = \frac{p_t}{r}$ $(t \in (0,1))$. From Taylor's theorem, for $g(x) \coloneqq x \ln x$ and $j = 0, 1$, it holds that

$$g(f_j) = g(f_t) + g'(f_t)(f_j - f_t) + (f_j - f_t)^2 \int_0^1 g''((1-s)f_t + sf_j)(1-s)ds.$$

Let

$$\begin{aligned}
\delta &\coloneqq (1-t)g(f_0) + tg(f_1) - g(f_t) \\
&= (1-t)t(f_1 - f_0)^2 \int_0^1 \left(\frac{t}{(1-s)f_t + sf_0} + \frac{1-t}{(1-s)f_t + sf_1} \right)(1-s)ds \\
&= (1-t)t(f_1 - f_0)^2 \int_0^1 \left(\frac{t}{f_{u_0}(t,s)} + \frac{1-t}{f_{u_1}(t,s)} \right)(1-s)ds,
\end{aligned}$$

where

$$\begin{aligned}
u_j(t,s) &\coloneqq (1-s)t + jt, \\
f_{\mu_j}(t,s) &\coloneqq (1-s)f_t + sf_j.
\end{aligned}$$

Then,

$$\begin{aligned}
\Delta &\coloneqq (1-t)H(p_0) + tH(p_1) - H(p_t) \\
&= \int \delta dr \\
&= (1-t)t \int_0^1 (1-s)ds \Big[tI(u_0(t,s)) + (1-t)I(u_1(t,s)) \Big],
\end{aligned}$$

where

$$\begin{aligned}
\|p_1 - p_0\| &\coloneqq \int_{\mathcal{X}} |dp_1(\boldsymbol{x}) - dp_0(\boldsymbol{x})| d\boldsymbol{x}, \\
H(p) &\coloneqq D_{GS}^{(\alpha,\lambda)}[p\|r] = \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{r(\boldsymbol{x})} d\boldsymbol{x}, \\
I(u) &\coloneqq \int \frac{(f_1 - f_0)^2}{f_u} dr.
\end{aligned}$$

Now, it is suffice to prove that $\Delta \geq \frac{t(1-t)}{2}\|p_1 - p_0\|^2$. For all $u \in (0,1)$, it is seen that p_1 is absolutely continuous with respect to p_u. Let $g_u \coloneqq \frac{p_1}{p_u} = \frac{f_1}{f_u}$. One obtains

$$\begin{aligned} I(u) &= \frac{1}{(1-u)^2}\int \frac{(f_1 - f_u)^2}{f_u} dr \\ &= \frac{1}{(1-u)^2}\int (g_u - 1)^2 dp_u \\ &\geq \frac{1}{(1-u)^2}\left(\int |g_u - 1| dp_u\right)^2 \\ &= \frac{1}{(1-u)^2}\|p_1 - p_u\|^2 = \|p_1 - p_0\|^2, \end{aligned}$$

and hence, for $j = 0, 1$,

$$\Delta \geq \frac{t(1-t)}{2}\|p_1 - p_0\|^2.$$

□

4.3 Natural α-geodesical skew divergence for exponential family

In this section, the exponential family is considered in which the probability density function is given by Eq. (2.5). In skew divergence, the probability distribution of the target is a weighted average of the two distributions. This implicitly assumes that interpolation of the two probability distributions is properly given by linear interpolation. Here, in the exponential family, the interpolation between natural parameters rather than the interpolation between probability distributions themselves is considered. Namely, the geodesic connecting two distributions $p(\boldsymbol{x}; \boldsymbol{\theta}_p)$ and $q(\boldsymbol{x}; \boldsymbol{\theta}_q)$ on the $\boldsymbol{\theta}$-coordinate system is considered:

$$\boldsymbol{\theta}(\lambda) = (1-\lambda)\boldsymbol{\theta}_p + \lambda\boldsymbol{\theta}_q, \tag{4.19}$$

where $\lambda \in [0,1]$ is the parameter. The probability distributions on the geodesic $\boldsymbol{\theta}(\lambda)$ are

$$\begin{aligned} p(\boldsymbol{x}; \lambda) &= p(\boldsymbol{x}; \boldsymbol{\theta}(\lambda)) \\ &= \exp\Big\{\lambda(\boldsymbol{\theta}_q - \boldsymbol{\theta}_p)\cdot\boldsymbol{x} + \boldsymbol{\theta}_p\cdot\boldsymbol{x} - \psi(\lambda)\Big\}. \end{aligned} \tag{4.20}$$

Hence, a geodesic itself is a one-dimensional exponential family, where λ is the natural parameter. A geodesic consists of a linear interpolation of the two distributions in the logarithmic scale because

$$\ln p(\boldsymbol{x}; \lambda) = (1-\lambda)\log p(\boldsymbol{x}; \boldsymbol{\theta}_p) + \lambda \log p(\boldsymbol{x}; \boldsymbol{\theta}_q) - \psi(\lambda). \tag{4.21}$$

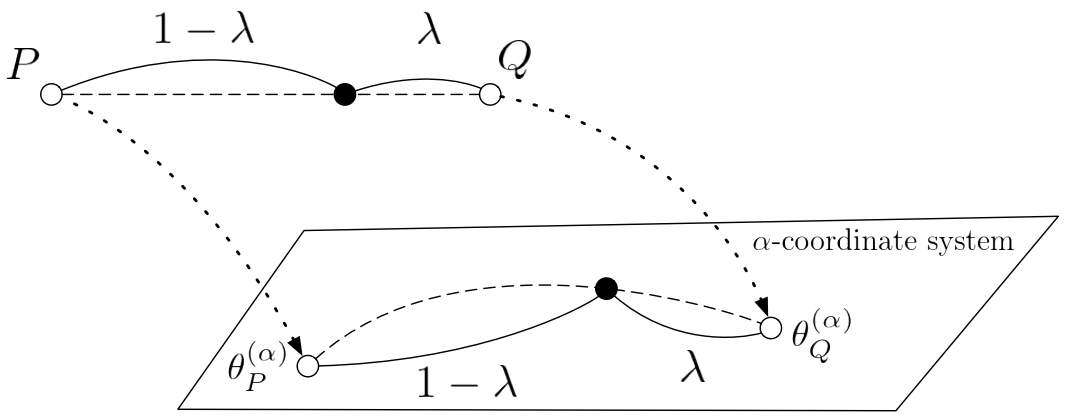

FIGURE 4.3: The geodesic between two probability distributions on the α-coordinate system.

This corresponds to the case $\alpha = 1$ on the f-interpolation with normalization factor $c(\lambda) = \exp\{-\psi(\lambda)\}$,

$$p(\boldsymbol{x}; \boldsymbol{\theta}(\lambda)) = m_f^{(\lambda,1)}(p(\boldsymbol{x}; \boldsymbol{\theta}_p), p(\boldsymbol{x}; \boldsymbol{\theta}_q)). \tag{4.22}$$

This induces the natural geodesic skew divergence with $\alpha = 1$ as

$$\begin{aligned}
D_{GS}^{(1,\lambda)}[p\|q] &= \int_{\mathcal{X}} p(\boldsymbol{x}) \log \left(\frac{p(\boldsymbol{x})}{m_f^{(\lambda,1)}(p(\boldsymbol{x}), q(\boldsymbol{x}))} \right) d\boldsymbol{x} \\
&= \int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - p(\boldsymbol{x}) \log \left(m_f^{(\lambda,1)}(p(\boldsymbol{x}), q(\boldsymbol{x})) \right) d\boldsymbol{x} \\
&= \int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - p(\boldsymbol{x}) \log \left(\exp\{(1-\lambda) \log p(\boldsymbol{x}) + \lambda \log q(\boldsymbol{x})\} \right) d\boldsymbol{x} \\
&= \int_{\mathcal{X}} \left(p(\boldsymbol{x}) \log p(\boldsymbol{x}) - (1-\lambda) p(\boldsymbol{x}) \ln p(\boldsymbol{x}) - \lambda p(\boldsymbol{x}) \log q(\boldsymbol{x}) \right) d\boldsymbol{x} \\
&= \int_{\mathcal{X}} \left(\lambda p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \lambda p(\boldsymbol{x}) \log q(\boldsymbol{x}) \right) d\boldsymbol{x} \\
&= \lambda \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \\
&= \lambda D_{KL}[p\|q],
\end{aligned}$$

and this is equal to the scaled KL divergence.

More generally, let $\theta_P^{(\alpha)}$ and $\theta_Q^{(\alpha)}$ be the parameter representations on the α-coordinate system of probability distributions P and Q. Then, the geodesics between them are represented as in Figure 4.3 and it induces the α-geodesical skew divergence.

4.4 Function space associated with the α-geodesical skew divergence

To discuss the functional nature of the α-geodesical skew divergence in more depth, the function space it constitutes is considered. For an α-geodesical skew divergence $f_q^{(\alpha,\lambda)}(p) = D_{GS}^{(\alpha,\lambda)}[p\|q]$ with one side of the distribution fixed, let the entire set be

$$\mathcal{F}_q = \left\{ f_q^{(\alpha,\lambda)} \mid \alpha \in \mathbb{R}, \lambda \in [0,1] \right\}. \tag{4.23}$$

For $f_q^{(\alpha,\lambda)} \in \mathcal{F}_q$, its semi-norm is defined by

$$\left\| f_q^{(\alpha,\lambda)} \right\|_p \coloneqq \int_{\mathcal{X}} \left(\left| f_q^{(\alpha,\lambda)} \right|^p d\mu \right)^{\frac{1}{p}}. \tag{4.24}$$

By defining addition and scalar multiplication for $f_q^{(\alpha,\lambda)}, g_q^{(\alpha,\lambda)} \in \mathcal{F}_q$, $c \in \mathbb{R}$ as follows, $\mathcal{F}_q$ becomes a semi-norm vector space:

$$(f_q^{(\alpha,\lambda)} + g_q^{(\alpha,\lambda)})(u) \coloneqq f_q^{(\alpha,\lambda)}(u) + g_q^{(\alpha,\lambda)}(u) = D_{GS}^{(\alpha,\lambda)}[u\|q] + D_{GS}^{(\alpha',\lambda')}[u\|q], \tag{4.25}$$

$$(cf)(u) \coloneqq c f_q^{(\alpha,\lambda)}(u) = c \cdot D_{GS}^{(\alpha,\lambda)}[u\|q]. \tag{4.26}$$

Theorem 4.14. *Let $\mathcal{N}$ be the kernel of $\|\cdot\|_p$ as follows:*

$$\mathcal{N} \coloneqq ker(\|\cdot\|_p) = \left\{ f_q^{(\alpha,\lambda)} \mid f_q^{(\alpha,\lambda)} = 0 \right\}. \tag{4.27}$$

Then the quotient space $\mathcal{V} \coloneqq (\mathcal{F}_q, \|\cdot\|_p)/N$ is a Banach space.

Proof. It is sufficient to prove that $f_q^{(\alpha,\lambda)}$ is integrable to the power of p and that $\mathcal{V}$ is complete. From Proposition 4.12, the α-geodesical skew divergence is bounded from above for all $\alpha \in \mathbb{R}$ and $\lambda \in [0,1]$. Since $f_q^{(\alpha,\lambda)}$ is continuous, we know that it is p-power integrable.

Let $\{f_n\}$ be a Cauchy sequence of $\mathcal{V}$:

$$\lim_{n,m\to\infty} \|f_n - f_m\|_p = 0.$$

Since $n(k),\ k = 1, 2, \ldots,$ can be taken to be monotonically increasing and

$$\|f_n - f_{n(k)}\|_p < 2^{-k}$$

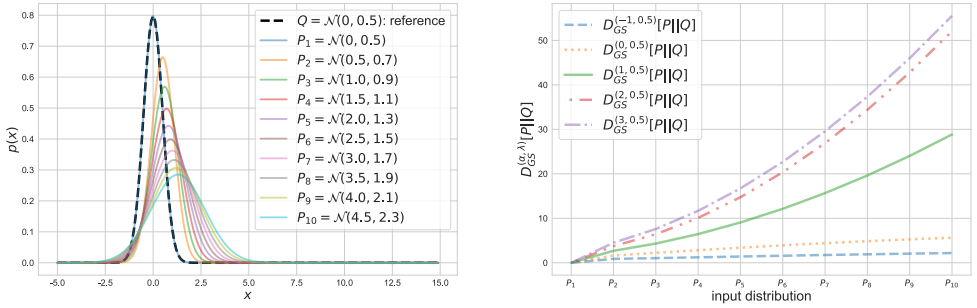

FIGURE 4.4: α-geodesical skew divergence between two normal distributions. The reference distribution is $Q = \mathcal{N}(0, 0.5)$. For $P_1, P_2, \ldots, P_j$, $(j = 1, 2, \ldots, 10)$, let their mean and variance be μ_j and σ_j^2, respectively, where $\mu_{j+1} - \mu_j = 0.5$ and $\sigma_{j+1}^2 - \sigma_j^2 = 0.2$.

with respect to $n > n(k)$, let

$$\|f_{n(k+1)} - f_{n(k)}\|_p < 2^{-k}.$$

If $g_n = |f_{n(1)}| + \sum_{j=1}^{n-1} |f_{n(j+1)} - f_{n(j)}| \in \mathcal{V}$, it is non-negatively monotonically increasing at each point, and from the subadditivity of the norm, $\|g_n\|_p \leq \|f_{n(1)}\|_p + \sum_{j=1}^{n-1} 2^{-j}$. From the monotonic convergence theorem, we have

$$\left\| \lim_{n\to\infty} g_n \right\|_p = \lim_{n\to\infty} \|g_n\|_p \leq \|f_{n(1)}\|_p + 1 < \infty.$$

That is, $\lim_{n\to\infty} g_n$ exists almost everywhere, and $\lim_{n\to\infty} g_n \in \mathcal{V}$. From $\lim_{n\to\infty} g_n < \infty$, we have

$$f_{n(1)} + \sum_{j=1}^{n-1} (f_{n(j+1)} - f_{n(j)}) = \lim_{n\to\infty} f_{n(1)}$$

converges absolutely almost everywhere to $|\lim_{n\to\infty} f_{n(n)}| \leq \lim_{n\to\infty} g_n, a.e.$. That is, $\lim_{n\to\infty} f_{n(n)} \in \mathcal{V}$. Then

$$\left| \lim_{n\to\infty} f_n - f_{n(n)} \right| \leq \lim_{n\to\infty} g_n$$

and from the superior convergence theorem, we can obtain

$$\lim_{m\to\infty} \left\| \lim_{n\to\infty} f_n - f_{n(m)} \right\|_p = 0$$

We have now confirmed the completeness of $\mathcal{V}$. □

Corollary 4.15. *Let*

$$\mathcal{F}_+ = \left\{ f_q^{(\alpha,\lambda)} \mid \alpha \in \mathbb{R}, \lambda \in (0, 1], q \in \mathcal{P} \right\}. \tag{4.28}$$

Then the space $\mathcal{V}_+ := (\mathcal{F}_+, \|\cdot\|_p)$ *is a Banach space.*

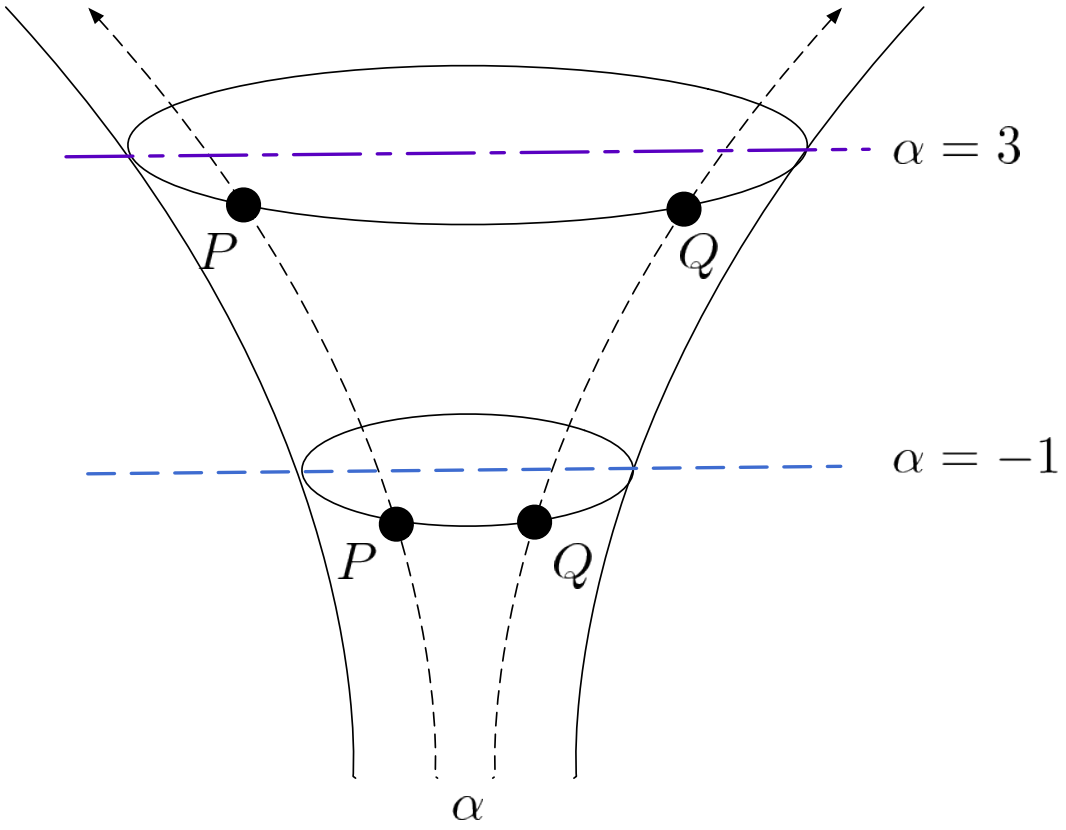

FIGURE 4.5: Coordinate system of $\mathcal{F}_q$ or $\mathcal{F}_+$. Such a coordinate system is not Euclidean.

Proof. If we restrict $\lambda \in (0, 1]$, $D_{GS}^{(\alpha,\lambda)}[u\|q] = 0$ if and only if $u = q$. Then, $\mathcal{V}_+$ has the unique identity element, and then $\mathcal{V}_+$ is a complete norm space. □

Consider the second argument Q of $D_{GS}^{(\alpha,\lambda)}(P||Q)$ is fixed, which is referred to as the reference distribution. Figure 4.4 shows values of the α-geodesical skew divergence for a fixed reference Q, where both P and Q are restricted to be Gaussian. In this figure, the reference distribution is $\mathcal{N}(0, 0.5)$ and the parameters of input distributions are varied in $\mu \in [0, 4.5]$ and $\sigma^2 \in [0.5, 2.3]$. From this figure, one can see that a larger value of α emphasizes the discrepancy between distributions P and Q. Figure 4.5 illustrates a coordinate system associated with the α-geodesical skew divergence for different α. As seen from the figure, for the same pair of distributions P and Q, the value of divergence with $\alpha = 3$ is larger than that with $\alpha = -1$.

4.5 Discussion and Possible Applications

In this chapter, we considered a generalization of the skew divergence and studied several properties of the generalized skew divergence. We presented that our generalization has

several known divergences as special cases. Moreover, as mentioned in Section 4.2, we see that our generalized skew divergence has desirable properties as a divergence, such as non-negativity.

In our generalization, we introduced a new parameter pair α and λ. From Figs. 4.1 and 4.2, we can confirm that the value of the generalized skew divergence varies monotonically and smoothly with respect to changes in these parameters. Optimizing these parameters could improve existing algorithms that rely on skew divergence, similar to the generalized covariate shift adaptation (Chapter 3). For example, the effectiveness of skew divergence has been reported for segmentation [49], disease detection [247], and natural language processing [315]. Furthermore, since the skew divergence is a modification of the KL-divergence by stabilizing its numerical difficulties, it is expected to be improved by replacing existing algorithms using the KL-divergence [6, 60, 92, 143, 144, 205, 262, 323]. The implementation of our generalized skew divergence in PyTorch [231], one of the automatic differentiation frameworks, is available on GitHub [1]. We hope that with our implementation, the possible applications discussed above will be realized.

[1] https://github.com/nocotan/geodesical_skew_divergence

Chapter 5

Conclusion and Future Work

In this study, we considered a manifold constructed by a set of probability distributions. In particular, this manifold is generally a non-Euclidean space and is known to be a Riemannian manifold. The framework that focuses on the geometric properties of a Riemannian manifold created by a set of probability distributions is called information geometry, and has many known applications. One of the advantages of using information geometry for discussion is that some statistical concepts can be identified with geometric objects. The identification of these two types of concepts has several benefits, such as a deeper understanding of statistical procedures and the induction of geometrically natural generalizations of existing algorithms. In this study, we considered a generalization of algorithms for handling covariate shift, a type of distribution shift, and a generalization of the skew divergence, one of the variants of the KL-divergence. Our generalizations are obtained by identifying the generalized mean with the selection of the curves. In particular, this curve is determined by a certain real parameter α and corresponds to a special curve called the α-geodesic. Since the α-geodesic is a straight curve in the α-coordinate system, our identification can be viewed as an identification of the choice of weighting strategy and the coordinate system in which the curve connecting two probability distributions is straight. This identification is discussed in Chapter 2.

The covariate shift assumes that the distributions of the input variables in the training and test data are different, and is one of the important research topics in machine learning. This is because the validity of most supervised learning is supported by the consistency of the ERM, and this property is not satisfied under the covariate shift

assumption. Generalized covariate shift adaptation (Chapter 3) is a generalization of importance-weighted ERM and its variants by introducing a new parameter pair. In this parameter pair, α corresponds to the shape of the curve connecting the two probability distributions and λ to the position on the curve. We experimented with two different optimization strategies for optimizing these parameters, Bayesian optimization and information criterion, and observed that in some cases their behavior differed. Our conjecture is that the stability of each optimization strategy is related, but there is still insufficient analysis to support this. Also, the theoretical properties of our generalized covariate shift adaptation for a given parameter pair have not yet been fully analyzed. Based on existing research and our numerical experiments, it can be expected that the parameter λ, which determines the position on the curve connecting the two probability distributions, corresponds to the confidence of the weighting. However, the role of the parameter α, which determines the shape of the curve, in optimization is still under discussion. The investigation of the differences in the behavior of optimization strategies and the properties of the generalized method depending on the parameters is a subject for future work. Our implementation of generalized covariate shift adaptation is available on GitHub [1] and more extensive experimentation is expected.

Finally, we considered a generalization of the skew divergence, one of the variants of the KL-divergence, in Chapter 4. The KL-divergence is known to be subject to numerical instability because of the inclusion of the density ratio between probability distributions in the integral. One idea to overcome this difficulty is skew divergence, which is achieved by replacing the denominator of the density ratio with the weighted average of two probability distributions. Our generalized skew divergence is based on α-geodesics, as is the generalization of the covariate shift adaptation above, and a similar parameter pair is introduced. We confirmed that generalized skew divergence has several desirable properties as divergence. However, its specific applications are still future issues to be addressed. We confirmed that the value of the generalized skew divergence varies monotonically and smoothly with respect to changes in these parameters, and it is expected to improve existing algorithms that rely on skew divergence by optimizing these parameters. The implementation of our generalized skew divergence in PyTorch is available on GitHub [2].

[1] https://github.com/nocotan/IGIWERM

[2] https://github.com/nocotan/geodesical_skew_divergence

Appendix A

Details for IGIWERM

This section describes the technical details of Chapter 3. In particular, we describe material here that would be too verbose to be included in the main body of the thesis.

A.1 Learning guarantee

We note the derivation of the learning bounds introduced in Eq. (3.40), following the paper by Cortes et al. [62]. First, we introduce the definition of the Rényi divergence, a necessary tool for theoretical analysis.

Definition A.1 (Rényi divergence [246])**.** For $\alpha > 0$, the Rényi divergence $D_R^{(\alpha)}[p\|q]$ between two probability distributions p and q is defined by

$$D_R^{\alpha}[p\|q] \coloneqq \frac{1}{\alpha - 1} \log_2 \int_{\mathcal{X}} p(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^{\alpha-1} d\boldsymbol{x}. \tag{A.1}$$

Next, the required theorems and their proofs are reproduced below. Here, following the paper by Cortes et al. [62], we use the notation $\hat{\mathbb{P}}$ to denote the empirical distribution based on a finite sample of size n, and $\hat{\mathbb{E}}$ to denote the expectation based on $\hat{\mathbb{P}}$.

Lemma A.2 ([62])**.** *For all* $\alpha > 0$ *and* $w(\boldsymbol{x}) = \frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}$,

$$\mathbb{E}_{(\boldsymbol{x},y)\sim p_{tr}} \left[w^2(\boldsymbol{x})\ell^2(h(\boldsymbol{x}), y)\right] \leq 2^{D_R^{\alpha+1}[p_{te}\|p_{tr}]} \mathbb{E}_{(\boldsymbol{x},y)\sim p_{te}} \left[\ell(h(\boldsymbol{x}), y)\right]^{1-\frac{1}{\alpha}}. \tag{A.2}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{x},y)\sim p_{tr}}\left[w^2(\boldsymbol{x})\ell^2(h(\boldsymbol{x}),y)\right] &= \int_{\mathcal{X}}\left(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\right)^2 \ell^2(h(\boldsymbol{x}),y)p_{tr}(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathcal{X}} p_{te}(\boldsymbol{x})^{\frac{1}{\alpha}}\left(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\right) p_{te}(\boldsymbol{x})^{\frac{\alpha-1}{\alpha}}\ell^2(h(\boldsymbol{x}),y)d\boldsymbol{x} \\
&\leq \left(\int_{\mathcal{X}} p_{te}(\boldsymbol{x})\left(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\right)^{\alpha} d\boldsymbol{x}\right)^{\frac{1}{\alpha}}\left(\int_{\mathcal{X}} p_{te}(\boldsymbol{x})\ell^{\frac{2\alpha}{\alpha-1}}(h(\boldsymbol{x}),y)d\boldsymbol{x}\right)^{\frac{\alpha-1}{\alpha}} \\
&= 2^{D_R^{\alpha+1}[p_{te}\|p_{tr}]}\left(\int_{\mathcal{X}} p_{te}(\boldsymbol{x})\ell(h(\boldsymbol{x}),y)\ell^{\frac{\alpha+1}{\alpha-1}}(h(\boldsymbol{x}),y)d\boldsymbol{x}\right) \\
&\leq 2^{D_R^{\alpha+1}[p_{te}\|p_{tr}]}\mathbb{E}_{(\boldsymbol{x},y)\sim p_{te}}\left[\ell(h(\boldsymbol{x}),y)\right]^{1-\frac{1}{\alpha}}.
\end{aligned}$$

□

Theorem A.3. *For any loss function ℓ and hypothesis class $\mathcal{H}$ such that $0 < \mathbb{E}_{(\boldsymbol{x},y)\sim p}\left[\ell(h(\boldsymbol{x}),y)\right] < +\infty$ for all $h \in \mathcal{H}$ and a probability distribution p,*

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\mathbb{E}[\ell(h(\boldsymbol{x}),y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\mathbb{E}}[\ell^2(h(\boldsymbol{x}),y)]} > \epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right] \\
&\quad\leq \mathbb{P}\left[\sup_{h\in\mathcal{H},t\in\mathbb{R}}\frac{\mathbb{P}[\ell(h(\boldsymbol{x}),y) > t] - \hat{\mathbb{P}}[\ell(h(\boldsymbol{x}),y) > t]}{\sqrt{\mathbb{P}[\ell(h(\boldsymbol{x}),y)]}} > \epsilon\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\mathbb{E}[\ell(h(\hat{\boldsymbol{x}}),y)] - \mathbb{E}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\hat{\mathbb{E}}}[\ell^2(h(\boldsymbol{x}),y)]} > \epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right] \\
&\quad\leq \mathbb{P}\left[\sup_{h\in\mathcal{H},t\in\mathbb{R}}\frac{\hat{\mathbb{P}}[\ell(h(\boldsymbol{x}),y) > t] - \mathbb{P}[\ell(h(\boldsymbol{x}),y) > t]}{\sqrt{\hat{\mathbb{P}}[\ell(h(\boldsymbol{x}),y)]}} > \epsilon\right].
\end{aligned}$$

Proof. The proof of the first statement is provided. The second statement can be shown in very similar way. Fix $\epsilon > 0$ and assume that for any $h \in \mathcal{H}$ and $t \geq 0$, the following holds.

$$\frac{\mathbb{P}[\ell(h(\boldsymbol{x}),y) > t] - \hat{\mathbb{P}}[\ell(h(\boldsymbol{x}),y) > t]}{\sqrt{\mathbb{P}[\ell(h(\boldsymbol{x}),y)]}} \leq \epsilon. \tag{A.3}$$

We show that this implies that for any $h \in \mathcal{H}$,

$$\frac{\mathbb{E}[\ell(h(\boldsymbol{x}),y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\mathbb{E}}[\ell^2(h(\boldsymbol{x}),y)]} > \epsilon\sqrt{2+\log\frac{1}{\epsilon}}.$$

By the properties of the Lebesgue integral, we can write

$$\mathbb{E}[\ell(h(\boldsymbol{x}),y)] = \int_0^{+\infty}\mathbb{P}[\ell(h(\boldsymbol{X}),y) > t]dt, \tag{A.4}$$

and

$$\hat{\mathbb{E}}[\ell(h(\boldsymbol{x}), y)] = \int_0^{+\infty} \hat{\mathbb{P}}[\ell(h(\boldsymbol{x}), y) > t]dt, \tag{A.5}$$

and similarly,

$$\mathbb{E}[\ell^2(h(\boldsymbol{x}), y)] = \int_0^{+\infty} \mathbb{P}[\ell^2(h(\boldsymbol{x}), y) > t]dt \tag{A.6}$$

$$= \int_0^{+\infty} 2t\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]dt. \tag{A.7}$$

In what follows, we use the shorter notation

$$I = \mathbb{E}[\ell^2(h(\boldsymbol{x}), y)].$$

Let $t_1 = \sqrt{\frac{1}{2}}\frac{1}{\epsilon}$. To bound $\mathbb{E}[\ell(h(\boldsymbol{x}), y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}), y)]$, we simply bound $\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t] - \hat{\mathbb{P}}[\ell(h(\boldsymbol{x}), y) > t]$ for large values of $t > t_1$, and use Eq. (A.4) for smaller values of t.

$$\mathbb{E}[\ell(h(\boldsymbol{x}), y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}), y)] = \int_0^{+\infty} \mathbb{P}[\ell(h(\boldsymbol{x}), y) > t] - \hat{\mathbb{P}}[\ell(h(\boldsymbol{x}), y) > t]dt \tag{A.8}$$

$$\leq \int_0^{+\infty} \epsilon\sqrt{\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]}dt + \int_0^{+\infty} \mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]dt. \tag{A.9}$$

For relatively small values of t, $\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]$ is close to one. Thus, if we define t_0 by $t_0 = \sqrt{\frac{1}{2}}$, we can write

$$\begin{aligned}
&\mathbb{E}[\ell(h(\boldsymbol{x}), y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}), y)] \\
&\quad \leq \int_0^{t_0} \epsilon dt + \int_{t_0}^{t_1} \epsilon\sqrt{\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]}dt + \int_{t_1}^{+\infty} \mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]dt && \text{(A.10)} \\
&\quad = \int_0^{+\infty} f(t)g(t)dt, && \text{(A.11)}
\end{aligned}$$

with

$$f(t) = \begin{cases} (2I)^{1/4}\epsilon & (0 \leq t \leq t_0) \\ \sqrt{2t\mathbb{P}[\ell(h(\boldsymbol{x}), y)] > t}\epsilon & (t_0 \leq t \leq t_1) \\ \sqrt{2t\mathbb{P}[\ell(h(\boldsymbol{x}), y)]}\epsilon & (t_1 \leq t), \end{cases} \tag{A.12}$$

and

$$g(t) = \begin{cases} \frac{1}{(2I)^{1/4}} & (0 \leq t \leq t_0) \\ \frac{1}{\sqrt{2t}} & (t_0 \leq t \leq t_1) \\ \sqrt{\frac{\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]}{2t}}\frac{1}{\epsilon} & (t_1 \leq t). \end{cases} \tag{A.13}$$

Now, by the Cauchy-Schwarz inequality,

$$\mathbb{E}[\ell(h(\boldsymbol{x}), y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}), y)] \leq \sqrt{\int_0^{+\infty} f(t)^2 dt} \sqrt{\int_0^{+\infty} g(t)^2 dt}. \tag{A.14}$$

The first integral on the right-hand side can be bounded as follows.

$$\begin{aligned}
\int_0^{+\infty} f(t)^2 dt &= \int_0^{t_0} \sqrt{2I}\epsilon^2 dt + \int_{t_0}^{+\infty} 2t\mathbb{P}[\ell(h(\boldsymbol{x}), y)]\epsilon^2 dt && \text{(A.15)} \\
&\leq \sqrt{2I} t_0 \epsilon^2 + \epsilon^2 I && \text{(A.16)} \\
&= 2\epsilon^2 I, && \text{(A.17)}
\end{aligned}$$

and since $t_1/t_0 = 1/\epsilon$, the second one can be computed and bounded as

$$\begin{aligned}
\int_0^{+\infty} g(t)^2 dt &= \int_0^{t_0} \frac{1}{\sqrt{2I}} dt + \int_{t_0}^{t_1} \frac{1}{2t} dt + \int_{t_1}^{+\infty} \frac{\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]}{2t\epsilon^2} && \text{(A.18)} \\
&= \frac{1}{2} + \frac{1}{2}\log\frac{1}{\epsilon} + \int_{t_1}^{+\infty} \frac{2t\mathbb{P}[\ell(h(\boldsymbol{x}), y)]}{4t^2\epsilon^2} dt && \text{(A.19)} \\
&\leq \frac{1}{2} + \frac{1}{2}\log\frac{1}{\epsilon} + \int_{t_1}^{+\infty} \frac{2t\mathbb{P}[\ell(h(\boldsymbol{x}), y) > t]}{4t_1^2\epsilon^2} dt && \text{(A.20)} \\
&\leq \frac{1}{2} + \frac{1}{2}\log\frac{1}{\epsilon} + \frac{1}{4t_1^2\epsilon^2} && \text{(A.21)} \\
&= 1 + \frac{1}{2}\log\frac{1}{\epsilon}. && \text{(A.22)}
\end{aligned}$$

Combining the bounds obtained for these integrals yields directly

$$\begin{aligned}
\mathbb{E}[\ell(h(\boldsymbol{x}), y)] - \mathbb{E}[\ell(h(\hat{\boldsymbol{x}}), y)] &\leq \sqrt{2\epsilon^2 I \left(1 + \frac{1}{2}\log\frac{1}{\epsilon}\right)} && \text{(A.23)} \\
&= t\sqrt{\left(2 + \log\frac{1}{\epsilon}\right)}\sqrt{I}, && \text{(A.24)}
\end{aligned}$$

which concludes the proof of the theorem. □

Theorem A.4 (Anthony and Shawe-Taylor [15])**.** *Let ℓ be the binary classification loss. Then, for any hypothesis class $\mathcal{H}$ of real-valued functions, the following holds.*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \frac{\mathcal{R}(h;p) - \hat{\mathcal{R}}(h;S_n)}{\sqrt{\mathcal{R}(h;p)}} > \epsilon\right] \leq 4\Pi_{\mathcal{H}}(2n)\exp\left(-\frac{n\epsilon^2}{4}\right), \tag{A.25}$$

where $\Pi_{\mathcal{H}}(n)$ is the value of the growth function (maximum number of classifications) for a sample of size n, using the hypothesis class $\mathcal{H}$.

From this theorem,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\hat{\mathcal{R}}(h;S_n)-\mathcal{R}(h;p)}{\sqrt{\hat{\mathcal{R}}(h;S_n)}}>\epsilon\right]\leq 4\Pi_{\mathcal{H}}(2n)\exp\left(-\frac{n\epsilon^2}{4}\right). \tag{A.26}$$

Combining these results with Theorem A.3 yields directly the following.

Theorem A.5. *Let $\mathcal{H}$ be a hypothesis class of real-valued functions and ℓ a loss function (not necessarily bounded) such that for all $h\in\mathcal{H}$, $0<\mathbb{E}[\ell^2(h(\boldsymbol{x}),y)]<+\infty$. Then, the following holds.*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\mathbb{E}[\ell(h(\boldsymbol{x}),y)]-\hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\mathbb{E}}[\ell^2(h(\boldsymbol{x}),y)]}>\epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right]\leq 4\Pi_{\mathcal{H}}(2n)\exp\left(-\frac{n\epsilon^2}{4}\right), \tag{A.27}$$

and

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]-\mathbb{E}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\hat{\mathbb{E}}}[\ell^2(h(\boldsymbol{x}),y)]}>\epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right]\leq 4\Pi_{\mathcal{H}}(2n)\exp\left(-\frac{n\epsilon^2}{4}\right). \tag{A.28}$$

Theorem A.6. *Let $\mathcal{H}$ be a hypothesis class of real-valued functions and ℓ a loss function (not necessarily bounded) such that for all $h\in\mathcal{H}$, $0<\mathbb{E}[\ell^2(h(\boldsymbol{x}),y)]<+\infty$. Assume that $\mathrm{Pdim}(\{\ell(h(\boldsymbol{x}),y)\mid h\in\mathcal{H}\})=d<\infty$. Then, the following holds.*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\mathbb{E}[\ell(h(\boldsymbol{x}),y)]-\hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\mathbb{E}}[\ell^2(h(\boldsymbol{x}),y)]}>\epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right]\leq 4\exp\left(d\log\frac{2\epsilon n}{d}-\frac{n\epsilon^2}{4}\right) \tag{A.29}$$

Proof. The results follows immediately by Sauer's lemma and the fact that the VC dimension of the family $\{\mathrm{sgn}(\ell(h(\boldsymbol{x}),y)-t)\mid h\in\mathcal{H},t\in\mathbb{R}\}$ is precisely the pseudo-dimension of $\{\ell(h(\boldsymbol{x}),y)\mid h\in\mathcal{H}\}$. □

Corollary A.7. *Let $\mathcal{H}$ be a hypothesis class of real-valued functions and ℓ a loss function (not necessarily bounded) such that for all $h\in\mathcal{H}$, $0<\mathbb{E}[\ell^2(h(\boldsymbol{x}),y)]<+\infty$. Assume that $\mathrm{Pdim}(\{\ell(h(\boldsymbol{x}),y)\mid h\in\mathcal{H}\})=d<\infty$. Then, the following holds.*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\frac{\mathbb{E}[\ell(h(\boldsymbol{x}),y)]-\hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]}{\sqrt{\mathbb{E}}[\ell^2(h(\boldsymbol{x}),y)]}>\epsilon\right]\leq 4\exp\left(d\log\frac{2\epsilon n}{d}-\frac{n\epsilon^{8/3}}{4^{5/3}}\right) \tag{A.30}$$

Corollary A.8. *Let $\mathcal{H}$ be a hypothesis class of real-valued functions and ℓ a loss function (not necessarily bounded) such that for all $h\in\mathcal{H}$, $0<\mathbb{E}[\ell^2(h(\boldsymbol{x}),y)]<+\infty$. Assume that $\mathrm{Pdim}(\{\ell(h(\boldsymbol{x}),y)\mid h\in\mathcal{H}\})=d<\infty$. Then, for any $\delta>0$, with probability at least*

$1-\delta$, *for any* $h \in \mathcal{H}$, *the following holds.*

$$\left|\mathbb{E}[\ell(h(\boldsymbol{x}),y)] - \hat{\mathbb{E}}[\ell(h(\boldsymbol{x}),y)]\right| \leq 2^{5/4} \max\left\{\sqrt{\mathbb{E}[\ell^2(h(\boldsymbol{x}),y)]}, \sqrt{\hat{\mathbb{E}}[\ell^2(h(\boldsymbol{x}),y)]}\right\} \sqrt[3/8]{\frac{p\log\frac{2me}{p} + \log\frac{4}{\delta}}{n}}. \tag{A.31}$$

Theorem A.9. *Let* $\mathcal{H}$ *be a hypothesis class such that* $\mathrm{Pdim}(\{\ell(h(\boldsymbol{x}),y) \mid h \in \mathcal{H}\}) = p < \infty$. *Assume that* $2^{D_R^\alpha[p_{te}\|p_{tr}]} < +\infty$ *and* $w(\boldsymbol{x}) \neq 0$ *for all* $\boldsymbol{x}$. *Then, for any* $\delta > 0$, *with probability at least* $1-\delta$,

$$\mathcal{R}(h;p_{te}) \leq \hat{\mathcal{R}}_w(h;S_n) + 2^{5/4}\sqrt{2^{D_R^\alpha[p_{te}\|p_{tr}]}} \sqrt[3/8]{\frac{p\log\frac{2me}{p} + \log\frac{4}{\delta}}{n}}, \tag{A.32}$$

where $\hat{\mathcal{R}}_w = \mathbb{E}[w(\boldsymbol{x})\ell(h(\boldsymbol{x}),y)]$.

Proof. Since $2^{D_R^\alpha[p_{te}\|p_{tr}]} < +\infty$, the second moment of $w(\boldsymbol{x})\ell(\boldsymbol{x},y)$ is finite and upper bounded by $2^{D_R^\alpha[p_{te}\|p_{tr}]}$. Thus, by Corollary A.7,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \frac{\mathcal{R}(h;p) - \hat{\mathcal{R}}_w(h;S_n)}{\sqrt{2^{D_R^\alpha[p_{te}\|p_{tr}]}}} > \epsilon\right] \leq 4\exp\left(d\log\frac{2\epsilon n}{d} - \frac{n\epsilon^{8/3}}{4^{5/3}}\right), \tag{A.33}$$

where d is the pseudo-dimension of the function class $\mathcal{H}'' = \{w(\boldsymbol{x})\ell(h(\boldsymbol{x}),y) \mid h \in \mathcal{H}\}$. We now show that $d = \mathrm{Pdim}(\{\ell(h(\boldsymbol{x}),y) \mid h \in \mathcal{H}\})$. Let $\mathcal{H}'$ denote $\{\ell(h(\boldsymbol{x}),y) \mid h \in \mathcal{H}\}$. Let $A = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ be a set shattered by $\mathcal{H}''$. Then, there exist real numbers $r_1, r_2, \ldots, r_k$ such that for any subset $B \subseteq A$ there exists $h \in \mathcal{H}$ such that

$$\forall \boldsymbol{x}_i \in B, \quad w(\boldsymbol{x}_i)\ell(h(\boldsymbol{x}_i),y) \geq r_i, \tag{A.34}$$

and

$$\forall \boldsymbol{x}_i \in A \setminus B, \quad w(\boldsymbol{x}_i)\ell(h(\boldsymbol{x}_i),y) < r_i. \tag{A.35}$$

Since by assumption $w(\boldsymbol{x}_i) > 0$ for all $i \in [1,k]$, this implies that

$$\forall \boldsymbol{x}_i \in B, \quad \ell(h(\boldsymbol{x}_i),y) \geq r_i/w(\boldsymbol{x}_i), \tag{A.36}$$

and

$$\forall \boldsymbol{x}_i \in A \setminus B, \quad \ell(h(\boldsymbol{x}_i),y) < r_i/w(\boldsymbol{x}_i). \tag{A.37}$$

Thus, $\mathcal{H}'$ shatters A with the witness $s_i = r_i/w(\boldsymbol{x}_i)$. Using the same observations, it is straightforward to see that conversely, any set shattered by $\mathcal{H}'$ is shattered by $\mathcal{H}''$. □

Appendix B

Other Related Literature

In this chapter, we introduce additional related literature. In particular, this chapter summarizes research related to distribution shift, density ratios, and importance weighting. Figure B.1 shows the overview of applications of importance weighting. We can see that importance weighting has a wide range of applications, including distribution shift adaptation, active learning, model calibration, and label noise correction. Therefore, we provide a cross-disciplinary literature review with importance weighting as the central focus.

B.1 Distribution Shift Adaptation

B.1.1 Covariate Shift

The covariate shift assumption was reported by Shimodaira [260] and is formulated as follows:

$$p_{tr}(\boldsymbol{x}) \neq p_{te}(\boldsymbol{x}),$$
$$p_{tr}(y|\boldsymbol{x}) = p_{te}(y|\boldsymbol{x}).$$

Under this assumption, we can see that the unbiasedness of Empirical Risk Minimization is not satisfied. This issue is known to be solvable by utilizing the density ratio between

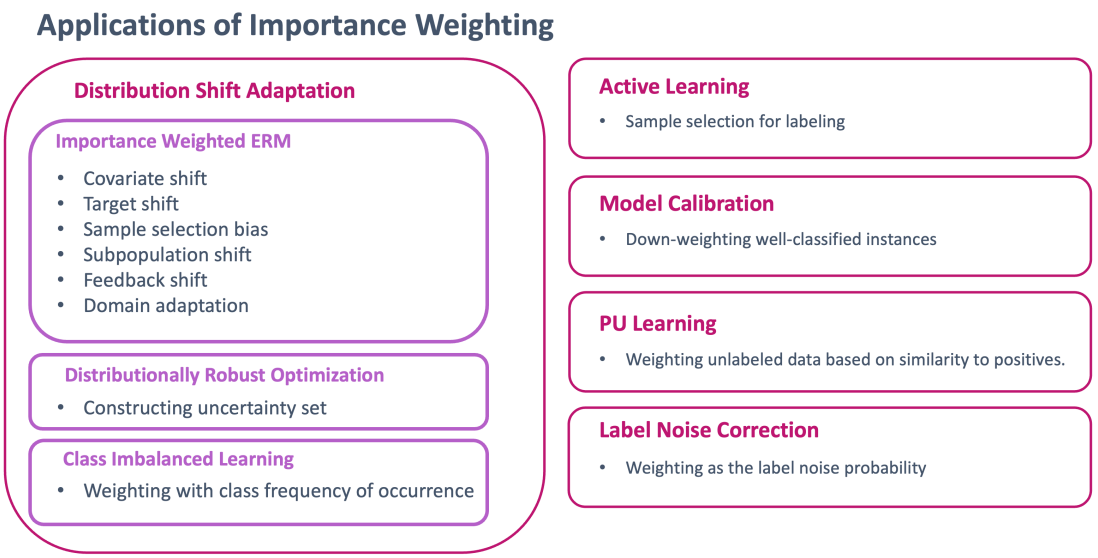

FIGURE B.1: Applications of importance weighting.

the learning distribution and the testing distribution, as follows.

$$\begin{aligned}
\mathbb{E}_{p_{tr}(\boldsymbol{x},y)}\left[\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\ell(h(\boldsymbol{x}),y)\right] &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\ell(h(\boldsymbol{x}),y)\cdot p_{tr}(\boldsymbol{x},y)d\boldsymbol{x}dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\ell(h(\boldsymbol{x}),y)\cdot p_{tr}(\boldsymbol{x})p_{tr}(y|\boldsymbol{x})d\boldsymbol{x}dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} p_{te}(\boldsymbol{x})\ell(h(\boldsymbol{x}),y)\cdot p_{tr}(y|\boldsymbol{x})d\boldsymbol{x}dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} p_{te}(\boldsymbol{x})\ell(h(\boldsymbol{x}),y)\cdot p_{te}(y|\boldsymbol{x})d\boldsymbol{x}dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}),y)\cdot p_{te}(\boldsymbol{x},y)d\boldsymbol{x}dy \\
&= \mathbb{E}_{p_{te}(\boldsymbol{x},y)}\left[\ell(h(\boldsymbol{x}),y)\right]. \qquad \text{(B.1)}
\end{aligned}$$

This is called the Importance Weighted Empirical Risk Minimization (IWERM) [260]. IWERM is theoretically valid, but it has been reported to be unstable in numerical experiments. To stabilize IWERM, Adaptive Importance Weighted Empirical Risk Minimization (AIWERM) [260], which replaces importance weighting with $\left(\frac{p_{te}(\boldsymbol{x})}{p_{tr}(\boldsymbol{x})}\right)^{\lambda}$, $\lambda \in [0,1]$, has been proposed, and its effectiveness has been demonstrated. Relative Importance Weighted Empirical Risk Minimization (RIWRM) [322] is another alternative to IWERM, achieving stability by using $\left(\frac{p_{te}(\boldsymbol{x})}{(1-\lambda)p_{tr}(\boldsymbol{x})+\lambda p_{te}(\boldsymbol{x})}\right)$ as importance weights, where $\lambda \in [0,1]$.

Importance weighting is also known to be effective in model selection. Cross-validation is an essential and practical strategy for model selection but is subject to bias under

covariate shift. Let

$$\mathcal{R}^{(n)} \coloneqq \mathbb{E}_{S_n,(\boldsymbol{x},\boldsymbol{y})}\left[\ell(h_{S_n}(\boldsymbol{x}), y)\right], \tag{B.2}$$

where h_{S_n} is a hypothesis trained by a sample S_n of sample size n. The Leave-One-Out Cross Validation (LOOCV) estimate $\hat{\mathcal{R}}^{(n)}_{\mathrm{LOOCV}}$ of the risk $\hat{\mathcal{R}}^{(n)}$ is

$$\hat{\mathcal{R}}^{(n)}_{\mathrm{LOOCV}} \coloneqq \frac{1}{n}\sum_{j=1}^{n} \ell(h_{S_n^j}(\boldsymbol{x}_j), y_j), \tag{B.3}$$

where $S_n^j = \{(\boldsymbol{x}_i, y_i)\}_{i\neq j}^n$. It is known that, if $p_{tr} = p_{te}$, LOOCV gives an almost unbiased estimate of the risk [191, 264] as

$$\mathbb{E}_{S_n}\left[\hat{\mathcal{R}}^{(n)}_{\mathrm{LOOCV}}\right] = \mathcal{R}^{(n-1)} \tag{B.4}$$

$$\approx \mathcal{R}^{(n)}. \tag{B.5}$$

However, this property is no longer true under the covariate shift. Importance Weighted Cross Validation (IWCV) [277], a variant of the cross-validation that applies importance weighting, overcomes this problem. The Leave-One-Out Importance-Weighted Cross Validation (LOOIWCV) is given as

$$\hat{\mathcal{R}}^{(n)}_{\mathrm{LOOCV}} \coloneqq \frac{1}{n}\sum_{j=1}^{n} \frac{p_{te}(\boldsymbol{x}_j)}{p_{tr}(\boldsymbol{x}_j)}\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j). \tag{B.6}$$

We can see that the validation error in the cross-validation procedure is weighted according to the importance of input vectors. For any $j \in \{1, \ldots, n\}$, we have

$$\begin{aligned}
\mathbb{E}_{S_n}\left[\frac{p_{te}(\boldsymbol{x}_j)}{p_{tr}(\boldsymbol{x}_j)}\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j)\right] &= \mathbb{E}_{S_n^j,\boldsymbol{x}_j,y_j}\left[\frac{p_{te}(\boldsymbol{x}_j)}{p_{tr}(\boldsymbol{x}_j)}\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j)\right] \\
&= \mathbb{E}_{S_n^j,y_j}\left[\int_{\mathcal{X}} \frac{p_{te}(\boldsymbol{x}_j)}{p_{tr}(\boldsymbol{x}_j)}\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j)p_{tr}(\boldsymbol{x}_j)d\boldsymbol{x}_j\right] \\
&= \mathbb{E}_{S_n^j,y_j}\left[\int_{\mathcal{X}} p_{te}(\boldsymbol{x}_j)\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j)d\boldsymbol{x}_j\right] \\
&= \mathbb{E}_{S_n^j,y}\left[\int_{\mathcal{X}} p_{te}(\boldsymbol{x})\ell(h_{S_n^j}(\boldsymbol{x}), y)d\boldsymbol{x}\right] \\
&= \mathbb{E}_{S_n^j,\boldsymbol{X},y}\left[\ell(h_{S_n^j}(\boldsymbol{x}), y)\right] \\
&= \mathcal{R}^{(n-1)} \\
&\approx \mathcal{R}^{(n)},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{S_n}\left[\hat{\mathcal{R}}^{(n-1)}_{\mathrm{LOOIWCV}}\right] &= \mathbb{E}_{S_n}\left[\frac{1}{n}\sum_{j=1}^{n}\frac{p_{te}(\boldsymbol{x}_j)}{p_{tr}(\boldsymbol{x}_j)}\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j)\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{S_n}\left[\frac{p_{te}(\boldsymbol{x}_j)}{p_{tr}(\boldsymbol{x}_j)}\ell(h_{S_n^j}(\boldsymbol{x}_j), y_j)\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathcal{R}^{(n-1)} \\
&= \mathcal{R}^{(n-1)} \\
&\approx \mathcal{R}^{(n)}.
\end{aligned}$$

Then, we can see that LOOIWCV gives an almost unbiased estimate of the risk even under the covariate shift. The above proof can be easily extended to an importance-weighted version of k-fold cross validation.

In contrast to IWERM, which weights the training data by the density ratio $p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$, robust algorithms, derived from the context of Distributionally Robust Optimization, weight the test data by $p_{tr}(\boldsymbol{x})/p_{te}(\boldsymbol{x})$ [56, 178, 179]. Martin et al. [192] proposed Double-Weighting Covariate Shift Adaptation, which performs weighting by considering a trade-off between both of these weighting strategies.

For another problem setting, Tibshirani et al. [286] demonstrated the effectiveness of importance weighted conformal prediction under covariate shift.

However, Kouw and Loog [154] experimentally report that they examined various estimates of IWERM, and all of them underestimate the expected risk. Also, classical results reported the utility of importance weighting for the case when the model is parametric. Gogolashvili et al. [90] demonstrated the necessity of importance weighting even in cases where the model is non-parametric and misspecified.

B.1.2 Target Shift

Target shift [337] scenario assumes that the target marginal distribution changes between the training and testing phases, while the target-conditioned covariate distribution remains unchanged.

$$\begin{aligned}
p_{tr}(y) &\neq p_{te}(y), \\
p_{tr}(\boldsymbol{x}|y) &= p_{te}(\boldsymbol{x}|y).
\end{aligned}$$

When the target variable is continuous, further challenges and research directions are being considered. Chan and Ng [51] uses the Expectation-Maximization (EM) method to infer the distribution $p(y)$, but their approach also involves the estimation of $p(\boldsymbol{x}|y)$. Nguyen et al. [213] propose a method for estimating importance weights $p_{te}(y)/p_{tr}(y)$ to address the shift in continuous target variables in a semi-supervised setting. Black Box Shift Estimation (BBSE) [177] estimates importance weights by solving the equation $\boldsymbol{C}\boldsymbol{w} = b$ using the confusion matrix $\boldsymbol{C}$ of a black box predictor f (where b represents the predictor's average output).

B.1.3 Sample Selection Bias

While closely related to the covariate shift assumption, many studies also treat similar problem settings as sample selection bias [28, 291, 300, 311].

Several studies [292, 334] introduced a random variable s to model sample selection bias, and considered the construction of the test distribution as follows:

$$p_{te}(x, y) = p(x, y) = \sum_{s} p(x, y, s), \tag{B.7}$$

where $s = 1$ means the example is selected, and $s = 0$ means the example is not selected. Under such assumptions, they demonstrate the effectiveness of importance weighting where $w(\boldsymbol{x}) = \frac{P(s=1)}{P(s=1|\boldsymbol{x})}$.

However, in scenarios with significant sample selection bias, Liu and Ziebart [178] point out that the errors associated with importance weighting can become excessively large.

B.1.4 Subpopulation Shift

Subpopulation shift [255, 326] in machine learning refers to a situation where the statistical properties of a particular subset or subpopulation of data change between the training and testing phases. This means that the distribution of data within a specific subgroup, such as a certain demographic, location, or any other defined subset, is different in the training and testing datasets.

Many studies propose to reweight the sample inverse proportionally to the subpopulation frequencies [47, 66, 185]. Alternatively, UMIX [107] is a variant of mixup data augmentation to estimate importance weights and achieve improved performance of the over-parameterized model under subpopulation shift.

B.1.5 Feedback Shift

In the advertising industry, predicting how many clicks are associated with purchases is crucial. However, in the real world, it is known that there is a time lag between clicks and actual purchases. This problem is referred to as a feedback shift or delayed feedback problem. Delayed feedback was initially researched by Chapelle [52] and has since led to the proposal of numerous techniques to address this problem [157, 250, 283, 330]. Furthermore, several studies have explored the concept of delayed feedback in the context of bandit problems [135, 237, 301].

Due to feedback shift, there are often cases where data that should have received positive labels is treated as negative. For example, due to delayed feedback, at the time the training instances were collected, purchases had not yet occurred. Yasui et al. [327] apply importance weighting to tackle the problem of feedback shift, called feedback shift importance weighting (FSIW).

B.2 Domain Adaptation

Domain adaptation [25, 65, 83, 232, 304, 310] in machine learning refers to the process of adapting a machine learning model trained on one domain (source domain) to make accurate predictions on another, often related but different, domain (target domain). Domains can differ in various ways, such as the distribution of data, feature representations, or the underlying data generation processes.

There is a substantial body of research that leverages the importance of weighting for domain adaptation. The main idea among these is to assign high weights to source domain data that are close to the target domain data. Lee et al. [169] proposed a method that approximates importance weighting using a distance kernel and applies it to domain adaptation. Mashayekhi [193] proposed estimating importance weighting through adversarial learning, where a generator is trained to assign weights to learning instances that deceive the discriminator. Xiao and Zhang [316] also proposed importance weighting that takes into account the difference in sample sizes between the source domain and the target domain. Azizzadenesheli et al. [20] derived a generalization error bound for domain adaptation using importance weighting.

Jiang and Zhai [132] utilize importance weighting for domain adaptation in NLP tasks. However, there have been reports of negative results in NLP tasks where domain adaptation based on importance weighting did not perform well [238]. This is suspected to be due to the nature of NLP task data, such as word occurrence rates and other factors. In light of these negative results, Xia et al. [314] pointed out that existing domain

adaptation strategies based on importance weighting primarily address sample selection bias but do not effectively resolve sample selection variance. To address this issue, they proposed several weighting methods that take into account the trade-off between bias and variance.

Additionally, there are subfields within general domain adaptation, such as the following:

- Multi-source domain adaptation: utilizes multiple source domains to achieve domain adaptation.
- Partial domain adaptation: the target domain has less number of classes compared to the source domain.
- Open-set domain adaptation: there exist unknown classes in both domains.
- Universal domain adaptation: requires no prior knowledge of the label sets.

B.2.1 Multi-Source Domain Adaptation

The goal of multi-source domain adaptation is to leverage the information from the multiple source domains to adapt the model to perform well on the target domain, even though the target domain may have a different data distribution. Sun et al. [280] propose a two-stage multi-source domain adaptation framework that computes importance weights for the instances from multiple sources to reduce both marginal and conditional probability differences between the source and target domains. Wen et al. [308] developed an algorithm that automatically explores the importance weights of data from multiple source domains.

B.2.2 Partial Domain Adaptation

Traditional domain adaptation techniques address the domain discrepancy in one of three ways: by acquiring feature representations that are invariant across domains, by producing features or samples specifically for target domains, or by converting samples between domains using generative models. They suppose that label sets are identical across domains. Partial domain adaptation problem [45, 47, 174] assumes that the target domain has less number of classes compared to the source domain. In partial domain adaptation, the target domain class space is a subspace of the source domain class space. Thus, aligning the whole source domain distribution p_{tr} and target domain distribution p_{te} is difficult.

Zhang et al. [336] proposed that introducing importance weighting into adversarial nets-based domain adaptation (IWAN) is effective for the partial domain adaptation problem. Domain adaptation methods based on adversarial neural networks [63, 93] utilize the discriminator which classifies whether input examples are from real distribution or generative distribution [186, 293]. The weighting function of IWAN is given as

$$w(\boldsymbol{x}) = 1 - D^*(\boldsymbol{x}) = \frac{1}{p_{tr}(\boldsymbol{x})/p_{te}(\boldsymbol{x}) + 1}, \tag{B.8}$$

where $D^*(\boldsymbol{x})$ is the optimal discriminator, $p_{tr}(\boldsymbol{x})$ is the source distribution and $p_{te}(\boldsymbol{x})$ is the target distribution. We can see that this weighting function is the unnormalized version of relative importance weighting (Eq. (2.34), [322]). Example Transfer Network (ETN) [46] extends IWAN by simultaneously optimizing domain-invariant features and importance weights for useful source instances. Selective Adversarial Network (SAN) [44] utilizes multiple adversarial networks with an importance weighting mechanism to select outsource examples in the outlier classes. There is also an extension of IWAN for regression and general domain adaptation[68].

B.2.3 Open-set Domain Adaptation

Recent works try to relax the assumption of identical label sets across domains by proposing open-set domain adaptation. Panareda Busto and Gall [227] introduced unknown classes in both domains while assuming that the shared classes between the two domains are identified during the training phase.

Liu et al. [183] pointed out that it is important to take into account the openness of the target domain, which is measured by the proportion of unknown classes in all target classes. Their proposed method uses a coarse-to-fine weighting mechanism to gradually distinguish between samples of unknown and known classes while assigning importance to their role in aligning feature distributions. Kundu et al. [162] proposed a metric called model inheritability to quantify how well a model trained on the source domain can perform on the target domain, and they suggest using this metric in importance weighting for the open-set domain adaptation. Garg et al. [88] also developed a method for open-set domain adaptation using the concept of Positive-Unlabeled (PU) learning [24, 148].

B.2.4 Universal Domain Adaptation

Universal domain adaptation is the most challenging problem setting in domain adaptation as it requires no prior knowledge about the label sets in the target domain. You et al. [331] have proposed the Universal Adaptation Network (UAN), which performs instance

weighting based on domain similarity and prediction uncertainty. Fu et al. [87] leverage model calibration techniques to enhance the effectiveness of importance weighting on source domain data based on the uncertainty in universal domain adaptation.

B.3 Active Learning

Active learning [115, 257] is a framework based on the belief that if a learner can select instances for labeling, the required sample size to achieve sufficient performance can be reduced. There are numerous methods within the realm of active learning, based on uncertainty [171, 172, 324, 344], diversity [4, 39, 325], representativeness [72, 124, 147], reducing expected error [136, 210] or maximizing expected model changes [41, 86, 137]. Among them, there is a subset of techniques that particularly focus on density ratios and importance weighting.

Importance Weighted Active Learning (IWAL) [33] is a seminal work that integrates importance weighting into the framework of active learning. IWAL assigns labels to unlabeled instances $\boldsymbol{x}_t$ based on their features and the history of labeled data, with a probability p_t. The weight of the instance is then regarded as $1/p_t$ during the learning process. In this strategy, the probability p_t is determined using the hypothesis space that includes the most divisive f and g in terms of achieving reasonable accuracy on the training data at time t, denoted as H_t. That is,

$$p_t = \max_{f,g \in H_{t+1}} \max_{y} \sigma(\ell(f(\boldsymbol{x}_t), y) - \ell(g(\boldsymbol{x}_t), y)), \tag{B.9}$$

$$H_{t+1} = \{h \in H_t; L_t(h) \leq L_t^* + \Delta_t\}, \tag{B.10}$$

where $L_t(h)$ and L_t^* are the empirical loss and the achievable minimum loss at time t, and Δ_t stands for some small quantity. IWAL has been proven to exhibit consistency through such importance weighting. Additionally, experimental results have shown its effectiveness. This framework has been reported to be effective in a wide range of applications, including image recognition [313], point cloud segmentation [312], sleep monitoring [120], and battery health management [55]. Beygelzimer et al. [34] proposed practical instantiations of IWAL by introducing a rejection threshold based on a deviation bound and reported that the IWAL framework can produce the reusable sample. Here, sample reusability [287, 288] is a concept that considers whether a labeled sample created in an active learning framework using one classifier can successfully train another classifier. However, Van Tulder [297] reported that it is impossible for the IWAL framework to produce reusable samples in later work.

In the context of model misspecification, similar discussions have been raised. Sugiyama [275] demonstrated that existing active learning methods can tolerate slight model misspecification and that importance weighting can handle larger misspecifications. Bach [21] investigated the asymptotic properties of active learning in generalized linear models under model misspecification and proposed efficient instance selection based on importance weighting according to their findings.

Active Learning by Learning (ALBL) [121] combines various instance selection algorithms for active learning using the multi-armed bandit framework to choose the most appropriate one for a given task. ALBL extends the IWAL for providing a proper reward function, called Importance Weighted Accuracy, for the multi-armed bandit algorithm within it.

One common challenge in active learning is the difficulty of performance evaluation. Active learning algorithms continuously select the most informative instances at each step to label, which can lead to a labeled dataset with a distribution that differs from the original data distribution. This is indeed a case of sampling bias. Kottke et al. [153] reported that the reweighted cross-validation, naturally induced by IWAL, cannot completely resolve these distribution differences. Zhao et al. [340] demonstrate that combining importance weighting with class-balanced sampling can reduce sample complexity for arbitrary label shifts. Mohamad et al. [203] proposed a stream-based active learning method using importance weighting under concept drift assumption. Su et al. [273] introduce the domain adaptation method that combines domain adversarial learning and active learning. In this method, importance weighting is utilized by approximating the density ratio using the output of the discriminator in adversarial learning. Furthermore, active testing [84, 152] is a known method for model evaluation using loss-proportional sampling.

Moreover, in active learning, it is crucial to consider the variability in annotator quality. In scenarios like crowdsourcing, where annotators may have varying levels of expertise, addressing the issue of annotator quality is indeed crucial. Zhao et al. [341] have proposed a pool-based active learning approach based on the concept of IWAL, which is robust to annotation noise.

Another challenge in active learning is the issue of computational cost. Agarwal et al. [3] utilize the IWAL framework to parallelize the exploration of data that should be labeled.

B.4 Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) [31, 69, 75, 91, 170, 242] aims to reduce the expected loss under the worst-case scenario, considering a probabilistic uncertainty set $\mathcal{U}(p_0)$ derived from observed samples and characterized by specific known properties of the actual data-generation distribution. That is, the goal of the DRO problem is

$$\text{minimize}_{h\in\mathcal{H}}\ \mathcal{R}(h;p_0) \coloneqq \sup_{q\in\mathcal{U}(p_0)} \mathbb{E}_{(\boldsymbol{x},y)\sim q(\boldsymbol{x},y)}\Big[\ell(h(\boldsymbol{x}),y)\Big], \tag{B.11}$$

for the hypothesis class $\mathcal{H}$ and the uncertainty set $\mathcal{U}(p_0)$ for training distribution p_0. The literature considers several uncertainty sets [26, 30, 38, 81], and aims to learn the model providing good performance against perturbations to the data-generating distribution. The problem of DRO can also be viewed as considering a worst-case evaluation for unknown distribution shifts.

One interpretation of the worst-case formulation in DRO is as importance-weighted loss minimization without a known target domain. For instance, by using some set of weighting functions W, we can construct the uncertainty set as

$$\mathcal{U}_w(p_0) = \{w(\boldsymbol{x}) \cdot p_0(\boldsymbol{x}); w \in W\}. \tag{B.12}$$

Duchi and Namkoong [76] provided the DRO framework by upweighting tail values of $\ell(h(\boldsymbol{x}),y)$ and giving a worst-case objective that focuses on hard regions of input domain $\mathcal{X}$. Awad and Atia [18] constructed an uncertainty set using the importance weighted source domain distribution as its center and demonstrated that it has a high probability of encompassing the true target domain distribution. As a practical optimization method for deep neural network models under data imbalance, a variant of momentum SGD called Attentional-Biased Stochastic Gradient Descent (ABSGD) [240] has been proposed. ABSGD applies importance weighting to each sample in the mini-batch and inherits many assets of the theoretical analysis of DRO.

Additionally, Agarwal and Zhang [2] have introduced Minimax Regret Optimization (MRO) as a variant of the DRO problem formulation. They have proven that for both DRO and MRO, the upper bound of their error is determined by the variance of the weighting.

B.5 Model Calibration

In many machine learning models, the output probabilities can be interpreted as confidence in their predictions. However, these probabilities are not always well-calibrated, meaning they may not accurately reflect the likelihood of the predicted outcomes. For example, a model might assign a probability of 0.7 to an event, but the true frequency of that event occurring is closer to 0.9. Model calibration typically involves adjusting the output probabilities through various techniques that aim to bring the predicted probabilities in line with the observed frequencies of events in the data [32, 77, 104, 114, 295].

One of the most famous methods for model calibration is focal loss. Focal loss was originally proposed for object detection [173, 176, 208, 290] in computer vision and was shown to be effective for model calibration in subsequent studies [207]. Focal loss tackles the model calibration problem by down-weighting the easy-to-classify examples (those with high confidence predictions) during training, effectively giving more emphasis to the difficult-to-classify examples. This is achieved through a modulating factor that is applied to the standard cross-entropy loss. This procedure can be viewed as an importance weighting $w(\boldsymbol{x}_i) = (1 - p_i)^\gamma$ that depends on the predictive probabilities p_i of the model for an instance $\boldsymbol{x}_i$ and $\gamma \in [0, 1]$.

The simplicity of the idea has led to the existence of many variants of focal loss. Cyclical focal loss [263] is a variant of the focal loss inspired by curriculum learning [27, 96, 106, 305], which applies weighting to each instance based on its classification difficulty during each epoch. AdaFocal [89] leverages the calibration characteristics of the focal loss and dynamically adjusts the parameter γ parameter for various sample groups in response to the model's previous iteration and its tendency to be over-confidence, as observed on the validation set. Dual focal loss [284] improves upon the focal loss by incorporating a balance between over-confidence and under-confidence. Adversarial Focal loss (AFL) [182] utilizes a separate adversarial network to produce a difficulty score for each input.

In addition, some theoretical results on focal loss have been reported. Charoenphakdee et al. [53] proved that the optimal classifier trained with the focal loss can achieve the Bayes-optimal classifier. Another study [212] showed that focal loss reduces the curvature of the loss surface of the model.

B.6 Positive-Unlabelled (PU) learning

PU learning, which stands for Positive-Unlabeled learning, is a type of machine learning paradigm where the goal is to build a classifier when you have a dataset with positive examples and unlabeled examples [24, 148]. The key challenge in PU learning is that you don't have labeled negative examples, which are typically used to train binary classifiers. In PU learning, the positive class is well-defined, but the negative class is composed of both true negatives and unlabeled examples. The main objective is to develop algorithms that can learn a reliable classifier despite the absence of labeled negatives. This is particularly useful in situations where obtaining a representative sample of negatives is difficult or costly.

Elkan and Noto [79] who first proposed the concept of PU Learning built their algorithm based on the following lemma.

Lemma B.1 ([79])**.** *Let $\boldsymbol{x}$ be an example and let $y \in \{0, 1\}$ be a binary label. Let $s = 1$ if the example $\boldsymbol{x}$ is labeled, and let $s = 0$ if $\boldsymbol{x}$ is unlabeled. Then, for the selected completely at random unlabeled example $\boldsymbol{x}$, we have*

$$p(y = 1|\boldsymbol{x}) = \frac{p(s = 1|\boldsymbol{x})}{p(s = 1|y = 1)}. \tag{B.13}$$

Proof. From the assumption, we have

$$p(s = 1|y = 1, \boldsymbol{x}) = p(s = 1|y = 1). \tag{B.14}$$

We also have

$$\begin{aligned} p(s = 1|\boldsymbol{x}) &= p(y = 1 \wedge s = 1|\boldsymbol{x}) \\ &= p(y = 1|\boldsymbol{x})p(s = 1|y = 1, \boldsymbol{x}) \\ &= p(y = 1|\boldsymbol{x})p(s = 1|y = 1). \end{aligned} \tag{B.15}$$

By dividing each side by $p(s = 1|y = 1)$, we can obtain the proof. □

For the unlabeled example,

$$\begin{aligned} p(y=1|\boldsymbol{x}, s=0) &= \frac{p(s=0|\boldsymbol{x}, y=1)p(y=1|\boldsymbol{x})}{p(s=0|\boldsymbol{x})} \\ &= \frac{\left(1-p(s=1|\boldsymbol{x}, y=1)\right)p(y=1|\boldsymbol{x})}{1-p(s=1|\boldsymbol{x})} \\ &= \frac{(1-c)p(y=1|\boldsymbol{x})}{1-p(s=1|\boldsymbol{x})} \\ &= \frac{(1-c)p(s=1|\boldsymbol{x})/c}{1-p(s=1|\boldsymbol{x})} \\ &= \frac{1-c}{c}\frac{p(s=1|\boldsymbol{x})}{1-p(s=1|\boldsymbol{x})}, \end{aligned} \tag{B.16}$$

where $c = p(s=1|y=1)$. Then,

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{x},y,s)}\left[h(\boldsymbol{x}, y)\right] &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}} h(\boldsymbol{x}, y)p(\boldsymbol{x}, y, x)d\boldsymbol{x}dyds \\ &= \int_{\mathcal{X}} p(\boldsymbol{x})\Bigg(p(s=1|\boldsymbol{x})h(\boldsymbol{x}, 1) \\ &\quad + p(s=0|\boldsymbol{x})\Big(p(y=1|\boldsymbol{x}, s=0)h(\boldsymbol{x}, 1) + p(y=0|\boldsymbol{x}, s=0)h(\boldsymbol{x}, 0)\Big)\Bigg). \end{aligned}$$

The plugin estimator of $\mathbb{E}_{p(\boldsymbol{x},y,s)}[h(\boldsymbol{x}, y)]$ is

$$\frac{1}{n_{tr}}\left(\sum_{\boldsymbol{x},s=1} h(\boldsymbol{x}, 1) + \sum_{(\boldsymbol{x},s=0)} w(\boldsymbol{x})h(\boldsymbol{x}, 1) + (1-w(\boldsymbol{x}))h(\boldsymbol{x}, 0)\right), \tag{B.17}$$

where

$$\begin{aligned} w(\boldsymbol{x}) &= p(y=1|\boldsymbol{x}, s=0) \\ &= \frac{1-c}{c}\frac{p(s=1|\boldsymbol{x})}{1-p(s=1|\boldsymbol{x})}. \end{aligned} \tag{B.18}$$

Thus, PU learning can be regarded as importance weighting according to the probability of labeling unlabeled data.

Because of this theoretical justification, many PU Learning studies have adopted this weighting strategy [57, 73, 74, 148, 181]. Gu et al. [99] propose Entropy Weight Allocation (EWA), which is an instance-dependent weighting method. EWA constructs an optimal transport from unlabeled instances to labeled instances and uses the transport distance to perform weighting. Acharya et al. [1] are applying the recently popularized learning framework called contrastive learning [142, 285] in the PU learning setting.

B.7 Label Noise Correction

Label noise correction [215, 216, 267] in machine learning refers to the process of identifying and rectifying errors or inaccuracies in the labeled training data used for supervised learning tasks. Label noise correction methods can be especially useful when dealing with datasets where manual labeling is error-prone or when using data from diverse sources with varying levels of accuracy.

As one of the most prominent methods for label noise correction, Patrini et al. [233] proposed a technique that uses a matrix corresponding to the probability of label noise occurrence to weight the loss function. This method can be viewed as applying importance weighting that downweights instances with a high probability of label noise occurrence. Noise Attention Learning [188] automatically models the probability of label noise using attention and employs it for weighting during gradient updates. In a similar vein, Attentive Feature Mixup (AFM) [235] combines an attention mechanism with mixup, a well-known data augmentation technique, to perform weighting based on the probability of label noise occurrence.

B.8 Density Ratio Estimation

Density ratio estimation is a method in machine learning employed to measure the distinction between the distributions of two sets of data. Its objective is to calculate the density ratio, denoted as $r(\boldsymbol{x})$, which expresses the ratio of the probability density functions between the test distribution $p_{te}(\boldsymbol{x})$ and the training distribution $p_{tr}(\boldsymbol{x})$, defined as $r(\boldsymbol{x}) = p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$. The estimated density ratio obtained in this manner plays a central role in the importance of weighting-based methods.

One of the most commonly used frameworks in density ratio estimation is moment matching. The core concept of moment matching involves aligning the moments of two distributions, $\hat{p}_{te}(\boldsymbol{x}) = \hat{r}(\boldsymbol{x})p_{tr}(\boldsymbol{x})$ and $p_{te}(\boldsymbol{x})$. One common example is matching their means, as shown by the equation:

$$\int \boldsymbol{x} \cdot \hat{r}(\boldsymbol{x})p_{tr}(\boldsymbol{x})d\boldsymbol{x} = \int \boldsymbol{x} \cdot p_{te}(\boldsymbol{x})d\boldsymbol{x}. \tag{B.19}$$

However, it is important to note that matching a finite number of moments does not necessarily lead to the true density ratio, even in asymptotic scenarios. Kernel mean

matching [97, 122] offers a solution by efficiently aligning all moments within the Gaussian Reproducing Kernel Hilbert Space (RKHS) $\mathfrak{H}$:

$$\min_{\hat{r}\in\mathfrak{H}} \left\| \int K(\boldsymbol{x},\cdot)\hat{r}(\boldsymbol{x})p_{tr}(\boldsymbol{x})d\boldsymbol{x} - \int K(\boldsymbol{x},\cdot)p_{tr}(\boldsymbol{x})d\boldsymbol{x} \right\|_{\mathfrak{H}}^2. \tag{B.20}$$

Kullback–Leibler Importance Estimation Procedure (KLIEP) [214, 278] minimize KL-divergence from $p_{te}(\boldsymbol{x})$ to $\hat{p}_{te}(\boldsymbol{x}) = \hat{r}(\boldsymbol{x})p_{tr}(\boldsymbol{x})$ as

$$\min_{\hat{r}} D_{KL}[p_{te}(\boldsymbol{x})\|\hat{p}_{te}] = \min_{\hat{r}} \int p_{te}(\boldsymbol{x}) \frac{p_{te}(\boldsymbol{x})}{\hat{r}(\boldsymbol{x})p_{tr}(\boldsymbol{x})} d\boldsymbol{x}.$$

Least-Squares Importance Fitting (LSIF) [139] minimize squared loss:

$$\min_{\hat{r}} \int \left(\hat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p_{tr}(\boldsymbol{x})d\boldsymbol{x}.$$

When the gap between the two distributions is substantial, the accuracy of density ratio estimation significantly deteriorates. Rhodes et al. [248] propose Telescoping Density Ratio Estimation (TRE) following a divide-and-conquer framework. TRE first generates a chain of intermediate datasets between two distributions, gradually transporting samples from the source distribution p_0 to the target distribution $q = p_m$:

$$\frac{p_0(\boldsymbol{x})}{p_m(\boldsymbol{x})} = \frac{p_0(\boldsymbol{x})p_1(\boldsymbol{x})}{p_1(\boldsymbol{x})p_2(\boldsymbol{x})} \cdots \frac{p_{m-2}(\boldsymbol{x})p_{m-1}(\boldsymbol{x})}{p_{m-1}(\boldsymbol{x})p_m(\boldsymbol{x})}. \tag{B.21}$$

Then, experimental results have shown that density ratio estimation along this chain leads to good outcomes. Moreover, Yin [329] revealed several statistical properties of TRE. Featurized Density Ratio Estimation (FDRE) [58] considers the use of invertible generative models to map two distributions to a common feature space for density ratio estimation. In this feature space, the two distributions become closer to each other in a way that allows for easy calculation of the density ratio. Furthermore, because the generative model used for mapping is invertible, it guarantees that the density ratio calculated in this common space is identical in the input space as well. Trimmed Density Ratio Estimation (TDRE) [184] is a robust method that automatically detects outliers, which can cause density ratio estimation values to become large, and mitigate their influence. Kumagai et al. [161] utilize a meta-learning framework for relative density ratio estimation.

B.9 Importance Weighting and Deep Learning

The behavior of Importance Weighted ERM in over-parametrized deep neural networks was unknown. In recent research, Byrd and Lipton [40] have reported that the effectiveness of importance weighting diminishes with prolonged iterative learning. Furthermore, they experimentally revealed that L2 regularization and batch normalization partially restore the effectiveness of importance weighting, but the extent of this effect depends on parameters unrelated to importance weighting.

They also provide an intuitive discussion that connects the experimental results obtained with the implicit bias [130, 131, 211, 269] of the gradient descent method. Xu et al. [317] took this argument further, providing theoretical support for their experimental results. From the theoretical analysis, they characterized the impact of importance weighting on the implicit bias of gradient descent and provided the generalization bound that hinges on the importance weights. They reported that, for finite finite-step training, the impact of importance weighting on the generalization bound depends on how it influences the margin and its role in balancing the source and target distributions.

Several studies have aimed at implicitly learning importance weightings using neural networks. Ren et al. [245] introduced an online meta-learning algorithm aimed at adjusting the weights of training examples, with the goal of training deep learning models that are more robust. Through experiments, they empirically demonstrated that their proposed method outperforms existing explicit weighting techniques. Fang et al. [82] reported that approximating the weighting function with a neural network induces bias. Therefore, they proposed addressing this issue by optimizing the weighting neural network and the classifier alternately. Holtz et al. [118] reported that learning through importance weighting improves robustness against adversarial attacks [50, 102, 123, 318]. They propose a method to acquire this weighting through an adversarial learning framework.

As an alternative to importance weighting, Lu et al. [189] propose importance tempering to enhance the decision boundary of over-parametrized neural networks. The central idea behind importance tempering is to introduce a temperature parameter into the softmax, which corresponds to importance weighting. They reported its effectiveness in training over-parametrized neural networks.

Appendix C

Introduction to Information Geometry

In this chapter, we begin by providing a brief introduction to Riemannian manifolds and differential geometry. Then, we provide an overview of statistical manifolds created by probability distributions, which are Riemannian manifolds with the Fisher information matrix as their metric. Finally, we introduce the dual structure and provide a concise overview of information geometry.

C.1 Riemannian Manifolds and Differential Geometry

A manifold is a multidimensional space that can be regarded as locally an Euclidean space. A point within a manifold can be identified by multiple parameter sets, where these parameters can be interpreted as local coordinate systems.

Here, tensors serve as valuable tools for describing the geometric properties of the manifold. One of the key properties of tensors regards their behavior under coordinate transformations. That is, if a tensor vanishes in one coordinate system, then it vanishes in all coordinate systems. Many geometric objects are tensors, and they are called by different names, such as metric, curvature, vector field, or 1-form.

In this section, we first provide an overview of manifold definitions and examples, tangent spaces, and differentiable mappings. Next, we introduce the definition of a Riemannian manifold and concepts such as linear connections, Levi-Civita connections, and related

notions. For more details, refer to textbooks on Riemannian manifolds and differential geometry [42, 101, 155, 158, 164, 166, 271]. The following concepts are the reproduction of known definitions and results in differential geometry.

C.1.1 Manifolds

For the construction of manifolds, we first start with a concept of metric space.

Definition C.1. Let $\mathcal{M}$ be a set of points. A distance function on $\mathcal{M}$ is defined as a mapping $d : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ with the following properties:

i) non-degenerate: $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$, for $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$;

ii) symmetric: $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$;

iii) satisfies the triangle inequality: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{M}$.

The pair $(\mathcal{M}, d)$ is called a metric space.

Example C.1. *Let $\mathcal{M} = \mathbb{R}^n$ for $n \in \mathbb{N}$. Consider $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, with $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$. Then the space $(\mathcal{M}, d_E)$ with the distance function $d_E(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^n (x_i - y_i)^2\right)^{\frac{1}{2}}$ is the metric space, and is called the Euclidean space.*

Definition C.2. Let $(\mathcal{M}, d)$ be a metric space. Consider $\boldsymbol{x} \in \mathcal{M}$ and $r > 0$. The ball of radius r and centered at $\boldsymbol{x}$ is the set $B_r(\boldsymbol{x}) = \{y \in \mathcal{M}; d(\boldsymbol{x}, \boldsymbol{y}) < r\}$. A subset U of $\mathcal{M}$ is called open if for any $\boldsymbol{x} \in U$, there is a $r > 0$ such that $B_r(\boldsymbol{x}) \subset U$.

The following proposition provides the equivalence of function continuity.

Proposition C.3. *Let $f : (\mathcal{M}, d_{\mathcal{M}}) \to (\mathcal{N}, d_{\mathcal{N}})$ be a mapping between two metric spaces. The following are equivalent:*

i) For any open set U in $\mathcal{N}$, the pullback $f^{-1}(U) = \{\boldsymbol{x} \in \mathcal{M}; f(\boldsymbol{x}) \in U\}$ is an open set in $\mathcal{M}$.

ii) For any convergent sequence $\boldsymbol{x}^{(k)} \to \boldsymbol{x}$ in $\mathcal{M}$, we have $f(\boldsymbol{x}^{(k)}) \to f(\boldsymbol{x})$.

A function $f : \mathcal{M} \to \mathcal{N}$ is called continuous if any of i) or ii) in Proposition C.3 holds. If f is invertible and both f and f^{-1} are continuous, then f is called a homeomorphism between $\mathcal{M}$ and $\mathcal{N}$.

Definition C.4. Let $U \subset \mathcal{M}$ be an open set. Then the pair (U, ϕ) is called a coordinate system on $\mathcal{M}$, if $\phi : U \to \phi(U) \subset \mathbb{R}^n$ is a homeomorphism of the open set U in $\mathcal{M}$ onto an open set $\phi(U)$ of $\mathbb{R}^n$. The coordinate functions f on U are defined as $x_j : U \to \mathbb{R}$, and $\phi(p) = (x_1(p), \ldots, x_n(p))$, namely $x_j = u_j \circ \phi$, where $u_j : \mathbb{R}^n \to \mathbb{R}$, $u_j(a_1, \ldots, a_n) = a_j$ is the j-th projection.

The integer n is called the dimension of the coordinate system.

Definition C.5. An atlas $\mathcal{A}$ of dimension n associated with the metric space $(\mathcal{M}, d)$ is a collection of coordinate systems $\{(U_\alpha, \phi_\alpha)\}_\alpha$ such that

i) $U_\alpha \subset \mathcal{M}, \forall\alpha, \bigcup_\alpha U_\alpha = \mathcal{M}$. That is, U_α covers $\mathcal{M}$.

ii) If $U_\alpha \cap U_\beta \neq \emptyset$, the restriction to $\phi_\alpha(U_\alpha \cap U_\beta)$ of the map

$$F_{\alpha\beta} = \phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta) \tag{C.1}$$

is differentiable from $\mathbb{R}^n$ to $\mathbb{R}^n$.

There exist several atlases on a given metric space $\mathcal{M}$ in general. Two atlases $\mathcal{A}$ and $\mathcal{A}'$ are called compatible if their union is an atlas on $\mathcal{M}$. The set of compatible atlases with a given atlas $\mathcal{A}$ can be partially ordered by inclusion, and its maximal element is called the complete atlas $\bar{\mathcal{A}}$.

Definition C.6. A differentiable manifold $\mathcal{M}$ is a metric space endowed with a complete atlas. The dimension n of the atlas is called the dimension of the manifold.

Example C.2. *Euclidean space $\mathbb{R}^n$ is the differentiable manifold, and the atlas has only one coordinate system, the identity map,* $\mathrm{Id} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{Id}(\boldsymbol{x}) = \boldsymbol{x}$.

Example C.3. *Any open set U of $\mathbb{R}^d$ is a differential manifold, with only one coordinate system,* (U, Id).

C.1.2 Tangent Space

First, we introduce the differentiable function on a manifold.

Definition C.7. A function $f : \mathcal{M} \to \mathbb{R}$ is said to be differentiable if for any coordinate system (U, ϕ) on $\mathcal{M}$ the function $f \circ \phi^{-1} : \phi(U) \to \mathbb{R}$ is differentiable. The set of all differentiable functions on the manifold $\mathcal{M}$ will be denoted by $\mathcal{F}(\mathcal{M})$.

From $\mathcal{F}(\mathcal{M})$, the tangent vectors on $\mathcal{M}$ is defined as follows.

Definition C.8. A tangent vector of $\mathcal{M}$ at a point $p \in \mathcal{M}$ is a function $X_p : \mathcal{F}(\mathcal{M}) \to \mathbb{R}$ such that

i) X_p is $\mathbb{R}$-linear

$$X_p(af + bg) = aX_p(f) + bX_p(g),\ \forall a, b \in \mathbb{R},\ \forall f, g \in \mathcal{F}(\mathcal{M}). \tag{C.2}$$

ii) The Leibniz rule is satisfied

$$X_p(fg) = X_p(f)g(p) + f(p)X_p(g),\ \forall f, g \in \mathcal{F}(\mathcal{M}). \tag{C.3}$$

Definition C.9. For a differentiable curve $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ on the manifold $\mathcal{M}$, with $\gamma(0) = p$, the tangent vector

$$X_p(f) = \frac{d(f \circ \gamma)}{dt}(0), \quad \forall f \in \mathcal{F}(\mathcal{M}) \tag{C.4}$$

is called the tangent vector to $\gamma(-\epsilon, \epsilon)$ at $p = \gamma(0)$ and is denoted by $\dot{\gamma}(0)$.

Here, consider the i-th coordinate curve γ. That is, there exists a coordinate system (U, ϕ) around $p = \gamma(0)$ in which $\phi(\gamma(t)) = (x_1^{(0)}, \ldots, x_i, \ldots, x_n^{(0)})$, where $\phi(p) = (x_1^{(0)}, \ldots, x_i^{(0)}, \ldots, x_n^{(0)})$. Then, the tangent vector to γ

$$\dot{\gamma}(0) = \left.\frac{\partial}{\partial x_i}\right|_p \tag{C.5}$$

is called a coordinate tangent vector at p. This can be defined equivalently as a derivation

$$\left.\frac{\partial}{\partial x_i}\right|_p (f) = \frac{\partial(f \circ \phi^{-1})}{\partial u_i}(\phi(p)), \quad \forall f \in \mathcal{F}(\mathcal{M}), \tag{C.6}$$

where $\phi = (x_1, \ldots, x_n)$ is a coordinate system around p and $u_1, \ldots, u_n$ are the coordinate functions on $\mathbb{R}^n$.

Definition C.10. The set of all tangent vectors at p to $\mathcal{M}$ is called the tangent space of $\mathcal{M}$ at p, and is denoted by $T_p\mathcal{M}$.

The tangent vector X_p acts on differentiable functions f on $\mathcal{M}$ as

$$X_p f = \sum_{i=1}^{n} X_i(p) \left.\frac{\partial f}{\partial x_i}\right|_p. \tag{C.7}$$

Definition C.11. A vector field X on $\mathcal{M}$ is a smooth map X that assigns to each point $p \in \mathcal{M}$ a vector X_p in $T_p\mathcal{M}$. For any functions $f \in \mathcal{F}(\mathcal{M})$ we define the real-valued function $(Xf)_p = X_p f$.

The set of all vector fields on $\mathcal{M}$ is denoted by $\mathcal{X}(\mathcal{M})$. In a local coordinate system, a vector field is given by $X = \sum X_i \frac{\partial}{\partial x_i}$, where the components $X_i \in \mathcal{F}(\mathcal{M})$ because they are given by $X_i = X(x_i)$, where x_i is the i-th coordinate function.

Next, given a vector field X, consider the ordinary differential equations

$$\frac{dc^k}{dt}(t) = X^k_{c(t)}, \quad q \leq k \leq n. \tag{C.8}$$

We can see that Eq. (C.8) can be solved locally around any point $x_0 = c(0)$.

Theorem C.12. *Given $x_0 \in \mathcal{M}$ and a nonzero vector field X on an open set $U \subset \mathcal{M}$, there exists an $\epsilon > 0$ such that Eq.* (C.8) *has a unique solution $c : [0, \epsilon) \to U$ satisfying $c(0) = x_0$.*

The solution $t \to c(t)$ is called the integral curve associated with the vector field X through the point x_0.

Definition C.13. A function $f \in \mathcal{F}(\mathcal{M})$ is called a first integral of motion for the vector field X if it remains constant along the integral curves of X.

Proposition C.14. *Let $f \in \mathcal{F}(\mathcal{M})$. Then f is the first integral of motion for the vector field X if and only if $X_{c(t)}(f) = 0$.*

Proof. Consider a local coordinate system $(x_1, \ldots, x_n)$ in which the vector field writes as $X = \sum_j X^j \frac{\partial}{\partial x_j}$. Then,

$$\begin{aligned} X_{c(t)}(f) &= \sum_j X^j_{c(t)} \frac{\partial f}{\partial x_j} \\ &= \sum_j \frac{dc_j}{dt}(t) \frac{\partial f}{\partial x_j} \\ &= \frac{d}{dt} f(c(t)). \end{aligned}$$

□

Next, we introduce an important operation on vector fields, called the Lie bracket.

Definition C.15. Lie bracket $[,] : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ is defined as

$$[X, Y]_p f \coloneqq X_p(Yf) - Y_p(Xf), \quad \forall f \in \mathcal{F}(\mathcal{M}), p \in \mathcal{M}. \tag{C.9}$$

The vector fields X and Y are commute if $[X, Y] = 0$. In local coordinate systems, the Lie bracket takes the form

$$[X, Y] = \sum_{i,j=1}^{n} \left(\frac{\partial Y_i}{\partial x_j} X_j - \frac{\partial X_i}{\partial x_j} Y_j \right) \frac{\partial}{\partial x_i}. \tag{C.10}$$

The Lie bracket satisfies the following properties

i) $\mathbb{R} - bilinearity$:

$$\begin{aligned} [aX + bY, Z] &= a[X, Z] + b[Y, Z], \\ [Z, aX + bY] &= a[Z, X] + b[Z, Y], \quad \forall a, b \in \mathbb{R}. \end{aligned}$$

ii) Skew-symmetry:

$$[X, Y] = -[Y, X].$$

iii) The Jacobi identity:

$$[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0.$$

iv) The Lie bracket is not $\mathcal{F}(\mathcal{M}) - linear$.

v) The Lie bracket satisfies

$$[fX, gY] = fg[X, Y] + f(Xg)Y - g(Yf)X, \quad \forall f, g \in \mathcal{F}(\mathcal{M}).$$

C.1.3 Differentiable Maps

Definition C.16. A map $F : \mathcal{M} \to \mathcal{N}$ between two manifolds $\mathcal{M}$ and $\mathcal{N}$ is differentiable about $p \in \mathcal{M}$ if for any coordinate systems (U, ϕ) on $\mathcal{M}$ about p and $(V, \psi) \in \mathcal{N}$ about $F(p)$, the map $\psi \circ F \circ \phi^{-1}$ is differentiable from $\phi(U) \subset \mathbb{R}^m$ to $\psi(V) \subset \mathbb{R}^n$.

Definition C.17. Let $F : \mathcal{M} \to \mathcal{N}$ be a differentiable map. For every $p \in \mathcal{M}$, the differential map dF at p is defined as

$$\begin{aligned} &dF_p : T_p\mathcal{M} \to T_{F(p)}\mathcal{N} \\ &(dF_p)(v)(f) = v(f \circ F), \quad \forall v \in T_p\mathcal{M}, \forall f \in \mathcal{F}(\mathcal{M}). \end{aligned} \tag{C.11}$$

The differential of a map satisfies the following properties:

i) dF_p is an $\mathbb{R}$-linear between the tangent spaces $T_p\mathcal{M}$ and $T_{F(p)}\mathcal{N}$:

$$dF_p(v+w) = dF_p(v) + dF_p(w), \quad \forall v, w \in T_p\mathcal{M};$$
$$dF_p(c \cdot v) = c \cdot dF_p(v), \quad \forall v \in T_p\mathcal{M}, \forall c \in \mathbb{R}.$$

ii) Let $\{\frac{\partial}{\partial x_j}\Big|_p\}$ and $\{\frac{\partial}{\partial y_j}\Big|_{F(p)}\}$ be bases associated with the tangent spaces $T_p\mathcal{M}$ and $T_{F(p)}\mathcal{N}$. Let $J_{kj} = \frac{\partial F_k}{\partial x_j}$ be the Jacobian matrix of the function $F = (F_1, \ldots, F_n)$ with respect to the coordinate systems $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ on $\mathcal{M}$ and $\mathcal{N}$. Then, dF_p can be represented locally by

$$dF_p\left(\frac{\partial}{\partial x_j}\Big|_p\right) = \sum_{k=1}^{n} J_{kj}(p)\ \frac{\partial}{\partial y_k}\Big|_{F(p)} \tag{C.12}$$

iii) If $\dim \mathcal{M} = \dim \mathcal{N} = n$, the following conditions are equivalent:

a) $dF_p : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$ is an isomorphism of vectorial spaces;

b) F is a local diffeomorphism in a neighborhood of p;

c) There are two coordinate systems $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ on $\mathcal{M}$ and $\mathcal{N}$ around p and on $\mathcal{N}$ around $F(p)$, such that the associated Jacobian is non-degenerate: $\det J_{kj}(p) \neq 0$.

iv) For $F : \mathcal{M} \to \mathcal{N}$, the differential dF commutes with the Lie bracket:

$$dF_p[v, w] = [dF_p(v), dF_p(w)], \quad \forall v, w \in T_p\mathcal{M}. \tag{C.13}$$

The differential of a function $f \in \mathcal{F}(\mathcal{M})$ is defined at any point p by $(df)_p : T_p\mathcal{M} \to \mathbb{R}$ as

$$(df)_p(v) = v(f), \quad \forall v \in T_p\mathcal{M}. \tag{C.14}$$

In local coordinate system $(x_1, \ldots, x_n)$, this takes the form

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i, \tag{C.15}$$

where $\{dx_i\}$ is the dual basis of $\{\frac{\partial}{\partial x_i}\}$ of $T_p\mathcal{M}$

$$dx_i\left(\frac{\partial}{\partial x_j}\right) = \delta_{ij}, \tag{C.16}$$

where δ_{ij} is the Kronecker delta. The space spanned by $dx_1, \ldots, dx_n$ is called the cotangent space of $\mathcal{M}$ at p, and is denoted by $T_p^*\mathcal{M}$. The elements of $T_p^*\mathcal{M}$ are called covectors.

Here, the differential *df* is an example of a 1-form. The 1-form ω on the manifold $\mathcal{M}$ is a mapping which assigns to each point $p \in \mathcal{M}$ an element $\omega_p \in T_p^*\mathcal{M}$. The 1-form can be written in the local coordinate system as

$$\omega = \sum_{i=1}^{n} \omega_i dx_i, \tag{C.17}$$

where $\omega_i = \omega(\frac{\partial}{\partial x_i})$ is the i-th coordinate of the form with respect to the basis $\{dx_i\}$. The set of all 1-forms on the manifold $\mathcal{M}$ is denoted by $\mathcal{X}^*(\mathcal{M})$.

C.1.4 Tensors

Let $T_p\mathcal{M}$ and $T_p^*\mathcal{M}$ be the tangent and cotangent spaces of $\mathcal{M}$ at p.

Definition C.18. A tensor of type (r, s) at $p \in \mathcal{M}$ is an $\mathcal{F}(\mathcal{M})$-multilinear function $T : (T_p^*\mathcal{M})^r \times (T_p\mathcal{M})^s \to \mathbb{R}$, where

$$(T_p^*\mathcal{M})^r = \underbrace{T_p^*\mathcal{M} \times \cdots \times T_p^*\mathcal{M}}_{r \text{ times}}, \tag{C.18}$$

$$(T_p\mathcal{M})^s = \underbrace{T_p\mathcal{M} \times \cdots \times T_p\mathcal{M}}_{s \text{ times}}. \tag{C.19}$$

A tensor field $\mathcal{T}$ of type (r, s) is a differential map, which assigns each point $p \in \mathcal{M}$ an (r, s)-tensor $\mathcal{T}_p$ on $\mathcal{M}$ at the point p.

The tensor field $\mathcal{T}$ is written by using local coordinate systems as

$$\mathcal{T} = \mathcal{T}^{i_i i_2 \ldots i_r}_{j_1 j_2 \ldots j_s} dx^{j_1} \otimes \cdots \otimes dx^{j_s} \otimes \frac{\partial}{\partial x^{j_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{i_r}}, \tag{C.20}$$

where $\otimes$ is the usual tensorial product operation, and $\{dx^{j_1 \otimes \cdots \otimes dx^{j_r}}\}_{j_1 < \cdots < j_r}$ and $\{\frac{\partial}{\partial x^{i_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{i_s}}\}_{i_1 < \cdots < i_s}$ are bases in the vectorial spaces $(T_p^*\mathcal{M})^r$ and $(T_p\mathcal{M})^s$.

The tensor field $\mathcal{T}$ acts on r 1-forms and a vector fields as

$$\begin{aligned}\mathcal{T}(\omega^1, \ldots, \omega^r, X_1, \ldots, X_s) &= T^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_s} dx^{j_1}(X_1) \otimes \cdots \otimes dx^{j_s}(X_s) \otimes \frac{\partial}{\partial x^{i_1}}(\omega^1) \otimes \cdots \otimes \frac{\partial}{\partial x^{i_r}}(\omega^r) \\ &= \mathcal{T}^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_s} X_1^{j_1} \ldots X_s^{j_s} \omega^1_{i_1} \ldots \omega^r_{i_r}.\end{aligned}$$

We say that the tensor $\mathcal{T}$ is s covariant and r contravariant.

C.1.5 Riemannian Manifolds

Definition C.19. A Riemannian metric g on a differentiable manifold $\mathcal{M}$ is a symmetric, positive definite 2-covariant tensor field. A Riemannian manifold is a differentiable manifold $\mathcal{M}$ endowed with a Riemannian metric g.

A Riemannian manifold is denoted by the pair $(\mathcal{M}, g)$. The Riemannian metric g can be considered as a positive definite scalar product $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ that depends differentially on the point $p \in \mathcal{M}$. In local coordinate systems, the Riemannian metric g is written as

$$g = g_{ij} dx^i dx^j, \tag{C.21}$$

with $g_{ij} = g(\partial_i, \partial_j)$.

A metric g induces a natural bijective correspondence between 1-forms and vector fields on a Riemannian manifold $\mathcal{M}$.

C.1.6 Linear Connections and Levi-Civita Connection

A linear connection allows the differentiation of a function, a vector field, or a tensor with respect to a given field.

Definition C.20. A linear connection ∇ on a differentiable manifold $\mathcal{M}$ is a map $\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ with the following properties:

i) $\nabla_X Y$ is $\mathcal{F}(\mathcal{M})$-linear in X;

ii) $\nabla_X Y$ is $\mathbb{R}$-linear in Y;

iii) it satisfies the Leibniz rule:

$$\nabla_X(fY) = (Xf)Y + f\nabla_X Y, \quad \forall f \in \mathcal{F}(\mathcal{M}). \tag{C.22}$$

For fixed vector fields X and Y, $\nabla_X Y$ is also a vector field on $\mathcal{M}$. In a local coordinate system $(x^1, \ldots, x^n)$ we can write

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k, \tag{C.23}$$

where Γ_{ij}^k are the coordinates of the connection with respect to the local base $\{\partial_i\}$, where $\partial_i = \frac{\partial}{\partial x^i}$. For $X = X^i\partial_i$ and $Y = Y^j\partial_j$, we have

$$\nabla_X Y = (\nabla_X Y)^k \partial_k, \tag{C.24}$$

where $(\nabla_X Y)^k = X^i(\partial_i Y^k + Y^j \Gamma_{ij}^k)$.

The differentiation of r-covariant tensor field $\mathcal{T}$ along a vector field X with respect to the linear connection ∇ as

$$(\nabla_X \mathcal{T})(Y_1, \ldots, Y_r) = X\mathcal{T}(Y_1, \ldots, Y_r) - \sum_{i=1}^{n} \mathcal{T}(Y_1, \ldots, \nabla_X Y_i, \ldots, Y_r). \tag{C.25}$$

Definition C.21. Let g be the Riemannian metric tensor. A linear connection ∇ is called metric connection if g is parallel with respect to ∇,

$$\nabla_X g = 0, \quad \forall X \in \mathcal{X}(\mathcal{M}). \tag{C.26}$$

Equivalently,

$$Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y), \quad \forall X, Y, Z \in \mathcal{X}(\mathcal{M}). \tag{C.27}$$

A linear connection is described by two other tensors, torsion and curvature.

Definition C.22. For a linear connection ∇, the torsion is defined as

$$\begin{aligned} &T : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M}) \\ &T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]. \end{aligned} \tag{C.28}$$

The torsion measures the noncommutativity of the derivation with respect to two vector fields. From the fact that $T(X, Y) = -T(Y, X)$, T is a 2-covariant skew-symmetric tensor. In the local coordinate system, we have

$$\begin{aligned} T_{ij} &= T(\partial_i, \partial_j) \\ &= \nabla_{\nabla_i} \partial_j - \nabla_{\partial_j} \partial_i - [\partial_i, \partial_j] \\ &= \left(\Gamma_{ij}^k - \Gamma_{ji}^k\right) \partial_k, \end{aligned}$$

and the torsion coordinates are given by $T_{ij}^k = \Gamma_{ij}^k - \Gamma_{ji}^k$.

A connection ∇ is called torsion-free if $T = 0$, or $\Gamma_{ij}^k = \Gamma_{ji}^k$.

Definition C.23. The curvature of the linear connection ∇ is given by

$$\begin{aligned} &R : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M}) \\ &R(X, Y, Z) = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \end{aligned} \tag{C.29}$$

It can be shown that the curvature R satisfies the following properties:

$$R(f_1 X, f_2 Y, f_3 Z) = f_1 f_2 f_3 R(X, Y, Z) \tag{C.30}$$

$$R(X_1 + X_2, Y, Z) = R(X_1, Y, Z) + R(X_2, Y, Z) \tag{C.31}$$

$$R(X, Y_1 + Y_2, Z) = R(X, Y_1, Z) + R(X, Y_2, Z) \tag{C.32}$$

$$R(X, Y, Z_1 + Z_2) = R(X, Y, Z_1) + R(X, Y, Z_2), \tag{C.33}$$

for all $f_1, f_2, f_3 \in \mathcal{F}(\mathcal{M})$ and $X, X_1, X_2, Y, Y_1, Y_2, Z, Z_1, Z_2 \in \mathcal{X}(\mathcal{M})$. R is a 3-covariant tensor field and skew-symmetric in the first pair of arguments: $R(X, Y, Z) = -R(Y, X, Z)$. The curvature tensor R is also denoted by $R(X, Y)Z$, and

$$R\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) \frac{\partial}{\partial x^k} = R^l_{ijk} \frac{\partial}{\partial x_l}. \tag{C.34}$$

in a local coordinate system.

Definition C.24. The Ricci curvature associated with the linear connection ∇ is given by

$$\mathrm{Ric} : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{F}(\mathcal{M}), \tag{C.35}$$

$$\mathrm{Ric}(Y, Z) = \mathrm{Trace}\left(X \to R(X, Y)Z\right). \tag{C.36}$$

It is known that there exists a unique metric connection with zero torsion on the Riemannian manifold. This metric connection $\nabla^{(0)}$ is called the Levi-Civita connection of the Riemannian manifold $(\mathcal{M}, g)$. In this section, we omit superscript and denote by ∇.

Theorem C.25. *On a Riemannian manifold, there is a unique torsion-free, metric connection ∇, which satisfies*

$$\begin{aligned} 2g(\nabla_X Y, Z) = &X g(Y, Z) + Y g(X, Z) - Z g(X, Y) \\ &+ g([X, Y], Z) - g([X, Z], Y) - g([Y, Z], X). \end{aligned} \tag{C.37}$$

Proof. First, we show that connection ∇ defined by formula C.37 is a metric and torsion-free connection. From the properties of vector fields and Lie brackets, we have

$$2g(\nabla_{fX} Y, Z) = 2f g(\nabla_X Y, Z), \quad \forall Z \in \mathcal{X}(\mathcal{M}), \tag{C.38}$$

and $\nabla_{fX}Y = f\nabla_X Y,\ \forall X, Y \in \mathcal{X}(\mathcal{M})$. Then, ∇ is $\mathcal{F}(\mathcal{M})$-linear. Also,

$$\begin{aligned}
2g(\nabla_X(fY), Z) &= Xg(fY, Z) + fYg(X, Z) - Zg(X, fY) \\
&\quad + g([X, fY], Z) - g([X, Z], fY) - g([fY, Z], X) \\
&= X(f)g(Y, Z) + fXg(Y, Z) + fYg(X, Z) - Z(f)g(X, Y) - fZg(X, Y) \\
&\quad + fg([X, Y], Z) + X(f)g(Y, Z) - fg([X, Z], Y) \\
&\quad - fg([Y, Z], X) + Z(f)g(Y, X) \\
&= 2fg(\nabla_X Y, Z) + 2X(f)g(Y, Z) \\
&= 2g(f\nabla_X Y + X(f)Y, Z).
\end{aligned}$$

Therefore, ∇ is a linear connection. Furthermore,

$$\begin{aligned}
2g(T(X, Y), Z) &= 2g(\nabla_X Y, Z) - 2g(\nabla_Y X, Z) - 2g([X, Y], Z) \\
&= Xg(Y, Z) + Yg(X, Z) - Zg(X, Y) \\
&\quad + g([X, Y], Z) - g([X, Z], Y) - g([Y, Z], X) \\
&\quad - Yg(X, Z) - Xg(Y, Z) + Zg(Y, X) \\
&\quad - g([Y, X], Z) + g([Y, Z], X) + g([X, Z], Y) \\
&\quad - 2g([X, Y], Z) \\
&= g(2[X, Y] - 2[X, Y], Z) \\
&= 0,
\end{aligned}$$

for all $X, Y, Z \in \mathcal{X}(\mathcal{M})$. From Eq. (C.37),

$$\begin{aligned}
2g(\nabla_Z, X, Y) + 2g(X, \nabla_Z Y) &= Zg(X, Y) + Xg(Z, Y) - Yg(Z, X) + g([Z, X], Y) \\
&\quad - g([Z, Y], X) - g([X, Y], Z) \\
&\quad + Zg(Y, X) + Yg(Z, X) - Xg(Z, Y) + g([Z, Y], X) \\
&\quad - g([Z, X], Y) - g([Y, X], Z) \\
&= 2Zg(X, Y).
\end{aligned}$$

Therefore, $g(\nabla_Z X, Y) + g(X, \nabla_Z Y) = Zg(X, Y)$, and ∇ is a metric connection.

Next, we prove that any metric and symmetric connection ∇ is given by Eq. (C.37). It suffices to consider in a local coordinate system $(x^1, \ldots, x^n)$. Let $X = \partial_i$, $Y = \partial_j$, $Z = \partial_k$. From $\Gamma_{ij}^k = g(\nabla_{\partial_i}\partial_j, \partial_k)$ and $g_{ij} = g(\partial_i, \partial_j)$, we have

$$2\Gamma_{ij}^p g_{pk} = \partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}. \tag{C.39}$$

Since ∇ is a metric connection,

$$\begin{aligned}\partial_i g_{jk} &= \Gamma^l_{ij} g_{lk} + \Gamma^r_{ik} g_{ir} \\ \partial_j g_{ki} &= \Gamma^l_{jk} g_{li} + \Gamma^r_{ji} g_{kr} \\ \partial_k g_{ij} &= \Gamma^l_{ki} g_{lj} + \Gamma^r_{kj} g_{ir}.\end{aligned}$$

The symmetry $\Gamma^k_{ij} = \Gamma^k_{ji}$ yields the Eq. (C.37). □

From the connection coefficient in Eq. (C.39), we have

$$\Gamma^l_{ij} = \frac{1}{2} g^{lk} \left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij} \right), \tag{C.40}$$

where (g^{lk}) denotes the inverse matrix of (g_{ij}). The coordinates Γ^l_{ij} of the Levi-Civita connection are called the Christoffel symbols of the second kind. The Christoffel symbols of the first kind are obtained as

$$\Gamma_{ij,k} = \Gamma^l_{ij} g_{lk}. \tag{C.41}$$

The curvature tensor of type $(1,3)$ associated with the Levi-Civita connection is called the Riemannian curvature tensor of type $(1,3)$. For local coordinate system, we have $R(\partial_i, \partial_j)\partial_k = R^l_{ijk}\partial_l$ and

$$R^r_{ijk} = \partial_i \Gamma^r_{jk} - \partial_j \Gamma^r_{ik} + \Gamma^r_{ih}\Gamma^h_{jk} - \Gamma^r_{jh}\Gamma^h_{ik}. \tag{C.42}$$

The $(0,4)$-type curvature tensor is given as

$$\begin{aligned}&R : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{F}(\mathcal{M}), \\ &R(X,Y,Z,W) = g(R(X,Y,Z),W),\end{aligned} \tag{C.43}$$

for all $X, Y, Z, W \in \mathcal{X}(\mathcal{M})$. In local coordinate system, $R(\partial_i, \partial_j, \partial_k, \partial_l) = R_{ijkl}$, and $R_{ijkl} = R^m_{ijkl} g_{ml}$.

C.2 Dual Connections and Dualistic Structure

In this section, we describe the concepts of dual connections, dual curvature, or other geometric objects.

C.2.1 Dual Connections

In this section, we introduce the concept of dual connections. Dual connections were first introduced by A.P. Norden in affine differential geometry literature under the name of conjugate connections [261]. They had also been independently introduced by Nagaoka and Amari in the context of information geometry [9].

Definition C.26. Let $(\mathcal{M}, g)$ be a Riemannian manifold. Two linear connections ∇ and ∇^* are called dual, with respect to the metric g, if

$$Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y), \quad \forall X, Y, Z \in \mathcal{X}(\mathcal{M}). \tag{C.44}$$

In local coordinate system $(x^1, \ldots, x^n)$, the duality condition can be expressed as

$$\partial_{x^k} g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}, \tag{C.45}$$

where $\Gamma_{ki,h} = g_{jm}\Gamma^m_{ki}$ and $\Gamma^*_{kj,i} = g_{im}\Gamma^{*m}_{kj}$ are the coordinate components of connections ∇ and ∇^*.

The triple (g, ∇, ∇^*) is called a dualistic structure on $\mathcal{M}$. A statistical manifold is a Riemannian manifold endowed with a dualistic structure, and it is a quadruple $(\mathcal{M}, g, \nabla, \nabla^*)$. Here, we introduce several important results of dual connections.

Proposition C.27. *Given a linear connection ∇ on the Riemannian manifold $(\mathcal{M}, g)$, there is a unique connection ∇^* dual to ∇.*

Proof. Let $\Gamma_{ij,l}$ be the components of the connection ∇, $\Gamma^*_{ij,l} = \partial_{x^i} g_{ij} - \Gamma_{il,j}$ and $\Gamma^{*k}_{ij} = \Gamma^*_{ij,l} g^{lk}$. Then, we construct the dual connection as

$$\nabla^*_{\partial_{x^i}} \partial_{x^j} = \Gamma^{*k}_{ij} \partial_{x^k}, \tag{C.46}$$

and this leads to the existence.

Also, from Eq. (C.45), the connection components $\Gamma^*_{kj,i}$ of ∇^* are uniquely determined by the metric g_{ij} and the connection coefficients $\Gamma_{ki,j}$ of ∇. □

Proposition C.28. *The following properties hold:*

i) $(\nabla^*)^* = \nabla$.

ii) ∇ *is a metrical connection on the Riemannian manifold* $(\mathcal{M}, g)$ *if and only if* $\nabla = \nabla^*$.

Proposition C.29. *Let R and R^* be the curvature tensors of ∇ and ∇^*. Then,*

i) $g(R(X,Y)Z,W) + g(R^*(X,Y)W,Z) = 0, \quad \forall X,Y,Z,W \in \mathcal{X}(\mathcal{M})$.

ii) $(\mathcal{M}, g, \nabla)$ *has constant curvature if and only if* $(\mathcal{M}, g, \nabla^*)$ *has same constant curvature.*

A connection ∇ is called flat in a given local coordinate system if its components vanish, $\Gamma_{ij}^k = 0$. Therefore, for two vector fields $X = X^i\partial_i$ and $Y = Y^j\partial_j$, the covariant derivative with respect to a flat connection is

$$\nabla_X Y = X^i(\partial_i Y^k + Y^j\Gamma_{ij}^k)\partial_k = X^i\partial_i Y^k\partial_k = X(Y^k)\partial_k. \tag{C.47}$$

This suggests that the torsion and the curvature of a flat connection are zero.

A statistical manifold $(\mathcal{S}, g, \nabla, \nabla^*)$ is called dually flat if both dual connections ∇ and ∇^* are flat.

Proposition C.30. *For a dually flat statistical manifold $(\mathcal{S}, g, \nabla, \nabla^*)$, we have the following properties.*

i) In any local coordinate system, the metric coefficients g_{ij} are constant.

ii) If γ is either a ∇- or ∇^*-autoparallel curve, $\gamma^k(t) = \alpha^k t + \beta^k$, with some constants α^k and β^k.

The following definition can be seen as an extension of the Levi-Civita connection.

Definition C.31. The connection ∇ is dual to ∇^* in strong sense, with respect to the metric g, if

i) $Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z^* Y)$,

ii) $\nabla_X Y - \nabla_Y^* X = [X,Y], \quad \forall X,Y,Z \in \mathcal{X}(\mathcal{M})$.

For torsion tensors T and T^* associated with the connections ∇ and ∇^*, we have

$$\begin{aligned} T(X,Y) + T^*(X,Y) &= \nabla_X Y - \nabla_Y X - [X,Y] + \nabla_X^* Y - \nabla_Y^* X - [X,Y] \\ &= (\nabla_X Y - \nabla_Y^* X - [X,Y]) - (\nabla_Y X - \nabla_X^* Y - [Y,X]) \\ &= 0. \end{aligned} \tag{C.48}$$

Proposition C.32. *If ∇ has a dual connection ∇^* in the strong sense, then its torsion is zero.*

Proposition C.33. *If ∇ and ∇^* are dual in the strong sense, then $T = T^* = 0$.*

The Levi-Civita connection ∇ is a strong auto-dual, that is, $\nabla^* = \nabla$ in the strong sense.

Theorem C.34. *If ∇ and ∇^* are dual connections in the strong sense, then $\nabla = \nabla^*$, and hence ∇ is the Levi-Civita connection.*

C.2.2 Relative Torsion Tensors

Let ∇ and ∇^* be the two connections, which are not dual in strong sense. Let

$$U(X,Y) = \nabla_X Y - \nabla^*_Y X - [X,Y]. \tag{C.49}$$

Since U is $\mathbb{R}$-linear in both variables, then for any smooth function $f \in \mathcal{F}(\mathcal{M})$, we have

$$\begin{aligned} U(fX,Y) &= \nabla_{fX} Y - \nabla^*_Y(fX) - [fX,Y] \\ &= f\nabla_X Y - f\nabla^*_Y X - Y(f)X - (fXY - Y(f)X - fYX) \\ &= f\left(\nabla_X Y - \nabla^*_Y X - [X,Y]\right) \\ &= fU(X,Y), U(X,fY) \qquad\qquad = fU(X,Y). \end{aligned} \tag{C.50}$$

Then, the mapping U is a $(1,2)$-type tensor on $\mathcal{M}$. The dual tensor to U is defined by

$$U^*(X,Y) = \nabla^*_X Y - \nabla_Y X - [X,Y]. \tag{C.51}$$

The tensors U and U^* are called the relative torsion tensors of connections ∇ and ∇^*.

Proposition C.35. *For the relative torsion tensors U and U^*, the following properties hold.*

i) $U^*(X,Y) = -U(Y,X)$.

ii) $U^{**} = U$.

iii) $U + U^* = T + T^*$, *where T and T^* are the torsions of ∇ and ∇^*.*

Proof. For any vector fields $X, Y \in \mathcal{X}(\mathcal{M})$,

i)

$$\begin{aligned} U^*(X,Y) &= \nabla^*_X Y - \nabla_Y X - [X,Y] \\ &= -\left(\nabla_Y X - \nabla^*_X Y + [X,Y]\right) \\ &= -U(X,Y). \end{aligned}$$

ii)

$$\begin{aligned} U^* * (X,Y) &= (-U(Y,X))^* \\ &= -U^*(Y,X) \\ &= U(X,Y). \end{aligned}$$

iii)

$$\begin{aligned} U(X,Y) &= \nabla_X Y - \nabla^*_Y X - [X,Y], \\ U^*(X,Y) &= \nabla^*_X Y - \nabla_Y X - [X,Y], \\ (U+U^*)(X,Y) &= (\nabla_X Y - \nabla_Y X - [X,Y]) + (\nabla^* XY - \nabla^*_Y X - [X,Y]) \\ &= (T+T^*)(X,Y). \end{aligned}$$

□

Corollary C.36. *We have $U + U^* = 0$ if and only if $T + T^* = 0$.*

Proposition C.37. *For dual connections ∇ and ∇^* with $U = 0$, we have $\nabla = \nabla^*$ and then ∇ is the Levi-Civita connection.*

Proposition C.38. *Let ∇ and ∇^* be two dual torsion-free connections, $T = T^* = 0$. Then the following relation holds*

$$\begin{aligned} g(U^*(X,Y),Z) &= g(U^*(X,Z),Y) \\ &= g(U^*(Z,Y),X) \\ &= g(U^*(Y,X),Z). \end{aligned}$$

C.2.3 Dual α-Connections

Consider the one-parameter family of connections given by the convex combination of two dual torsion-free connections ∇ and ∇^* with respect to the metric g,

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^* + \frac{1-\alpha}{2}\nabla, \quad \alpha \in \mathbb{R}. \tag{C.52}$$

The connection $\nabla^{(\alpha)}$ is called the α-connection.

Proposition C.39. *$\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual connections with respect to the metric g.*

Proof. From the definition and duality of connections, we have

$$\begin{aligned} g(\nabla_Z^{(\alpha)}X, Y) &= \frac{1+\alpha}{2}g(\nabla_Z^* X, Y) + \frac{1-\alpha}{2}g(\nabla_Z X, Y) \\ &= \frac{1+\alpha}{2}\left(Zg(X,Y) - g(X, \nabla_Z Y)\right) + \frac{1-\alpha}{2}g(\nabla_Z X, Y) \\ &= \frac{1+\alpha}{2}Zg(X,Y) - \frac{1+\alpha}{2}g(X, \nabla_Z Y) + \frac{1-\alpha}{2}g(\nabla_Z X, Y). \end{aligned}$$

Similarly,

$$g(\nabla_Z^{(-\alpha)}Y, X) = \frac{1-\alpha}{2}Zg(Y,X) - \frac{1-\alpha}{2}g(Y, \nabla_Z X) + \frac{1+\alpha}{2}g(\nabla_Z Y, X).$$

Then, we have

$$g(\nabla_Z^{(\alpha)}X, Y) + g(\nabla_Z^{(-\alpha)}Y, X) = Zg(X,Y).$$

□

Here, we have the following relations for the connection coefficients

$$\begin{aligned} \Gamma_{ij,k}^{(\alpha)} &= \frac{1+\alpha}{2}\Gamma_{ij,k}^* + \frac{1-\alpha}{2}\Gamma_{ij,k} \\ &= \frac{1+\alpha}{2}\left(\partial_{x^i} g_{jk} - \Gamma_{ik,j}\right) + \frac{1-\alpha}{2}\Gamma_{ij,k} \\ &= \frac{1+\alpha}{2}\partial_{x^i} g_{jk} - \frac{1+\alpha}{2}\Gamma_{ik,j} + \frac{1-\alpha}{2}\Gamma_{ij,k}. \end{aligned} \tag{C.53}$$

For the flat connection ∇ with $\Gamma_{ij,k} = 0$, the coefficients of the α-connection is

$$\Gamma_{ij,k}^{(\alpha)} = \frac{1+\alpha}{2}\partial_{x^i} g_{jk}. \tag{C.54}$$

Also, for flat ∇^*, we have $\Gamma_{ki,j} = \partial_{x^k} g_{ij}$ and

$$\Gamma_{ij,k}^{(\alpha)} = \frac{1+\alpha}{2}\partial_{x^i} g_{jk} - \frac{1+\alpha}{2}\partial_{x^i} g_{kj} + \frac{1-\alpha}{2}\partial_{x^i} g_{jk} \tag{C.55}$$

$$= \frac{1-\alpha}{2}\partial_{x^i} g_{jk}. \tag{C.56}$$

Proposition C.40. *The α-connection can be written in one of the following equivalent forms*

$$\nabla^{(\alpha)} = (1-\alpha)\nabla^{(0)} + \alpha\nabla^{(1)} \tag{C.57}$$
$$= (1+\alpha)\nabla^{(0)} - \alpha\nabla^{(-1)} \tag{C.58}$$
$$= \nabla^{(0)} + \frac{1}{2}\alpha(\nabla^{(1)} - \nabla^{(-1)}). \tag{C.59}$$

Proof. From a straightforward computation, we have

$$\begin{aligned}
\nabla^{(\alpha)} &= \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)} \\
&= \frac{1+\alpha}{2}\nabla^{(1)} + \left(\frac{1-\alpha}{2}\nabla^{(-1)} + \frac{1-\alpha}{2}\nabla^{(1)}\right) - \frac{1-\alpha}{2}\nabla^{(1)}, \\
&= \alpha\nabla^{(1)} + (1-\alpha)\nabla^{(0)} \\
\nabla^{(\alpha)} &= \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)} \\
&= \left(\frac{1+\alpha}{2}\nabla^{(1)} + \frac{1+\alpha}{2}\nabla^{(-1)}\right) - \frac{1+\alpha}{2}\nabla^{(-1)} + \frac{1-\alpha}{2}\nabla^{(-1)} \\
&= (1+\alpha)\nabla^{(0)} - \alpha\nabla^{(-1)}, \\
\nabla^{(\alpha)} &= \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)} \\
&= \frac{1}{2}(\nabla^{(1)} + \nabla^{(-1)}) + \frac{\alpha}{2}(\nabla^{(1)} - \nabla^{(-1)}) \\
&= \nabla^{(0)} + \frac{\alpha}{2}(\nabla^{(1)} - \nabla^{(-1)}).
\end{aligned}$$

□

C.2.4 Difference Tensor and Curvature Vector Field

For two torsion-free dual connections ∇ and ∇^*, the difference tensor is defined as

$$K(X,Y) \coloneqq \nabla^*_X Y - \nabla_X Y. \tag{C.60}$$

The difference tensor K is symmetric:

$$\begin{aligned}
K(X,Y) - K(Y,X) &= (\nabla^*_X Y - \nabla^*_Y X) + (\nabla_Y X - \nabla_X Y) \\
&= [X,Y] + [Y,X] \\
&= 0.
\end{aligned} \tag{C.61}$$

The difference tensor K can be also expressed in terms of the Levi-Civita connection $\nabla^{(0)}$ as

$$K(X,Y) = 2(\nabla^{(0)}_X Y - \nabla_X Y) = 2(\nabla^*_X Y - \nabla^{(0)}_X Y). \tag{C.62}$$

From

$$\begin{aligned}\nabla^{(-\alpha)} - \nabla^{(\alpha)} &= \left(\frac{1-\alpha}{2}\nabla + \frac{1+\alpha}{2}\nabla^*\right) - \left(\frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^*\right) \\ &= \alpha(\nabla^* - \nabla) \\ &= \alpha K,\end{aligned}$$

we can write

$$K(X,Y) = \frac{1}{\alpha}(\nabla^{(-\alpha)}_X Y - \nabla^{(\alpha)}_X Y). \tag{C.63}$$

Taking $\alpha \to 0$ yields

$$K(X,Y) = -2\,\frac{d}{d\alpha}\nabla^{(\alpha)}\Big|_{\alpha=0}. \tag{C.64}$$

C.2.5 Skewness Tensor

The skewness tensor on the Riemannian manifold $(\mathcal{M}, g)$ with respect to the linear connection ∇ is defined by $C = \nabla g$, that is

$$C(X,Y,Z) \coloneqq (\nabla g)(X,Y,Z) \tag{C.65}$$

$$= Xg(Y,Z) - g(\nabla_X Y, Z) - g(Y, \nabla_X Z), \tag{C.66}$$

and in a local coordinate system, we have

$$C_{ijk} = \partial_{x^i} g_{jk} - \Gamma_{ij,k} - \Gamma_{ik,j}, \tag{C.67}$$

where $C_{ijk} = C(\partial_{x^i}, \partial_{x^j}, \partial_{x^k})$.

Proposition C.41. *The skewness tensor and the difference tensor are related by*

$$C(X,Y,Z) = g(K(X,Y), Z). \tag{C.68}$$

Proposition C.42. *The skewness tensor is totally symmetric,*

$$C_{ijk} = C_{ikj} = C_{jik} = C_{jki} = C_{kij} = C_{kji}. \tag{C.69}$$

Proposition C.43. *We have the following relation*

$$g(\nabla_X^{(\alpha)} Y, Z) = g(\nabla_X^{(0)} Y, Z) + \frac{\alpha}{2} C(X, Y, Z). \tag{C.70}$$

C.3 Statistical Models and Statistical Manifolds

Information geometry aims to geometrically describe statistical models [8, 13, 14, 19, 221]. In this section, we first introduce the two central classes of statistical models, namely the exponential family and the mixture family. Next, we demonstrate that the Fisher information matrix becomes the Riemannian metric, and we construct geometric objects such as connections, skewness tensors, and autoparallel curves on the statistical manifold.

C.3.1 Statistical Models

C.3.1.1 Exponential Family

Consider $n+1$ real-valued smooth functions $C(\boldsymbol{x})$, $F_i(\boldsymbol{x})$ on $\mathcal{X} \subset \mathbb{R}^d$ such that

$$1, F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x}) \tag{C.71}$$

are linearly independent. Then, the exponential family of probability distributions is defined as

$$p(\boldsymbol{x}; \boldsymbol{\theta}) \coloneqq \exp\left\{C(\boldsymbol{x}) + \theta^i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{C.72}$$

where $\boldsymbol{x} \in \mathcal{X}$, and

$$\psi(\boldsymbol{\theta}) = \ln\left(\int_{\mathcal{X}} \exp\left\{C(\boldsymbol{x}) + \theta^i F_i(\boldsymbol{x}) d\boldsymbol{x}\right\}\right) \tag{C.73}$$

is the normalization function. Eq. (C.72) can be written equivalently as

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\theta^i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{C.74}$$

with $h(\boldsymbol{x}) = e^{C(\boldsymbol{x})}$.

The statistical manifold $\mathcal{S} = \{p(\boldsymbol{x}; \boldsymbol{\theta})\}$ is called an exponential family and $\theta^i \in \Theta$ are its natural parameters. The parameter space Θ is an open subset of $\mathbb{R}^n$, and the dimension of Θ is called the order of the exponential family.

Proposition C.44. *Assume $\mathbb{E}_{\boldsymbol{\theta}}[F_i] < \infty$, where $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ denotes the expectation with respect to $p_{\boldsymbol{\theta}}$. Then, we have*

$$\frac{\partial}{\partial \theta^j}\psi(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[F_j] \tag{C.75}$$

Proof. From the normalization $\int_{\mathcal{X}} p(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x} = 1$, we have

$$\begin{aligned}
\int_{\mathcal{X}} \frac{\partial}{\partial \theta^j} p(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x} &= 0 \\
\int_{\mathcal{X}} p(\boldsymbol{x};\boldsymbol{\theta}) \left(F_j(\boldsymbol{x}) - \frac{\partial}{\partial \theta^j}\psi(\boldsymbol{\theta})\right) d\boldsymbol{x} &= 0 \\
\int_{\mathcal{X}} p(\boldsymbol{x};\boldsymbol{\theta})F_j(\boldsymbol{x})d\boldsymbol{x} &= \frac{\partial}{\partial \theta^j}\psi(\boldsymbol{\theta}) \int_{\mathcal{X}} p(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x} \\
\mathbb{E}_{\boldsymbol{\theta}}[F_j] &= \frac{\partial}{\partial \theta^j}\psi(\boldsymbol{\theta}).
\end{aligned}$$

□

Since

$$\frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x};\boldsymbol{\theta}) = F_j(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}[F_j], \tag{C.76}$$

we have

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \theta^i} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x};\boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[(F_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}[F_i])(F_j(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}[F_j])\right] \tag{C.77}$$

$$= \mathrm{Cov}_{\boldsymbol{\theta}}(F_i, F_j). \tag{C.78}$$

Many of the usual probability distributions belong to the exponential family.

Example C.4 (Exponential distribution)**.** *For $n = 1$ and*

$$\begin{aligned}
C(x) &= 0, \\
F_1(x) &= -x, \\
\theta^1 &= \theta, \\
\psi(\theta) &= -\ln\theta,
\end{aligned}$$

we can obtain the exponential distribution

$$p(x;\theta) = \theta e^{-\theta x}. \tag{C.79}$$

Example C.5 (Gaussian distribution)**.** *For*

$$\begin{aligned} C(x) &= 0, \\ F_1(x) &= x, \\ F_2(x) &= x^2, \\ \theta^1 &= \frac{\mu}{\sigma^2}, \\ \theta^2 &= -\frac{1}{2\sigma^2}, \\ \psi(\theta) &= -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\ln\left(-\frac{\pi}{\theta^2}\right), \end{aligned}$$

we can obtain the Gaussian distribution

$$p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{C.80}$$

Example C.6 (Poisson distribution)**.** *For* $n = 1$ *and*

$$\begin{aligned} C(x) &= -\ln x!, \\ F_1(x) &= x, \\ \theta &= \ln\lambda, \\ \psi(\theta) &= \lambda = e^{\theta}, \end{aligned}$$

we can obtain the Poisson distribution

$$p(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}. \tag{C.81}$$

Example C.7 (Gamma distribution)**.** *For*

$$\begin{aligned} C(x) &= -\ln x, \\ F_1(x) &= \ln x, \\ F_2(x) &= x, \\ \theta^1 &= \alpha, \\ \theta^2 &= -\frac{1}{\beta}, \\ \psi(\theta) &= \ln(\beta^\alpha\Gamma(\alpha)) = \ln\left(\left(-\frac{1}{\theta^2}\right)^\alpha\Gamma(\theta^1)\right), \end{aligned}$$

we can obtain the Gamma distribution

$$p(x;\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}, \tag{C.82}$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \quad \alpha > 0.$$

Example C.8 (Beta distribution)**.** *For $n = 2$ and*

$$\begin{aligned} C(x) &= -\ln(x(1-x)), \\ F_1(x) &= \ln x, \\ F_2(x) &= \ln(1-x), \\ \theta^1 &= a, \\ \theta^2 &= b, \\ \psi(\theta) &= \ln B(\theta^1, \theta^2), \end{aligned}$$

we can obtain the Beta distribution

$$p(x; a, b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \tag{C.83}$$

where

$$B(a, B) = \int_0^1 t^{a-1}(1-t)^{b-1} dt, \quad a, b > 0.$$

C.3.1.2 Mixture Family

For linearly independent n smooth functions $F_i : \mathcal{X} \to \mathbb{R}$ with

$$\int_{\mathcal{X}} F_j(\boldsymbol{x}) d\boldsymbol{x} = 0, \quad 1 \le j \le n, \tag{C.84}$$

and a smooth function $C(\boldsymbol{x})$ with

$$\int_{\mathcal{X}} C(\boldsymbol{x}) d\boldsymbol{x} = 1, \tag{C.85}$$

the mixture family of probability distributions is defined as follows:

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = C(\boldsymbol{x}) + \theta^i F_i(\boldsymbol{x}). \tag{C.86}$$

The statistical manifold $\mathcal{S} = \{p(\boldsymbol{x}; \boldsymbol{\theta})\}$ is called a mixture family, and θ^i are its mixture parameters.

Proposition C.45. *For the mixture family, the following properties hold:*

i)

$$\frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{F_j(\boldsymbol{x})}{p(\boldsymbol{x}; \boldsymbol{\theta})}. \tag{C.87}$$

ii)

$$\frac{\partial}{\partial \theta^i} \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) = -\frac{F_i(\boldsymbol{x}) F_j(\boldsymbol{x})}{p(\boldsymbol{x}; \boldsymbol{\theta})^2}. \tag{C.88}$$

iii)

$$\frac{\partial}{\partial \theta^i} \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) = -\frac{\partial}{\partial \theta^i} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x}; \boldsymbol{\theta}). \tag{C.89}$$

Example C.9 (Mixture of $n+1$ distributions)**.** *For linearly independent $n+1$ probability distributions $p_1, p_2, \ldots, p_n, p_{n+1}$, and $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^n; \theta^j > 0; \sum_{j=1}^n \theta^j = 1\}$, the following weighted average is also a probability distribution*

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \theta^i p_i(\boldsymbol{x}) + \left(1 - \sum_{i=1}^n \theta^i\right) p_{n+1}(\boldsymbol{x}) \tag{C.90}$$

$$= p_{n+1}(\boldsymbol{x}) + \sum_{i=1}^n \theta^i (p_i(\boldsymbol{x}) - p_{n+1}(\boldsymbol{x})), \tag{C.91}$$

which becomes a mixture family with $C(\boldsymbol{x}) = p_{n+1}(\boldsymbol{x})$ and $F_i(\boldsymbol{x}) = p_i(\boldsymbol{x}) - p_{n+1}(\boldsymbol{x})$.

C.3.2 Fisher Metric

Riemannian metric plays a central role in Riemannian manifolds. For example, the metric tensor induces the length or angle between tangent vectors. In this section, we introduce how the Fisher information matrix behaves as a Riemannian metric.

The Fisher information matrix is defined as

$$g_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial \theta^i} \ln(\boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) \right] \tag{C.92}$$

$$= \int_{\mathcal{X}} \frac{\partial}{\partial \theta^i} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{x}; \boldsymbol{\theta}) \cdot p(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x}. \tag{C.93}$$

Proposition C.46. *The Fisher information matrix can be represented in terms of the square root of probability densities as*

$$g_{ij}(\boldsymbol{\theta}) = 4 \int_{\mathcal{X}} \frac{\partial}{\partial \theta^i} \sqrt{p(\boldsymbol{x}; \boldsymbol{\theta})} \frac{\partial}{\partial \theta^j} \sqrt{p(\boldsymbol{x}; \boldsymbol{\theta})} d\boldsymbol{x}. \tag{C.94}$$

Proof.

$$\begin{aligned} g_{ij}(\boldsymbol{\theta}) &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta^i} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial}{\partial\theta^j} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \cdot p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \frac{1}{p(\boldsymbol{x};\boldsymbol{\theta})} \frac{\partial}{\partial\theta^i} p(\boldsymbol{x};\boldsymbol{\theta}) \cdot \frac{1}{p(\boldsymbol{x};\boldsymbol{\theta})} \frac{\partial}{\partial\theta^j} p(\boldsymbol{x};\boldsymbol{\theta}) \cdot p(\boldsymbol{x};\boldsymbol{\theta}) \\ &= 4\int_{\mathcal{X}} \frac{1}{2\sqrt{p(\boldsymbol{x};\boldsymbol{\theta})}} \frac{\partial}{\partial\theta^i} p(\boldsymbol{x};\boldsymbol{\theta}) \cdot 2\sqrt{p(\boldsymbol{x};\boldsymbol{\theta})} \frac{\partial}{\partial\theta^j} p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} \\ &= 4\int_{\mathcal{X}} \frac{\partial}{\partial\theta^i} \sqrt{p(\boldsymbol{x};\boldsymbol{\theta})} \frac{\partial}{\partial\theta^j} \sqrt{p(\boldsymbol{x};\boldsymbol{\theta})} d\boldsymbol{x}. \end{aligned}$$

□

Proposition C.47. *The Fisher information matrix on any statistical model is symmetric, positive definite, and non-degenerate. That is, the Fisher information matrix is a Riemannian metric.*

Proposition C.48. *The Fisher information matrix can be written as the negative expectation of the Hessian of the log-likelihood function*

$$g_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial\theta^i} \frac{\partial}{\partial\theta^j} \ell(\boldsymbol{\theta}) \right] \tag{C.95}$$

$$= -\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial\theta^i} \frac{\partial}{\partial\theta^j} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \right]. \tag{C.96}$$

Example C.10 (Exponential family)**.** *For the exponential family*

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left\{ C(\boldsymbol{x}) + \theta^i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) \right\}, \tag{C.97}$$

the Fisher information matrix is written as

$$g_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial\theta^i} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial}{\partial\theta^j} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \right] = \mathrm{Cov}_{\boldsymbol{\theta}}(F_i, F_j),$$

$$\begin{aligned} g_{ij}(\boldsymbol{\theta}) &= -\mathbb{E} \left[\frac{\partial}{\partial\theta^i} \frac{\partial}{\partial\theta^j} \ln p(\boldsymbol{x};\boldsymbol{\theta}) \right] \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial\theta^i} \frac{\partial}{\partial\theta^j} \psi(\boldsymbol{\theta}) \right] \qquad \text{(C.98)} \\ &= \frac{\partial}{\partial\theta^i} \frac{\partial}{\partial\theta^j} \psi(\boldsymbol{\theta}). \qquad \text{(C.99)} \end{aligned}$$

C.3.3 Geometric Objects on Statistical Manifold

Theorem C.49. *The Fisher metric is invariant under reparametrizations of the sample space.*

Proof. Let X be a random variable with the sample space $\mathcal{X} \subseteq \mathbb{R}^n$ and probability density function $p(\cdot;\boldsymbol{\theta})$. Consider the invertible transform function $f : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^n$, and

denote by $\tilde{p}(\cdot;\boldsymbol{\theta})$ the probability density function associated with the random variable $Y = f(X)$. Then, we have

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{y};\boldsymbol{\theta})\left|\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}\right|. \tag{C.100}$$

Since

$$\ln \tilde{p}(\boldsymbol{y};\boldsymbol{\theta}) = \ln \tilde{p}(f(\boldsymbol{x});\boldsymbol{\theta}), \tag{C.101}$$

$$\ln p(\boldsymbol{x};\boldsymbol{\theta}) = \ln \tilde{p}(\boldsymbol{y};\boldsymbol{\theta}) + \ln\left|\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}\right|, \tag{C.102}$$

we have

$$\frac{\partial}{\partial\theta^i}\ln p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\partial}{\partial\theta^i}\ln \tilde{p}(\boldsymbol{y};\boldsymbol{\theta}). \tag{C.103}$$

Therefore,

$$\begin{aligned} g_{ij}(\boldsymbol{\theta}) &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta^i}\ln p(\boldsymbol{x};\boldsymbol{\theta})\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x}l\boldsymbol{\theta})d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta^i}\ln \tilde{p}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta^j}(\boldsymbol{y};\boldsymbol{\theta})\tilde{p}(f(\boldsymbol{x});\boldsymbol{\theta})\left|\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}\right| d\boldsymbol{x} \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta^i}\ln \tilde{p}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta^j}\ln \tilde{p}(\boldsymbol{y};\boldsymbol{\theta})\tilde{p}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\ &= \tilde{g}_{ij}(\boldsymbol{\theta}). \end{aligned}$$

□

Theorem C.50. *The Fisher metric is covariant under reparametrizations of the parameters space.*

Proof. For two coordinate systems $\boldsymbol{\theta} = (\theta^1,\ldots,\theta^n)$ and $\boldsymbol{\xi} = (\xi^1,\ldots,\xi^n)$ with the invertible relationship $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\xi})$. Let $\tilde{p}(\boldsymbol{x};\boldsymbol{\xi}) = p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{xii}))$. Then, we have

$$\frac{\partial}{\partial\xi^i}\tilde{p}(\boldsymbol{x};\boldsymbol{\xi}) = \frac{\partial\theta^k}{\partial\xi^i}\frac{\partial}{\partial\theta^k}p(\boldsymbol{x};\boldsymbol{\theta}), \tag{C.104}$$

and

$$\begin{aligned} \tilde{g}_{ij}(\boldsymbol{\theta}) &= \int_{\mathcal{X}} \frac{1}{\tilde{p}(\boldsymbol{x};\boldsymbol{\xi})}\frac{\partial}{\partial\xi^i}p(\boldsymbol{x};\boldsymbol{\xi})\frac{\partial}{\partial\xi^j}p(\boldsymbol{x};\boldsymbol{\xi})d\boldsymbol{x} \\ &= \left(\int_{\mathcal{X}} \frac{1}{p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{\xi}))}\frac{\partial}{\partial\theta^k}p(\boldsymbol{x};\boldsymbol{\theta})\frac{\partial}{\partial\theta^r}p(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x}\right)\frac{\partial\theta^k}{\partial\xi^i}\frac{\partial\theta^r}{\partial\xi^j} \\ &= g_{kr}(\boldsymbol{\xi})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{\xi})}\frac{\partial\theta^k}{\partial\xi^i}\frac{\partial\theta^r}{\partial\xi^j}. \end{aligned}$$

□

From the above theorems, we can see that

i) g_{ij} is invariant under reparametrizations of the sample space $\mathcal{X}$;

ii) g_{ij} is covariant under reparametrizations of the parameters space.

For Fisher metric g_{ij}, the Christoffel symbols of first kind are given by

$$\Gamma_{ij,k} = \frac{1}{2}\left(\frac{\partial}{\partial\theta^i}g_{jk} + \frac{\partial}{\partial\theta^k}g_{ik} - \frac{\partial}{\partial\theta^k}g_{ij}\right). \tag{C.105}$$

Proposition C.51. *The following equivalent expressions of the Christoffel symbols hold:*

i)

$$\begin{aligned}2\Gamma_{ij,k}(\boldsymbol{\theta}) =& \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right)\frac{\partial}{\partial\theta^k}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right]\\ &- \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta^k}\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right)\frac{\partial}{\partial\theta^i}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right]\\ &- \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta^k}\frac{\partial}{\partial\theta^i}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right)\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right]\\ &- \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x};\boldsymbol{\theta})\frac{\partial}{\partial\theta^k}\right)\right].\end{aligned}$$

ii)

$$\Gamma_{ij,k}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x};\boldsymbol{\theta}) + \frac{1}{2}\frac{\partial}{\partial\theta^i}\ln p(\boldsymbol{x};\boldsymbol{\theta}\frac{\partial}{\partial\theta^j}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right)\frac{\partial}{\partial\theta^k}\ln p(\boldsymbol{x};\boldsymbol{\theta})\right].$$

iii)

$$\Gamma_{ij,k} = 4\int_{\mathcal{X}}\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}\sqrt{p(\boldsymbol{x};\boldsymbol{\theta})}\cdot\frac{\partial}{\partial\theta^k}\sqrt{p(\boldsymbol{x};\boldsymbol{\theta})}d\boldsymbol{x}.$$

Fisher metric induces the Riemannian metric on $\mathcal{S}$.

$$g(X_{\boldsymbol{\theta}}, Y_{\boldsymbol{\theta}}) \coloneqq \sum_{i,j=1}^{n} g_{ij}(\boldsymbol{\theta})a^i(\boldsymbol{\theta})b^j(\boldsymbol{\theta}), \quad p(\boldsymbol{x};\boldsymbol{\theta}) \in \mathcal{S}, \tag{C.106}$$

where $X_{\boldsymbol{\theta}} = \sum_{i=1}^{n} a^i(\boldsymbol{\theta})\frac{\partial}{\partial\theta^i}p(\boldsymbol{x};\boldsymbol{\theta})$ and $Y_{\boldsymbol{\theta}} = \sum_{j=1}^{n} b^j(\boldsymbol{\theta})\frac{\partial}{\partial\theta^j}p(\boldsymbol{x};\boldsymbol{\theta})$ are vector fields on $\mathcal{S}$. The tensor g is called the Fisher-Riemannian metric. Its associated Levi-Civita connection is denoted by $\nabla^{(0)}$ and is defined by

$$g\left(\nabla^{(0)}_{\frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta^i}}\frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta^j}, \frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta^k}\right) = \Gamma_{ij,k}. \tag{C.107}$$

C.3.4 Dually Flat Connection

The Fisher metric g induces the $\nabla^{(1)}$-connection as

$$g(\nabla^{(1)}_{\partial_i}\partial_j, \partial_k) = \mathbb{E}_{\boldsymbol{\theta}}[(\partial_i\partial_j \ln p(\boldsymbol{x};\boldsymbol{\theta}))\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta})], \tag{C.108}$$

where $\partial_i = \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta^i}$. Or equivalently,

$$\Gamma^{(1)}_{ij,k}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[(\partial_i\partial_j \ln p(\boldsymbol{x};\boldsymbol{\theta}))\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta})]. \tag{C.109}$$

Proposition C.52. *The statistical model given by the exponential family is $\nabla^{(1)}$-flat.*

Proof.

$$\begin{aligned}
\Gamma^{(1)}_{ij,k}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}[(\partial_i\partial_j \ln p(\boldsymbol{x};\boldsymbol{\theta}))\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta})] \\
&= -\mathbb{E}_{\boldsymbol{\theta}}[\partial_i\partial_j\psi(\boldsymbol{\theta})\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta})] \\
&= -\partial_i\partial_j\psi(\boldsymbol{\theta})\mathbb{E}_{\boldsymbol{\theta}}[\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta})] \\
&= 0.
\end{aligned}$$

□

Then, we can see that the torsion and curvature of an exponential family with respect to connection $\nabla^{(1)}$ vanish everywhere.

Proposition C.53. *The mixture family is $\nabla^{(1)}$-flat if and only if is $\nabla^{(0)}$-flat.*

The Fisher metric g also induces the $\nabla^{(-1)}$-connection on a statistical model $\mathcal{S}$ as

$$g(\nabla^{(-1)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(-1)}_{ij,k} = \mathbb{E}_{\boldsymbol{\theta}}[(\partial_i\partial_j \ln p(\boldsymbol{x};\boldsymbol{\theta}) + \partial_i \ln p(\boldsymbol{x};\boldsymbol{\theta})\partial_j \ln p(\boldsymbol{x};\boldsymbol{\theta}))\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta})]. \tag{C.110}$$

Proposition C.54. *The mixture family is $\nabla^{(-1)}$-flat.*

Proposition C.55. *The exponential family is $\nabla^{(-1)}$-flat if and only if is $\nabla^{(0)}$-flat.*

Proposition C.56. *The following relation holds.*

$$\nabla^{(0)} = \frac{1}{2}\left(\nabla^{(-1)} + \nabla^{(1)}\right). \tag{C.111}$$

Proposition C.57. *The following relations hold*

$$\nabla^{(\alpha)} = (1-\alpha)\nabla^{(0)} + \alpha\nabla^{(1)} \tag{C.112}$$

$$= (1+\alpha)\nabla^{(0)} - \alpha\nabla^{(-1)} \tag{C.113}$$

$$= \nabla^{(0)} + \frac{\alpha}{2}(\nabla^{(1)} - \nabla^{(-1)}), \tag{C.114}$$

$$\nabla^{(0)} = \frac{1}{2}(\nabla^{(-\alpha)} + \nabla^{(\alpha)}). \tag{C.115}$$

Corollary C.58. *For the statistical model $\mathcal{S}$, the following statements are equivalent:*

i) $\mathcal{S}$ *is* $\nabla^{(\alpha)}$*-flat for all* $\alpha \in \mathbb{R}$*;*

ii) $\mathcal{S}$ *is both* $\nabla^{(1)}$ *and* $\nabla^{(-1)}$*-flat;*

iii) $\mathcal{S}$ *is both* $\nabla^{(0)}$ *and* $\nabla^{(1)}$*-flat;*

iv) $\mathcal{S}$ *is both* $\nabla^{(0)}$ *and* $\nabla^{(-1)}$*-flat.*

C.3.5 Skewness Tensor on Statistical Manifold

For two connections $\nabla^{(\alpha)}$ and ∇^{β}, the generalized difference tensor is defined as follows:

$$K^{(\alpha,\beta)}(X,Y) = \nabla_X^{(\beta)}Y - \nabla_X^{(\alpha)}Y. \tag{C.116}$$

Proposition C.59. *There is a 3-covariant, totally symmetric tensor T such that for any distinct α, β we have*

$$g(K^{(\alpha,\beta)}(X,Y),Z) = \frac{\alpha-\beta}{2}T(X,Y,Z), \tag{C.117}$$

where g is the Fisher metric.

The 3-covariant, symmetric tensor T with components

$$T(\partial_i,\partial_j,\partial_k) = T_{ijk} = \mathbb{E}_{\boldsymbol{\theta}}[\partial_i \ln p(\boldsymbol{x};\boldsymbol{\theta})\partial_j \ln p(\boldsymbol{x};\boldsymbol{\theta})\partial_k \ln p(\boldsymbol{x};\boldsymbol{\theta}) \tag{C.118}$$

is called the skewness tensor.

Theorem C.60. *The metric g and the skewness tensor T are related by*

$$\alpha T = \nabla^{(\alpha)} g. \tag{C.119}$$

Corollary C.61. *For any $\epsilon > 0$, the skewness tensor T is given by*

$$T = \frac{\nabla^{(\alpha+\epsilon)} g - \nabla^{(\alpha)} g}{\epsilon}. \tag{C.120}$$

C.3.6 Autoparallel Curves

Consider a smooth curve $\boldsymbol{\theta}(t)$ in the parameter space $\Theta \subset \mathbb{R}^n$, so that each component $\boldsymbol{\theta} = (\theta^1(t), \ldots, \theta^n(t))$ is smooth. A curve on the statistical manifold $\mathcal{S} = \{p(\boldsymbol{x}; \boldsymbol{\theta})\}$ is defined as

$$\gamma(t) = \iota(\boldsymbol{\theta}(t)) = p(\boldsymbol{x}; \boldsymbol{\theta}(t)), \quad t \in [0, T]. \tag{C.121}$$

The velocity of the curve $\gamma(t)$ is given by

$$\dot{\gamma}(t) = \frac{d}{dt} p(\boldsymbol{x}; \boldsymbol{\theta}(t)). \tag{C.122}$$

A curve $\gamma : [0, T] \to \mathcal{S}$ is called $\nabla^{(\alpha)}$-autoparallel if $\dot{\gamma}(t)$ is parallel transported along $\gamma(t)$:

$$\nabla^{(\alpha)}_{\dot{\gamma}(t)} \dot{\gamma}(t) = 0, \quad \forall t \in [0, T]. \tag{C.123}$$

Bibliography

[1] A. Acharya, S. Sanghavi, L. Jing, B. Bhushanam, D. Choudhary, M. Rabbat, and I. Dhillon. Positive unlabeled contrastive learning. *arXiv preprint arXiv:2206.01206*, 2022.

[2] A. Agarwal and T. Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory*, pages 2704–2729. PMLR, 2022.

[3] A. Agarwal, L. Bottou, M. Dudik, and J. Langford. Para-active learning. *arXiv preprint arXiv:1310.8243*, 2013.

[4] S. Agarwal, H. Arora, S. Anand, and C. Arora. Contextual diversity for active learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 137–153. Springer, 2020.

[5] N. Ahmed, J. Neville, and R. R. Kompella. Network sampling via edge-based node selection with graph induction. 2011.

[6] M. Al-Jarrah, E. Alsusa, and C. Masouros. A Unified Performance Framework for Integrated Sensing-Communications Based on KL-Divergence. *IEEE Transactions on Wireless Communications*, 22(12):9390–9411, 2023.

[7] P. Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

[8] S. Amari. Information geometry. *Contemporary Mathematics*, 203:81–96, 1997.

[9] S. Amari and H. Nagaoka. Differential geometry of smooth families of probability distributions. *METR*, 82:49–98, 1982.

[10] S.-i. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[11] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[12] S.-I. Amari. α-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

[13] S.-i. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[14] S.-i. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

[15] M. Anthony and J. Shawe-Taylor. A result of vapnik with applications. *Discrete Applied Mathematics*, 47(3):207–217, 1993.

[16] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38, 2017.

[17] K. M. Audenaert. Quantum skew divergence. *Journal of Mathematical Physics*, 55(11):112202, 2014.

[18] A. S. Awad and G. K. Atia. Distributionally robust domain adaptation. *arXiv preprint arXiv:2210.16894*, 2022.

[19] N. Ay, J. Jost, H. Vân Lê, and L. Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.

[20] K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019.

[21] F. Bach. Active learning for misspecified generalized linear models. *Advances in neural information processing systems*, 19, 2006.

[22] M. Basseville. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

[23] M. Bayarri and G. García-Donato. Generalization of jeffreys divergence-based priors for bayesian hypothesis testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):981–1003, 2008.

[24] J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109:719–760, 2020.

[25] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[26] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[27] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[28] R. A. Berk. An introduction to sample selection bias in sociological data. *American sociological review*, pages 386–398, 1983.

[29] D. Bertsekas and J. N. Tsitsiklis. *Introduction to probability*, volume 1. Athena Scientific, 2008.

[30] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

[31] D. Bertsimas, M. Sim, and M. Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618, 2019.

[32] K. Beven and A. Binley. The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6(3):279–298, 1992.

[33] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

[34] A. Beygelzimer, D. Hsu, N. Karampatziakis, J. Langford, and T. Zhang. Efficient active learning. In *ICML 2011 Workshop on On-line Trading of Exploration and Exploitation*, 2011.

[35] B. Bigi. Using kullback-leibler distance for text categorization. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings*, pages 305–319. Springer, 2003.

[36] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[37] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[38] J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[39] K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 59–66, 2003.

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

[41] W. Cai, Y. Zhang, and J. Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th international conference on data mining*, pages 51–60. IEEE, 2013.

[42] O. Calin and C. Udrişte. *Geometric modeling in probability and statistics*, volume 121. Springer, 2014.

[43] L. Cao, Z. Zhao, and D. Wang. Clustering algorithms. In *Target Recognition and Tracking for Millimeter Wave Radar in Intelligent Transportation*, pages 97–122. Springer, 2023.

[44] Z. Cao, M. Long, J. Wang, and M. I. Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 135–150, 2018.

[46] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[47] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2985–2994, 2019.

[48] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[49] B. M. Carvalho, E. Garduno, and I. O. Santos. Skew divergence-based fuzzy segmentation of rock samples. In *Journal of Physics: Conference Series*, volume 490, page 012010, 2014.

[50] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

[51] Y. S. Chan and H. T. Ng. Word sense disambiguation with distribution estimation. In *IJCAI*, volume 5, pages 1010–5, 2005.

[52] O. Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105, 2014.

[53] N. Charoenphakdee, J. Vongkulbhisal, N. Chairatanakul, and M. Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5202–5211, 2021.

[54] K. C. Chatzisavvas, C. C. Moustakidis, and C. Panos. Information entropy, information distances, and complexity in atoms. *The Journal of chemical physics*, 123 (17):174111, 2005.

[55] H. Chen and X. R. Li. Distributed active learning with application to battery health management. In *14th International conference on information fusion*, pages 1–7. IEEE, 2011.

[56] X. Chen, M. Monfort, A. Liu, and B. D. Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

[57] X. Chen, C. Gong, and J. Yang. Cost-sensitive positive and unlabeled learning. *Information Sciences*, 558:229–245, 2021.

[58] K. Choi, M. Liao, and S. Ermon. Featurized density ratio estimation. In *Uncertainty in Artificial Intelligence*, pages 172–182. PMLR, 2021.

[59] P. Cimiano and J. Völker. Towards large-scale, open-domain and ontology-based named entity classification. In *Proceedings of RANLP*, volume 5, pages 166–172, 2005.

[60] A. Clim, R. D. Zota, and G. TinicĂ. The kullback-leibler divergence used in machine learning algorithms for health care applications and hypertension prediction: a literature review. *Procedia Computer Science*, 141:448–453, 2018.

[61] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

[62] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[63] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[64] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[65] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[66] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[67] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005.

[68] A. de Mathelin, G. Richard, F. Deheeger, M. Mougeot, and N. Vayatis. Adversarial weighting for domain adaptation in regression. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 49–56. IEEE, 2021.

[69] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3): 595–612, 2010.

[70] L. Deng and Y. Liu. *Deep learning in natural language processing*. Springer, 2018.

[71] E. Deza, M. M. Deza, M. M. Deza, and E. Deza. *Encyclopedia of distances*. Springer, 2009.

[72] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, and D. Tao. Exploring representativeness and informativeness for active learning. *IEEE transactions on cybernetics*, 47(1):14–26, 2015.

[73] M. Du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR, 2015.

[74] M. C. Du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27, 2014.

[75] J. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[76] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[77] K. Eckhardt, N. Fohrer, and H.-G. Frede. Automatic model calibration. *Hydrological Processes: An International Journal*, 19(3):651–658, 2005.

[78] S. Eguchi and O. Komori. Path connectedness on a space of probability density functions. In *Lecture Notes in Computer Science*, pages 615–624. Springer International Publishing, 2015. doi: 10.1007/978-3-319-25040-3_66. URL https://doi.org/10.1007/978-3-319-25040-3_66.

[79] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[80] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.

[81] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

[82] T. Fang, N. Lu, G. Niu, and M. Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems*, 33:11996–12007, 2020.

[83] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[84] S. Farquhar, Y. Gal, and T. Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

[85] P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[86] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 562–577. Springer, 2014.

[87] B. Fu, Z. Cao, M. Long, and J. Wang. Learning to detect open classes for universal domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 567–583. Springer, 2020.

[88] S. Garg, S. Balakrishnan, and Z. Lipton. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, 35:22531–22546, 2022.

[89] A. Ghosh, T. Schaaf, and M. Gormley. Adafocal: Calibration-aware adaptive focal loss. *Advances in Neural Information Processing Systems*, 35:1583–1595, 2022.

[90] D. Gogolashvili, M. Zecchin, M. Kanagawa, M. Kountouris, and M. Filippone. When is importance weighting correction needed for covariate shift adaptation? *arXiv preprint arXiv:2303.04020*, 2023.

[91] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

[92] J. Goldberger, S. Gordon, H. Greenspan, et al. An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures. In *ICCV*, volume 3, pages 487–493, 2003.

[93] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[94] C. H. Gooi and J. Allan. Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 9–16, 2004.

[95] J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[96] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. Pmlr, 2017.

[97] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5, 2009.

[98] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[99] W. Gu, T. Zhang, and H. Jin. Entropy weight allocation: Positive-unlabeled learning via optimal transport. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 37–45. SIAM, 2022.

[100] B. Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

[101] H. W. Guggenheimer. *Differential geometry*. Courier Corporation, 2012.

[102] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.

[103] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[104] H. V. Gupta, K. J. Beven, and T. Wagener. Model calibration and uncertainty estimation. *Encyclopedia of hydrological sciences*, 2006.

[105] K. Gurney. *An introduction to neural networks*. CRC press, 2018.

[106] G. Hacohen and D. Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019.

[107] Z. Han, Z. Liang, F. Yang, L. Liu, L. Li, Y. Bian, P. Zhao, B. Wu, C. Zhang, and J. Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35:37704–37718, 2022.

[108] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities. By GH Hardy, JE Littlewood, G. Pólya.*. University Press, 1952.

[109] A. Hassan, R. Damper, and M. Niranjan. On acoustic emotion recognition: compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013.

[110] A. Hassan, F. Riaz, and A. Shaukat. Scale and rotation invariant texture classification using covariate shift methodology. *IEEE Signal Processing Letters*, 21(3): 321–324, 2014.

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

[112] S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.

[113] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[114] M. C. Hill. Methods and guidelines for effective model calibration. In *Building partnerships*, pages 1–10. 2000.

[115] H. Hino. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*, 2020.

[116] H. Hino, S. Akaho, and N. Murata. Geometry of em and related iterative algorithms. *Information Geometry*, pages 1–39, 2022.

[117] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[118] C. Holtz, T.-W. Weng, and G. Mishne. Learning sample reweighting for accuracy and adversarial robustness. *arXiv preprint arXiv:2210.11513*, 2022.

[119] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[120] H. S. Hossain, N. Roy, and M. A. A. H. Khan. Sleep well: a sound sleep monitoring framework for community scaling. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 1, pages 44–53. IEEE, 2015.

[121] W.-N. Hsu and H.-T. Lin. Active learning by learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[122] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

[123] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[124] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010.

[125] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 581–589, 2007.

[126] A. R. Islam. Machine learning in computer vision. In *Applications of Machine Learning and Artificial Intelligence in Education*, pages 48–72. IGI Global, 2022.

[127] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[128] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[129] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[130] Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[131] Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[132] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. ACL, 2007.

[133] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13:455–492, 1998.

[134] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[135] P. Joulani, A. Gyorgy, and C. Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

[136] K.-S. Jun and R. Nowak. Graph-based active learning: A new look at expected error minimization. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1325–1329. IEEE, 2016.

[137] C. Käding, A. Freytag, E. Rodner, A. Perino, and J. Denzler. Large-scale active learning with approximations of expected model output changes. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 179–191. Springer, 2016.

[138] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[139] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

[140] J. D. Kelleher, B. Mac Namee, and A. D'arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.

[141] S. Khalighi, T. Sousa, and U. Nunes. Adaptive automatic sleep stage classification under covariate shift. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2259–2262. IEEE, 2012.

[142] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[143] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *arXiv preprint arXiv:2105.08919*, 2021.

[144] M. Kimura. Generalized t-sne through the lens of information geometry. *IEEE Access*, 9:129619–129625, 2021.

[145] M. Kimura and H. Hino. α-geodesical skew divergence. *Entropy*, 23(5):528, 2021.

[146] M. Kimura and H. Hino. Information geometrically generalized covariate shift adaptation. *Neural Computation*, 34(9):1944–1977, 2022.

[147] M. Kimura, K. Wakabayashi, and A. Morishima. Batch prioritization of data labeling tasks for training classifiers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 163–167, 2020.

[148] R. Kiryo, G. Niu, M. C. Du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30, 2017.

[149] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[150] T. Kohonen. Essentials of the self-organizing map. *Neural Networks*, 37:52–65, 2013.

[151] A. N. Kolmogorov and G. Castelnuovo. *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei, 1930.

[152] J. Kossen, S. Farquhar, Y. Gal, and T. Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR, 2021.

[153] D. Kottke, J. Schellinger, D. Huseljic, and B. Sick. Limitations of assessing active learning performance at runtime. *arXiv preprint arXiv:1901.10338*, 2019.

[154] W. M. Kouw and M. Loog. On regularization parameter estimation under covariate shift. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 426–431. IEEE, 2016.

[155] E. Kreyszig. *Differential geometry*. Courier Corporation, 2013.

[156] S. Krügel, A. Ostermaier, and M. Uhl. The moral authority of chatgpt. *arXiv preprint arXiv:2301.07098*, 2023.

[157] S. I. Ktena, A. Tejani, L. Theis, P. K. Myana, D. Dilipkumar, F. Huszár, S. Yoo, and W. Shi. Addressing delayed feedback for continuous training with neural networks in ctr prediction. In *Proceedings of the 13th ACM conference on recommender systems*, pages 187–195, 2019.

[158] W. Kühnel. *Differential geometry*, volume 77. American Mathematical Soc., 2015.

[159] S. Kulinski and D. I. Inouye. Towards explaining distribution shifts. In *International Conference on Machine Learning*, pages 17931–17952. PMLR, 2023.

[160] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[161] A. Kumagai, T. Iwata, and Y. Fujiwara. Meta-learning for relative density-ratio estimation. *Advances in Neural Information Processing Systems*, 34:30426–30438, 2021.

[162] J. N. Kundu, N. Venkat, A. Revanur, R. V. Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12376–12385, 2020.

[163] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*, 2023.

[164] S. Lang. *Differential and Riemannian manifolds*, volume 160. Springer Science & Business Media, 2012.

[165] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[166] J. M. Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

[167] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32, 1999.

[168] L. Lee. On the effectiveness of the skew divergence for statistical language analysis. In *International Workshop on Artificial Intelligence and Statistics*, pages 176–183. PMLR, 2001.

[169] W. Lee, J. Lee, and S. Park. Instance weighting domain adaptation using distance kernel. *Industrial Engineering & Management Systems*, 17(2):334–340, 2018.

[170] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[171] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[172] M. Li and I. K. Sethi. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1251–1261, 2006.

[173] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[174] J. Liang, Y. Wang, D. Hu, R. He, and J. Feng. A balanced and uncertainty-aware approach for partial domain adaptation. In *European conference on computer vision*, pages 123–140. Springer, 2020.

[175] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[176] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[177] Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[178] A. Liu and B. Ziebart. Robust classification under sample selection bias. *Advances in neural information processing systems*, 27, 2014.

[179] A. Liu and B. D. Ziebart. Robust covariate shift prediction with general losses and feature views. *arXiv preprint arXiv:1712.10043*, 2017.

[180] A. Liu, X. Hu, L. Wen, and P. S. Yu. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*, 2023.

[181] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE international conference on data mining*, pages 179–186. IEEE, 2003.

[182] C. Liu, X. Dong, M. Potter, H.-M. Chang, and R. Soni. Adversarial focal loss: Asking your discriminator for hard examples. *arXiv preprint arXiv:2207.07739*, 2022.

[183] H. Liu, Z. Cao, M. Long, J. Wang, and Q. Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2927–2936, 2019.

[184] S. Liu, A. Takeda, T. Suzuki, and K. Fukumizu. Trimmed density ratio estimation. *Advances in neural information processing systems*, 30, 2017.

[185] W. Liu and S. Chawla. Class confidence weighted k nn algorithms for imbalanced data sets. In *Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part II 15*, pages 345–356. Springer, 2011.

[186] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[187] D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[188] Y. Lu, Y. Bo, and W. He. Noise attention learning: Enhancing noise robustness by gradient scaling. *Advances in Neural Information Processing Systems*, 35:23164–23177, 2022.

[189] Y. Lu, W. Ji, Z. Izzo, and L. Ying. Importance tempering: Group robustness for overparameterized models. *arXiv preprint arXiv:2209.08745*, 2022.

[190] B. D. Lund and T. Wang. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, 40(3):26–29, 2023.

[191] A. Luntz. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 1969.

[192] J. I. S. Martin, S. Mazuelas, and A. Liu. Double-weighting for covariate shift adaptation. In *International Conference on Machine Learning*, pages 30439–30457. PMLR, 2023.

[193] N. Mashayekhi. *An Adversarial Approach to Importance Weighting for Domain Adaptation*. PhD thesis, 2022.

[194] H. Matsuzoe, J.-i. Takeuchi, and S.-i. Amari. Equiaffine structures on statistical manifolds and bayesian statistics. *Differential Geometry and its Applications*, 24 (6):567–578, 2006.

[195] D. A. McAllester. Some PAC-Bayesian Theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

[196] R. W. McGee. Capitalism, socialism and chatgpt. *Available at SSRN 4369953*, 2023.

[197] R. W. McGee. Is chat gpt biased against conservatives? an empirical study. *An Empirical Study (February 15, 2023)*, 2023.

[198] P. Melville and V. Sindhwani. Recommender systems. *Encyclopedia of machine learning*, 1:829–838, 2010.

[199] M. Menéndez, J. Pardo, L. Pardo, and M. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[200] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[201] T. M. Mitchell. Machine learning, 1997.

[202] J. Mockus. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2:117, 1998.

[203] S. Mohamad, M. Sayed-Mouchaweh, and A. Bouchachia. Active learning for classifying data streams with unknown number of classes. *Neural Networks*, 98:1–15, 2018.

[204] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[205] P. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. *Advances in neural information processing systems*, 16, 2003.

[206] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1): 521–530, 2012.

[207] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

[208] M. Mulyanto, M. Faisal, S. W. Prakosa, and J.-S. Leu. Effectiveness of focal loss for minority classification in network intrusion detection systems. *Symmetry*, 13 (1):4, 2020.

[209] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[210] S. Mussmann, J. Reisler, D. Tsai, E. Mousavi, S. O'Brien, and M. Goldszmidt. Active learning with expected error reduction. *arXiv preprint arXiv:2211.09283*, 2022.

[211] M. S. Nacson, S. Gunasekar, J. Lee, N. Srebro, and D. Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*, pages 4683–4692. PMLR, 2019.

[212] H. Naganuma and M. Kimura. Necessary and sufficient hypothesis of curvature: Understanding connection between out-of-distribution generalization and calibration. *ICLR workshop on Domain Generalization*, 2023.

[213] T. D. Nguyen, M. Christoffel, and M. Sugiyama. Continuous target shift adaptation in supervised learning. In *Asian Conference on Machine Learning*, pages 285–300. PMLR, 2016.

[214] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[215] B. Nicholson, J. Zhang, V. S. Sheng, and Z. Wang. Label noise correction methods. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE, 2015.

[216] B. Nicholson, V. S. Sheng, and J. Zhang. Label noise correction and application in crowdsourcing. *Expert Systems with Applications*, 66:149–162, 2016.

[217] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[218] F. Nielsen. A family of statistical symmetric divergences based on jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.

[219] F. Nielsen. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters*, 20(7):657–660, 2013.

[220] F. Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

[221] F. Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10): 1100, 2020.

[222] F. Nielsen. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221, 2020.

[223] F. Nielsen. On a variational definition for the jensen-shannon symmetrization of distances based on the information radius. *Entropy*, 23(4):464, 2021.

[224] R. Nishii and S. Eguchi. Image classification based on markov random field models with jeffreys divergence. *Journal of multivariate analysis*, 97(9):1997–2008, 2006.

[225] D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[226] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[227] P. Panareda Busto and J. Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017.

[228] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[229] W. Park, D. Kim, Y. Lu, and M. Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[230] K. R. Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 2005.

[231] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and

S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[232] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[233] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[234] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[235] X. Peng, K. Wang, Z. Zeng, Q. Li, J. Yang, and Y. Qiao. Suppressing mislabeled data via grouping and self-attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 786–802. Springer, 2020.

[236] A. L. Perry, P. J. Low, J. R. Ellis, and J. D. Reynolds. Climate change and distribution shifts in marine fishes. *science*, 308(5730):1912–1915, 2005.

[237] C. Pike-Burke, S. Agrawal, C. Szepesvari, and S. Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

[238] B. Plank, A. Johannsen, and A. Søgaard. Importance weighting and unsupervised domain adaptation of pos taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, 2014.

[239] I. Portugal, P. Alencar, and D. Cowan. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 97:205–227, 2018.

[240] Q. Qi, Y. Xu, W. Yin, R. Jin, and T. Yang. Attentional-biased stochastic gradient descent. *Transactions on Machine Learning Research*, 2022.

[241] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning.* Mit Press, 2008.

[242] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[243] C. Rao and T. Nayak. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, 31(5):589–593, 1985.

[244] P. P. Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

[245] M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

[246] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

[247] P. Revathi and M. Hemalatha. Cotton leaf spot diseases detection utilizing feature selection with skew divergence method. *International Journal of scientific engineering and technology*, 3(1):22–30, 2014.

[248] B. Rhodes, K. Xu, and M. U. Gutmann. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

[249] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[250] A. Safari, R. M. Altman, and T. M. Loughin. Display advertising: Estimating conversion probability efficiently. *arXiv preprint arXiv:1710.08583*, 2017.

[251] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[252] T. Sakai and N. Shimizu. Covariate shift adaptation on learning from positive and unlabeled data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4838–4845, 2019.

[253] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81(10.5555):26853, 1986.

[254] S. Santurkar, L. Schmidt, and A. Madry. A classification-based study of covariate shift in gan distributions. In *International Conference on Machine Learning*, pages 4480–4489. PMLR, 2018.

[255] S. Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

[256] A. Satti, C. Guan, D. Coyle, and G. Prasad. A covariate shift minimisation method to alleviate non-stationarity effects for an adaptive brain-computer interface. In *2010 20th International Conference on Pattern Recognition*, pages 105–108. IEEE, 2010.

[257] B. Settles. Active learning literature survey. 2009.

[258] B. Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

[259] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[260] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244, 2000.

[261] U. Simon. Affine differential geometry. In *Handbook of differential geometry*, volume 1, pages 905–961. Elsevier, 2000.

[262] A. Smith, P. A. Naik, and C.-L. Tsai. Markov-switching model selection using kullback–leibler divergence. *Journal of Econometrics*, 134(2):553–577, 2006.

[263] L. N. Smith. Cyclical focal loss. *arXiv preprint arXiv:2202.08978*, 2022.

[264] A. J. Smola and B. Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

[265] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[266] K. Solanki, K. Sullivan, U. Madhow, B. Manjunath, and S. Chandrasekaran. Provably secure steganography: Achieving zero KL divergence using statistical restoration. In *2006 International Conference on Image Processing*, pages 125–128. IEEE, 2006.

[267] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[268] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[269] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[270] M. Spüler, W. Rosenstiel, and M. Bogdan. Principal component based covariate shift adaption to reduce non-stationarity in a meg-based brain-computer interface. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–7, 2012.

[271] J. J. Stoker. *Differential geometry*, volume 20. John Wiley & Sons, 2011.

[272] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[273] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.

[274] A. Subbaswamy, R. Adams, and S. Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021.

[275] M. Sugiyama. Active learning for misspecified models. *Advances in neural information processing systems*, 18, 2005.

[276] M. Sugiyama. *Introduction to statistical machine learning*. Morgan Kaufmann, 2015.

[277] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (5), 2007.

[278] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

[279] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[280] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. A two-stage weighting framework for multi-source domain adaptation. *Advances in neural information processing systems*, 24, 2011.

[281] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[282] J. Takeuchi and S.-i. Amari. α-parallel prior and its properties. *IEEE transactions on information theory*, 51(3):1011–1023, 2005.

[283] M. Tallis and P. Yadav. Reacting to variations in product demand: An application for conversion rate (cr) prediction in sponsored search. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1856–1864. IEEE, 2018.

[284] L. Tao, M. Dong, and C. Xu. Dual focal loss for calibration. *arXiv preprint arXiv:2305.13665*, 2023.

[285] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[286] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[287] K. Tomanek. Resource-aware annotation through active learning. 2010.

[288] K. Tomanek and K. Morik. Inspecting sample reusability for active learning. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 169–181. JMLR Workshop and Conference Proceedings, 2011.

[289] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[290] G. S. Tran, T. P. Nghiem, V. T. Nguyen, C. M. Luong, J.-C. Burie, et al. Improving accuracy of lung nodule classification using deep learning with focal loss. *Journal of healthcare engineering*, 2019, 2019.

[291] V.-T. Tran. *Selection bias correction in supervised learning with importance weight*. PhD thesis, Université de Lyon, 2017.

[292] V.-T. Tran and A. Aussem. Correcting a class of complete selection bias with external data based on importance weight estimation. In *International Conference on Neural Information Processing*, pages 111–118. Springer, 2015.

[293] 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*, pages 7167–7176, 2017.

[294] K. Ueki, M. Sugiyama, and Y. Ihara. Perceived age estimation under lighting condition change by covariate shift adaptation. In *2010 20th International Conference on Pattern Recognition*, pages 3400–3403. IEEE, 2010.

[295] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

[296] J. E. Van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

[297] G. Van Tulder. Sample reusability in importance-weighted active learning. 2012.

[298] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[299] V. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[300] F. Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169, 1998.

[301] C. Vernade, O. Cappé, and V. Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

[302] M. M.-C. Vidovic, L. P. Paredes, H.-J. Hwang, S. Amsu, J. Pahl, J. M. Hahne, B. Graimann, D. Farina, K.-R. Müller, et al. Covariate shift adaptation in emg pattern recognition for prosthetic device control. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pages 4370–4373. IEEE, 2014.

[303] F. Wang, B. C. Vemuri, and A. Rangarajan. Groupwise point pattern registration using a novel cdf-based jensen-shannon divergence. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 1283–1288. IEEE, 2006.

[304] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[305] X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

[306] D. Wei, K. N. Ramamurthy, and K. R. Varshney. Health insurance market risk assessment: Covariate shift and k-anonymity. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 226–234. SIAM, 2015.

[307] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[308] J. Wen, R. Greiner, and D. Schuurmans. Domain aggregation networks for multi-source domain adaptation. In *International conference on machine learning*, pages 10214–10224. PMLR, 2020.

[309] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

[310] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[311] C. Winship and R. D. Mare. Models for sample selection bias. *Annual review of sociology*, 18(1):327–350, 1992.

[312] T.-H. Wu, Y.-C. Liu, Y.-K. Huang, H.-Y. Lee, H.-T. Su, P.-C. Huang, and W. H. Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15510–15519, 2021.

[313] W. Wu, M. Guo, Y. Liu, and R. Xu. Active learning with optimal distribution for image classification. In *2011 International Conference on Multimedia Technology*, pages 132–136. IEEE, 2011.

[314] R. Xia, Z. Pan, and F. Xu. Instance weighting for domain adaptation via trading off sample selection bias and variance. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden*, pages 13–19, 2018.

[315] F. Xiao, Y. Wu, H. Zhao, R. Wang, and S. Jiang. Dual skew divergence loss for neural machine translation. *arXiv preprint arXiv:1908.08399*, 2019.

[316] N. Xiao and L. Zhang. Dynamic weighted learning for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15242–15251, 2021.

[317] D. Xu, Y. Ye, and C. Ruan. Understanding the role of importance weighting for deep learning. *arXiv preprint arXiv:2103.15209*, 2021.

[318] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17:151–178, 2020.

[319] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[320] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative Density-Ratio estimation for robust distribution comparison. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 594–602. Curran Associates, Inc., 2011.

[321] M. Yamada, L. Sigal, and M. Raptis. No bias left behind: Covariate shift adaptation for discriminative 3d pose estimation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 674–687. Springer, 2012.

[322] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

[323] X. Yang, X. Yang, J. Yang, Q. Ming, W. Wang, Q. Tian, and J. Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34:18381–18394, 2021.

[324] Y. Yang and M. Loog. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2646–2651. IEEE, 2016.

[325] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113:113–127, 2015.

[326] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

[327] S. Yasui, G. Morishita, F. Komei, and M. Shibata. A feedback shift correction in predicting conversion rates under delayed feedback. In *Proceedings of The Web Conference 2020*, pages 2740–2746, 2020.

[328] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021.

[329] J. Yin. *On the Improvement of Density Ratio Estimation via Probabilistic Classifier–Theoretical Study and Its Applications*. PhD thesis, The University of British Columbia (Vancouver, 2023.

[330] Y. Yoshikawa and Y. Imai. A nonparametric delayed feedback model for conversion rate prediction. *arXiv preprint arXiv:1802.00255*, 2018.

[331] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[332] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13 (3):55–75, 2018.

[333] D. Yu, K. Yao, H. Su, G. Li, and F. Seide. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7893–7897. IEEE, 2013.

[334] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.

[335] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[336] J. Zhang, Z. Ding, W. Li, and P. Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[337] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

[338] S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

[339] T. Zhang, I. Yamane, N. Lu, and M. Sugiyama. A one-step approach to covariate shift adaptation. In *Asian Conference on Machine Learning*, pages 65–80. PMLR, 2020.

[340] E. Zhao, A. Liu, A. Anandkumar, and Y. Yue. Active learning under label shift. In *International Conference on Artificial Intelligence and Statistics*, pages 3412–3420. PMLR, 2021.

[341] L. Zhao, G. Sukthankar, and R. Sukthankar. Importance-weighted label prediction for active learning with noisy annotations. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3476–3479. IEEE, 2012.

[342] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[343] Z.-H. Zhou. *Machine learning*. Springer Nature, 2021.

[344] J. Zhu, H. Wang, B. K. Tsou, and M. Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6):1323–1331, 2009.

[345] X. Zhu and A. B. Goldberg. *Introduction to semi-supervised learning*. Springer Nature, 2022.

[346] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.