WHY DOES DEEP LEARNING WORKS ?

MAJOR TECHNICAL PROJECT (DP 401P)

to be submitted by

BHAVYA BHATT

for the

END-SEMESTER EVALUATION

under the supervision of

DR. SAMAR AGNIHOTRI

SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

KAMAND-175005, INDIA

JULY, 2020

Redefined Targets

School of Computing and Electrical Engineering

- **Name :** Bhavya Bhatt
- **Roll No. :** B16016
- **Branch:** Computer Science and Engineering
- **Project Title:** Why does deep learning works?
- **Supervisor:** Dr. Samar Agnihotri

0.1 Objectives

Original	Revised
Implementation of EDGE method	Explore theoretical MI lower bounds
Unification of theoretical approaches	Review information geometry literature

ABSTRACT

The report give an exhaustive and rigorous introduction to the use of information theory in deep learning. Starting with the most exceptional and original information theoretic framework based on information bottleneck principle first introduced by Prof. Naftali Tishby. Information bottleneck principle has provided a strong and bold predictions about learning dynamics of deep learning, an area which was believed to be inaccessible and black box before the introduction of IB. We provide a self-sufficient and complete introduction to all the major area of IB research along with the modification on original theory. After that we provide some starting motivating arguments inspired from IB theory and general covariance principle in statistics which hints us towards an information geometric framework of deep learning. We move on to give a rigorous self sufficient introduction to the field of information geometry. This report also exposes to a complete mathematical prerequisites including basics of differential geometry required to conduct research in the field. We have also provided an extensive and rather long introduction to the mathematical theory of connections which is indispensable in the field of theoretical deep learning. The expositions on the theory of connections treated in this report is by far the most rigours introduction in the literature. We give the closure by providing one application of information geometry in deep learning namely natural gradient descent which poses a fundamental change in traditional gradient descent used to train deep neural networks.

Keywords: *Mutual Information, Information Theory, Information Bottleneck Principle, Deep Learning, Information Geometry, Theory of Connections*

Contents

Introduction

1.1 Abstract

Since the success of deep learning, where it outperforms almost every machine learning task including computer vision, speech recognition, natural language processing, representation learning and many more and after so many years when it first shone like a charm, there lies many subtle mysterious regarding theoretical aspects of deep learning which revolves around the questions dealing with generalization, memorization, regularization, hidden representations etc. Many great researchers have put forward their views and perspectives on how to approach towards a theoretical framework for deep learning which has explanations and interpretations of what we observe in practice while training a deep learning model. One of the most promising ideas which is there for quite some time now is the application of information theory in deep learning using which researchers have shown interesting state-of-the-art results comparable to results from standard practices in deep learning. Many people believe that information theory and deep learning is just two different sides of the same coin. This report investigates extensively different types of theoretical formulations among which a heavy emphasis is on the information theoretic description. We will be discussing the complete

framework of information geometry and it's application in machine learning and deep learning.

1.2 Background and Literature Survey

A sketch of IB principle for deep learning was first introduced in the paper [3] and after that another paper [4] by the same author introduced a complete theory of two phase training process giving precise meaning to generalization and compression ability of a network. Currently there are three main areas in which IB-theory is been applied which are as follows:

- Using IB-theory to analyse training dynamics for standard loss functions and optimization algorithms.
- Use IB-theory to obtain performance bounds and experiment with them in practice.
- Use IB-based training paradigms for DNN, for example Deep Variational Information Bottleneck encoder, training using HSIC objective function, HSIC sigma networks etc.

The paper used tanh activation function for their experiment claimed to have observed two phase training dynamics namely generalization phase and then compression phase. According to the paper, in generalization phase the network rapidly increases it's generalization abilities (I(H, Y)) for all the layers at the cost of increasing complexity of representation (I(X, H)). After that it enters compression phase in which the network decreases the complexity measure without much changing generalization ability.

The important part of the above computation is estimation of mutual information which is challenging task in itself. There exist a lot of mutual information estimation methods and it was observed that the two phase observation which was claimed in [4] was sensitive to the methods used for estimating MI. This observation was given in [7] and critically attacked the original paper [4]. Further many more accurate methods were published as given in [10] and they confirmed the phenomenon of two phase training process. This motivated researchers to use IB-theory not just for analysing the training dynamics but rather borrow the objective function itself as loss function to train the network. These approaches are called variational information bottleneck models and one such example is given in [13] and another example which uses HSIC measure instead of MI is given in .

The central idea of IB-theory is the attempt to give precise definition of optimal hidden representations and using this define measure of generalization and compression abilities. These optimal hidden representations are expressed in terms of probability distribution which inherently assumes stochastic neural networks. Many of the researchers have shown the inadequacy of using MI as criterion to measure generalization and compression in case of non-stochastic neural networks (or deterministic networks). These deterministic networks suffer from critical theoretical issues when uses MI as criterion which is described in more detail in [12]. The same paper gives a rough sketch of what an optimal representation should truly capture in context of deterministic neural network. These sketches are based on the need for a representation to not only capture generalization and compression (as done in traditional IB-Theory) but also more exotic properties like simple decision rule and robustness to noise in the data. So most of the effort in this

field is now concentrated on coming up with new criterion (like MI in traditional IB-Theory) which are expected to apparently solve the issues as described in [12]. The research for this new 'ideal' criterion will advance both the field i.e. one involving training dynamics analysis and other in which IB criterion is used as objective function to train the neural networks.

1.3 Objective and scope of the Work

1.3.1 Mathematical Preliminary

The IB theory assumes that $p_{XY}(x, y)$ (from which finite samples of dataset is generated) is given. Given this, IB-theory introduces an optimal representation of variable X by the following lagrangian

$$L = I(X, \hat{X}) - \beta I(\hat{X}, Y)$$

where I is mutual information between the arguments. The above objective attains it's minimum value for the minimal sufficient statistics of X. The minimum sufficient statistics is defined as the compressed representation of X which is maximally independent from X given that it keeps most of the information about Y. The optimal probability distribution of compressed representation $\hat{X}$ is defined as follows [1]

$$p^*(\hat{X}|X) = \arg\min_{p(\hat{X}|X)} L(p(\hat{X}|X))$$

Here, β is trade-off parameter between generalization and compression terms in the IB objective. The figure below shows the true (black) and empirical

[1] for more details, see references

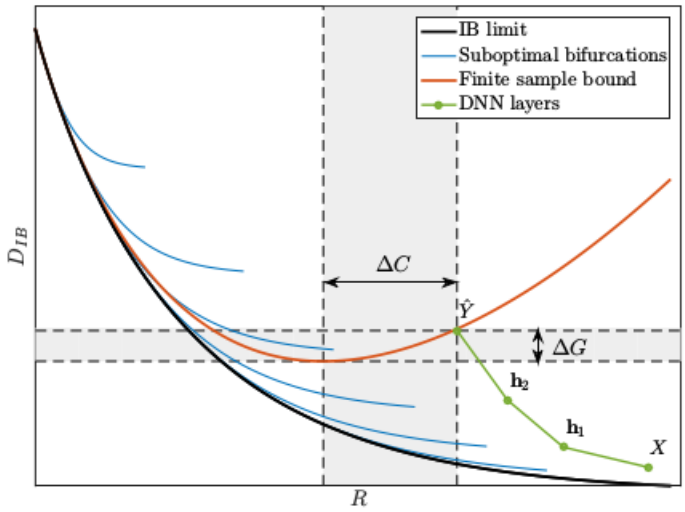

Figure 1.1: A qualitative information plane, with a hypothesized path of the layers in a typical DNN (green line) on the training data. The black line is the optimal achievable IB limit, and the blue lines are sub-optimal IB bifurcations, obtained by forcing the cardinality of X or remaining in the same representation. The red line corresponds to the upper bound on the out-of-sample IB distortion (mutual information on Y), when training from a finite sample. While the training distortion may be very low (the green points) the actual distortion can be as high as the red bound. This is the reason why one would like to shift the green DNN layers closer to the optimal curve to obtain lower complexity and better generalization. Another interesting consequence is that getting closer to the optimal limit requires stochastic mapping between the layers. Figure is taken from [3]

(red) optimal information plane curve for different trade-off parameters β (slope of the curve is β^{-1}).

1.4 Methodology

- Implement deep variational information bottleneck models and benchmark their performance with other state-of-the-art models.

- The second part of the project is to theoretically study the properties of 'ideal' criterion measure and come up with realization of it using techniques from information theory, signal processing (noise robustness), statistical mechanics (bifurcations in case second-order dependence in data as predicted by IB-Theory), high dimensional probability theory (finite sample bounds) and stochastic processes (study the dynamics of stochasticity in context of neural networks).

- The final aim of the project is to explore other approaches to the theoretical framework of deep learning. The expected outcome is provide a unified study of these approaches and how can they be possibly sew together in a grand framework which can be used as a guide to more advanced theoretical studies in the field of deep learning.

1.5 Results

Now some of the results obtained for a particular architecture is described as follows: Criterion of dependence used in the experiment is HSIC (Hilbert-Schmidt Independence Criterion). Plots in Figure 1.2 are obtained between HSIC(X, H) and HSIC(H, Y). Plots in Figure 1.3 are between epochs vs HSIC(X, H) and HSIC(H, Y). The architecture used is Conv2d(16, 3, 1, 1) [2]

[2]the semantics is (filter, kernel, stride, padding)

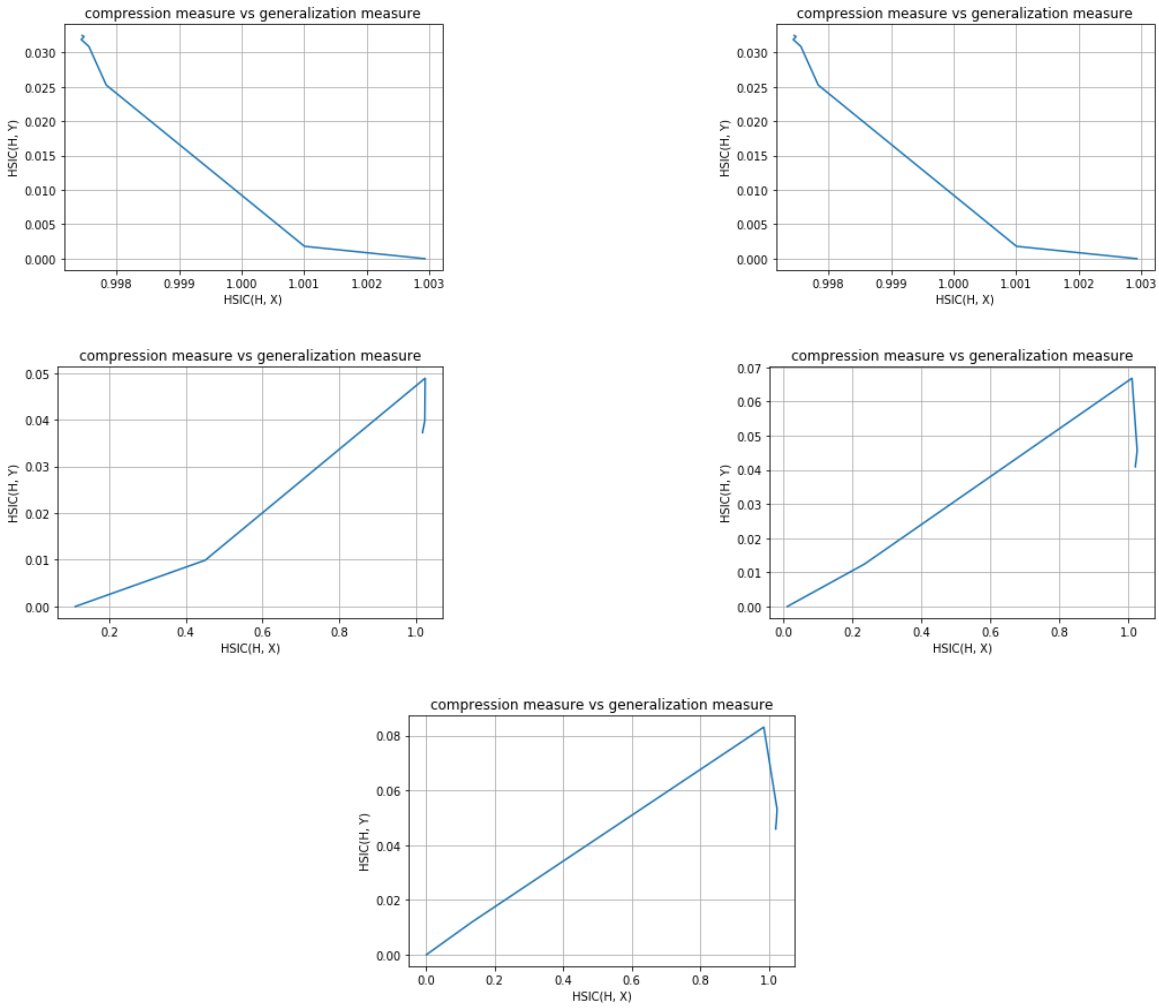

Figure 1.2: Plots between HSIC(X, H) and HSIC(H, Y) for all parametric layers - starting with first layer from top left

- (Flatten) - 500 - 200 - 50 - 10, in which relu activation function is used for hidden layers (including Conv layer) and softmax for output layer with no explicit regularization used. This is done in order to test how much current loss function (cross entropy loss is used in the experiment) is capable of capturing generalization and compression abilities while training.

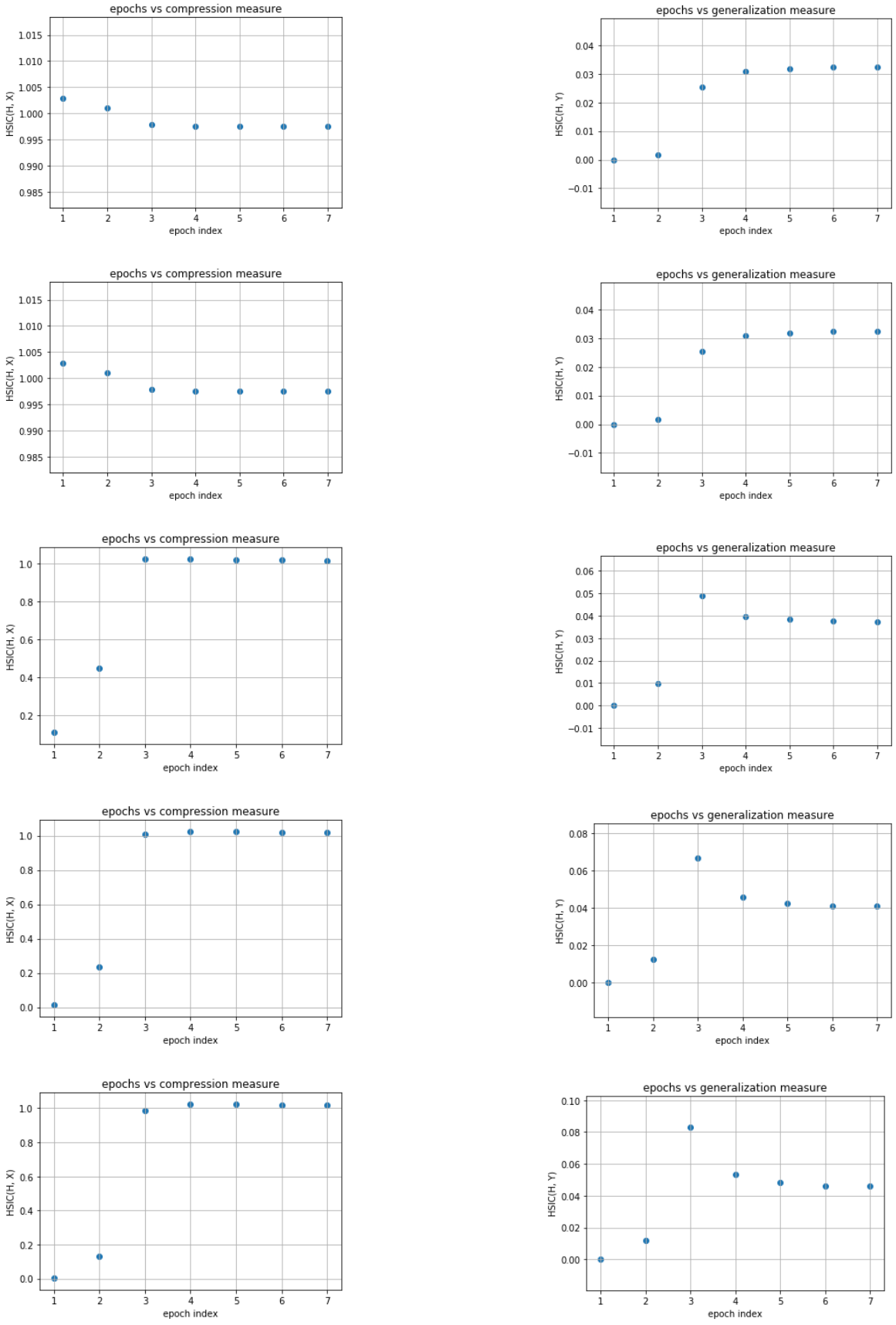

Figure 1.3: Plots between epochs vs HSIC(X, H) (left) and epochs vs HSIC(H, Y) (right) for all parametric layers

1.6 The famous IB Controversy

After the elegant theory of information bottleneck principle was introduced it was welcomed with a great enthusiasm and applaud in the deep learning community. The reason was because of it's shear potential to explain learning dynamics in details which was almost inaccessible and seems impossible black box before, and that too from an information theoretic approach which is a very established field in itself and has a strong formalism of it's own. But soon it was attacked by many authors and research group from all over the world because of non-reproduciblity of results which were quotes in the original IB paper [3]. One of the strong attacking paper [7] which claimed that all the IB predictions are not at all general and are an artifact of careful setup, architecture and experimentation techniques. They denied all the three core predictions provided by IB

- Deep neural networks undergo two distinct phases consisting of an initial fitting phase and a subsequent compression phase.
- The compression phase is related to the generalization ability of deep neural networks.
- The compression phase occurs due to the diffusion-like behavior of stochastic gradient descent.

This came out to be very interesting and long discussion between the authors of the two contradictory papers [3] and [7]. The claim of Tishby et al. was that the authors of attacking papers are not estimating mutual information with enough precision so as to observe the predictions of IB. Later the controversy was resolved by the paper [10] in which they introduced a very

robust and till now the most accurate mutual information estimator which reproduced all the predictions of IB. The method they used was based on dependency graphs and a brief about the method will be dealt in next section. After this another neural network based mutual information estimator was published in the paper [11] which was based on lower bound minimization of estimator function using neural networks. This paper also claimed the IB predictions are indeed true and information plane trajectory do show two phase dynamics. These interesting works on mutual information estimation solved the controversy in favour of IB principle and provided with a strong opinion of IB theory as one of the candidates to explain generalization in deep learning. Even after such enthusiasm for IB principle in the community it is still very naive for practical purposes and the algorithms based on IB learning dynamics are computationally very expensive without much scope of parallelization. Also to observe IB predictions and use it to improve learning of deep neural networks which are of practical importance we require a very accurate estimator with tighter lower bounds for large datasets which is still an open problem. More discussion on this is provided in the next section.

1.7 Problems with IB Principle

The important part of the above computation is estimation of mutual information which is challenging task in itself. There exist a lot of mutual information estimation methods and it was observed that the two phase observation which was claimed in [4] was sensitive to the methods used for estimating MI. This motivated researchers to use IB-theory not just for analysing the training dynamics but rather borrow the objective function itself as loss

function to train the network. These approaches are called variational information bottleneck models and one such example is given in [13] and another example which uses HSIC measure instead of MI is given in. The central idea of IB-theory is the attempt to give precise definition of optimal hidden representations and using this to define measure of generalization and compression abilities. These optimal hidden representations are expressed in terms of probability distribution which inherently assumes stochastic neural networks. Many of the researchers have shown the inadequacy of using MI as criterion to measure generalization and compression in case of non-stochastic neural networks (or deterministic networks). These deterministic networks suffer from critical theoretical issues when uses MI as criterion which is described in more detail in [12]. The same paper gives a rough sketch of what an optimal representation should truly capture in context of deterministic neural network. These sketches are based on the need for a representation to not only capture generalization and compression (as done in traditional IB-Theory) but also more exotic properties like simple decision rule and robustness to noise in the data. So most of the effort in this field is now concentrated on coming up with new criterion (like MI in traditional IB-Theory) which are expected to apparently solve the issues as described in [12]. The research for this new 'ideal' criterion will advance both the field i.e. one involving training dynamics analysis and other in which IB criterion is used as objective function to train the neural networks.

- The central idea of IB-theory is the attempt to give precise definition of optimal hidden representations and using this define measure of generalization and compression abilities.

- These optimal hidden representations are expressed in terms of probability distribution which inherently assumes stochastic neural networks.

- Mutual Information being a simple theoretical measure of dependence is a notoriously difficult task to estimate in practice when there is only fewer samples available of the random variable X and Y.

- Many of the researchers have shown the inadequacy of using MI as criterion to measure generalization and compression in case of non-stochastic neural networks (or deterministic networks). These deterministic networks suffer from critical theoretical issues when uses MI as criterion which is described in more detail in [12].

- The same paper gives a rough sketch of what an optimal representation should truly capture in context of deterministic neural network. These sketches are based on the need for a representation to not only capture generalization and compression (as done in traditional IB-Theory) but also more exotic properties like simple decision rule and robustness to noise in the data.

- So most of the effort in this field is now concentrated on coming up with new criterion (like MI in traditional IB-Theory) which are expected to apparently solve the issues as described in [12].

- The research for this new 'ideal' criterion will advance both the field i.e. one involving training dynamics analysis and other in which IB criterion is used as objective function to train the neural networks.

1.8 Relation of IB with Other approaches

The Information Bottleneck principle provides us with new learning bounds which are stronger than what is provided by statistical learning theory using trivial inequalities like Chernoff bounds etc. These bounds are give in the foundational papers [3]. On the other hand Arora et al. have also done foundational and parallel work in the theory of deep learning which has main emphasis on complexity theory of learning algorithms. They have been successful in providing a great insight in the asymptotic limits of the deep neural networks and how learning compression operates in these limits on which more can be found in these interesting papers [8] and [9]. These papers shows an exceptional similarity with the work on complexity bounds provided by IB principle. This hints us towards the unification of these seemingly different approaches and formalism. Complexity theory is derived from the statistical bounds provided by the learning algorithms and we have seen that statistical modelling has a deep connection with information theory and this can be a motivational argument to attempt for unification of these two approaches. A detailed work of Arora et al. can be found in the recent and exceptionally interesting papers [5], [6], [8] which give a general introduction to the convergence theory of deep learning.

1.9 General Workflow

This section describes the complete workflow in the IB theory research. There are following two main application of the IB Theory on which more can be found in [12].

1.9.1 Performance Analysis using IB Measures

This use case involves using IB measures to evaluate performance of the standard training algorithms from a information theoretic perspective. The analysis includes measure $I(X, H_i)$ and $I(H_i, Y)$ for every hidden layer H_i for all batches for epochs. This forms a trajectory of training in information plane ($I(X, H)$ vs $I(H, Y)$). The trajectory is then compared with the optimal curve on which information bottleneck objective is minimum and then measures of generalization and compression are defined as the difference between the two curves. This gives us a reasonable measure of how well the standard training algorithms have been able to implicitly imposed regularization during the training.

1.9.2 Training NN using IB Theory

This use case is the first direct application of the IB Theory to train neural networks. The objective of information bottleneck principle is directly used as loss function for training such networks. There exist two varieties of networks which directly uses IB objective. First is stochastic networks which are similar to variational autoencoders but rather than using reconstruction loss and KL Divergence they use variational lower bound of the IB objective for a encoding which is minimal sufficient statistics according to IB theory. One of the first paper to use such an approach to autoencoders were by Google [13]. The second class is deterministic networks which uses IB objective as their loss function. One of the most innovative recent advances in this direction is given in the paper [17]. The paper assumes for a particular layer the loss objective is only dependent on the parameters of that

layer itself and because of the nature of the IB objective (which takes only the input, the output and the hidden layer itself) this can be done. Now if this is the case then each layer is independent of one another and can be trained parallely with any backpropagation. In this algorithm each layer is forced to reach the optimal IB curve during it's training which means all the layers at the end are approximate minimal sufficient statistics of the output. This is the condition which is defined to be the state of the network containing as much "information" as possible about the output label.

1.10 Motivation for information geometric framework

In this section we briefly discuss some of the motivating arguments for a geometric formalism of information theory and in turn reformulate the framework of deep learning.

- Information bottleneck principle is one of those ensuring works that proves that information theory and deep learning is deeply interconnected and this calls for a more sophisticate and elegant theoretical framework in which this relationship is immediately obvious.

- Another motivating argument is the need for coordinate free language where the model is specified not by a particular choice of parameters but rather something more geometric. This is because we know that the same model can be reparametrized to a completely new set of parameters and this changes the learning dynamics which may converge to a completely different optimal parameters but this should not

be the case as the model is exactly the same. Having said that we need a framework where the learning dynamics should not change on reparametrization of the model.

- Another strong motivating argument is the intrinsic non-euclidean nature of information theory. We know that a common distance measure between two probabilistic models is defined by divergences in information theory and when we consider infinitesimal limit of this distance measure we get a metric tensor which is non-euclidean and hence this too directs us towards a geometric formalism of information theory.

In the subsequent chapter we will develop the subject of information geometry which precisely achieves the above goal of reformulating information theory in more geometric setting using the concepts of differential geometry. In next two chapters we will give an exhaustive mathematical introduction to the differential geometric prerequisites required for the subject of information geometry. This mathematical machinery is inevitably essential for anyone who is willing to contribute something positive in the field of theoretical deep learning. In further chapters we try to provide some recent and interesting application of this framework to the problems of deep learning.

Information Geometry

3.1 Introduction

In this chapter we will layout the modern language of information theory in terms of elegant mathematics of differential geometry. The field is called Information Geometry and it has found vast applications in the field of deep learning. Information Geometry is a coordinate-free formulation of decision-making problems in statistics and machine learning. In this chapter we will give an introduction to information manifolds and important structures which we can define on that to perform calculus related computations in a covariant coordinate-free way. We have taken the material heavily from [14] which provides an excellent modern introduction to the field of information geometry and anyone willing to get a more elaborate exposition should refer this.

3.2 Overview of Differential Geometry

manifolds, tangent space and vectors and forms, transition function, covariant derivative and parallel transport, curvature and torsion, levi-civita connections

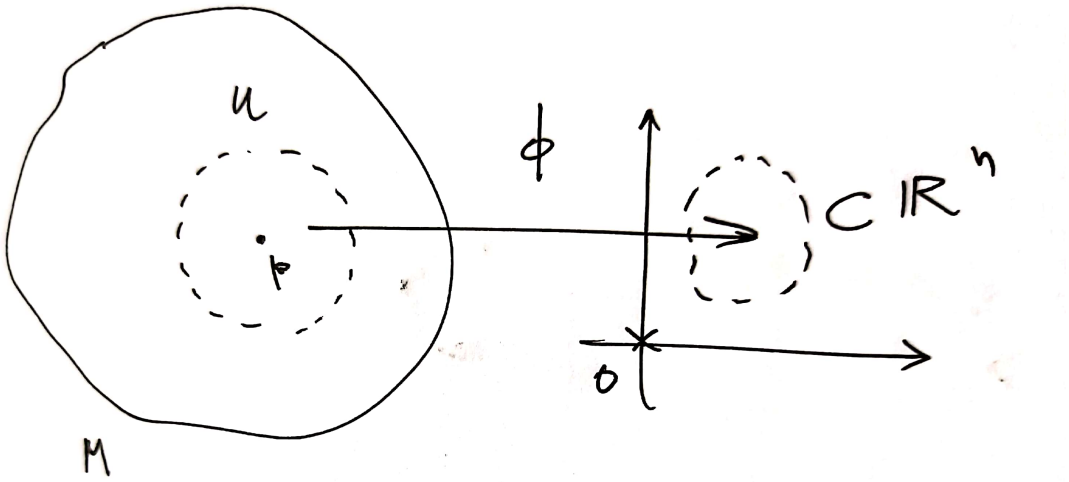

Figure 3.4: A geometric depiction of manifold and local homeomorphism to $\mathbb{R}^n$

3.3 Manifolds

Manifolds are topological spaces (which means that we have a notion of open sets on the space and have the ability to comment on related notions like continuity, connectedness, compactness) which are **locally** homeomorphic to euclidean space $\mathbb{R}^n$. We stress upon this statement as manifolds are only locally look alike of euclidean spaces but globally they can have a highly non-trivial properties. These global properties are captured by topological notions like homotopy and homology groups from algebraic topology. In manifold theory we will be dealing with differential structures on the space which allows us to do calculus on the space in a coordinate free form. As calculus treats only local quantities we have no way to determine global non-trivial properties of the underlying topological space and hence we will not go into that direction. Having said that we move on to formally define manifolds. Let

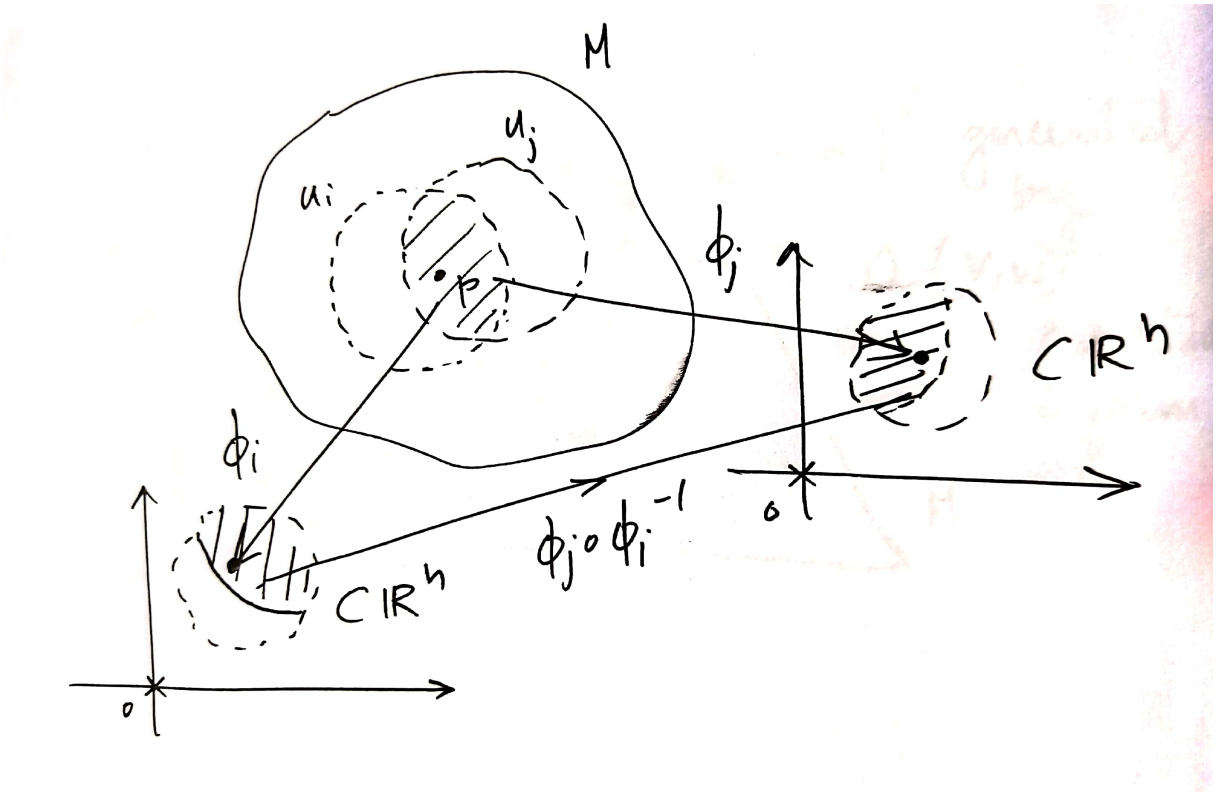

Figure 3.5: Two overlapping charts U_i, U_j and their respective mappings ϕ_i, ϕ_j and transition function $\phi_j \circ \phi_i$

$(M, \mathfrak{U})$ be a topological space with topology $\mathfrak{U}$, then M is a manifold if it satisfies following properties:

- For an open cover $\{U_i\}$ of M, $\bigcup_i U_i = M$, there exists homeomorphism $\psi_i \colon U_i \to P \subset \mathbb{R}^n$ and so a manifold is equipped with pair (U_i, ψ_i). The ψ_i is called as coordinate charts.

- For two different pairs (U_i, ψ_i) and (U_j, ψ_j), the mappings $\psi_j \circ \psi_i^{-1} \colon \psi_i(U_i \cap U_j) \to \psi_j(U_i \cap U_j)$ and $\psi_i \circ \psi_j^{-1} \colon \psi_j(U_i \cap U_j) \to \psi_i(U_i \cap U_j)$ should be infinitely differentiable.

The latter condition ensures that calculus related statements in one chart holds in any other chart too without any ambiguity. Collection of charts fol-

lowing above compatiblity condition is called as an atlas. If there exists some other atlas which has charts compatible with the charts of the first altas, then their union is also an atlas and we say the two atlases are compatible (this is an equivalence relation). In this way we can form a maximal altas which defines a differentiable structure on the manifold. Intutively a differential structure provides a way to carry on calculus on the manifold with ambiguity.

3.4 Vectors

We now define the notion of tangent space which will provide us with the basic objects called as vectors using which we can layout a theory of calculus on manifolds. Let us consider a curve $\gamma\colon [0,1] \to M$ and consider a point $p = \gamma(0)$. Given a smooth function $f \in C\infty(M)$ we define a tangent vector V to this curve at point p as

$$V\colon C^{\infty}(M) \to \mathbb{R}$$

We can obtain a coordinate representation of this definition by considering that the curve lies in the open set U and we have a coordinate chart $x\colon U \to \mathbb{R}^n$. Then we have,

$$\begin{aligned} V[f] &= \frac{\mathrm{d}f \circ \gamma(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0} \\ &= \frac{\mathrm{d}f \circ x^{-1} \circ x \circ \gamma(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0} \\ &= \frac{\mathrm{d}x^{\mu} \circ \gamma(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0} \frac{\partial f \circ x^{-1}(x^{\mu})}{\partial x^{\mu}}\Big|_{p} \\ &= \frac{\mathrm{d}x^{\mu}(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0} \frac{\partial}{\partial x^{\mu}}\Big|_{p}(f) \end{aligned}$$

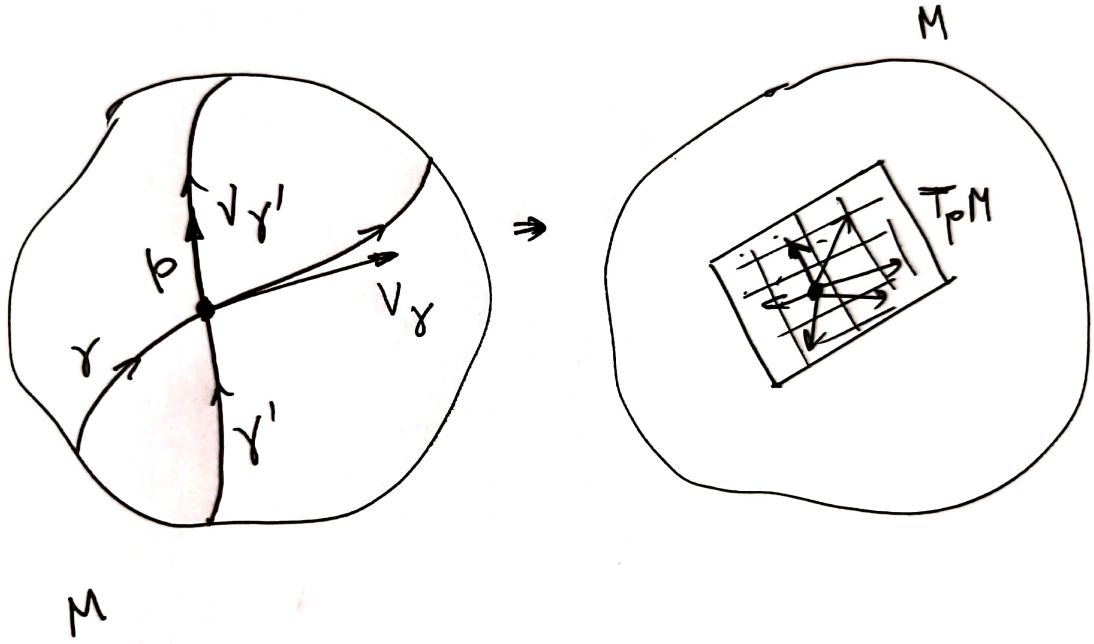

Figure 3.6: An illustrative diagram of a vector to the curve on a manifold and resultant tangent space formed by all such curves passing from p

and so we can write tangent vector in coordinate basis $\frac{\partial}{\partial x^\mu}\Big|_p$ as

$$V_\gamma = \frac{\mathrm{d}x^\mu(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0} \frac{\partial}{\partial x^\mu}\Big|_p$$

We now define a tangent space T_pM as

$$T_pM = \{V_\gamma | \gamma(0) = p\}$$

where γ is all possible smooth curves passing through point p at $\lambda = 0$. This space has a structure of vector space with addition and scalar multiplication corresponding to $\mathbb{R}$.

$$V[f] \coloneqq \frac{\mathrm{d}f \circ \gamma(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0}$$

3.5 Covectors

We now know that T_pM has a vector space structure and so we can thus define a covector space of real valued mapping $\omega\colon T_pM \to \mathbb{R}$ as

$$\omega(V) \in \mathbb{R}$$

The space of all such ω is called as cotangent space T_p^*M

$$T_p^*M = \{\omega | \omega(V) \in \mathbb{R}, V \in T_pM\}$$

3.5.1 One-forms

We now define another type of mathematical objects called as forms and in particular one-forms which will be useful in obtaining basis for T_p^*M. First consider a smooth function $f \in C^\infty(M)$, we define an object $df \in T_p^*M$ which is defined by it's action on vectors as

$$df(V) \coloneqq V[f] \in \mathbb{R}$$

This object is called as gradient of a function f. We now define dual basis ω_α of T_p^*M, which are dual to the coordinate basis of T_pM as

$$\omega_\alpha(\frac{\partial}{\partial x^\beta}\Big|_p) \coloneqq \delta^\alpha_\beta$$

We can already guess the dual basis to be dx^α as

$$\begin{aligned} dx^\alpha(\frac{\partial}{\partial x^\beta}\Big|_p) &= \frac{\partial x^\alpha}{\partial x^\beta}\Big|_p \\ &= \delta^\alpha_\beta \end{aligned}$$

Following this we can know obtain components of any vector in a particular coordinate basis as

$$\begin{aligned} dx^\alpha(V) &= dx^\alpha(V^\beta \partial_\alpha) \\ &= V^\beta dx^\alpha(\partial_\alpha) \\ &= V^\beta \frac{\partial x^\alpha}{\partial x^\beta}\Big|_p \\ &= V^\beta \delta^\alpha_\beta \\ &= V^\alpha \end{aligned}$$

where ∂_α is shorthand notation for $\frac{\partial}{\partial x^\alpha}\Big|_p$. We can invert this and obtain any $\omega = \omega_\alpha dx^\alpha \in T^*_p M$ in dual coordinate basis dx^α as

$$\begin{aligned} \omega(\partial_\alpha) &= \omega_\beta dx^\beta(\partial_\alpha) \\ &= \omega_\beta \frac{\partial x^\beta}{\partial x^\alpha}\Big|_p \\ &= \omega_\alpha \end{aligned}$$

We have thus constructed two vector spaces $T_p M$ and $T^*_p M$ on a point $p \in M$ and their natural coordinate induced basis $\{\partial_\alpha, dx^\alpha\}$.

3.6 Tensors

We can define a more general multi-linear (k, l) mapping called as tensor

$$T : T^*_p M \otimes \cdots \otimes T^*_p M \otimes T_p M \otimes \cdots \otimes T_p M \to \mathbb{R}$$

with k slots for T^*_M and l slots for $T_p M$. In tensor coordinate basis formed by outer product we can express a (k, l) tensor as

$$T = T^{\mu_1 \ldots \mu_k}_{\nu_1 \ldots \nu_l} \partial_{\mu_1} \otimes \cdots \otimes \partial_{\mu_k} \otimes dx^{\nu_1} \otimes \cdots \otimes dx^{\nu_l}$$

3.7 Metric tensor fields

We have tangent space T_pM at each point $p \in M$ and so we can endow it with a metric space structure which will let us talk about length of curves on the manifold. We can endow completely unrelated metric space structure on tangent spaces at different points on the manifold but we are interested in the case when the assignment of metric is smooth as we vary over points on the manifold. So considering smooth metric space structure on the whole manifold gives us a notion of metric tensor field $g\colon T_pM \times T_pM \to \mathbb{R}$ as

$$g(U, V) = \langle U, V \rangle$$

for any $U, V \in T_pM$. In coordinate basis, we can represent metric tensor field as

$$g = g_{\alpha\beta}(p) dx^\alpha \otimes dx^\beta \tag{3.1}$$

and smoothness is the condition that $g_{\alpha\beta}(p)$ as a scalar function is infinitely differentiable. Availability of a metric tensor field provides us with a canonical mapping between T_pM and T_p^*M. Let us consider $V \in T_pM$ and the mapping $g(V,)\colon T_pM \to \mathbb{R}$ then we have

$$g(V,)\colon V \mapsto g_{\alpha\beta}V^\alpha \in T_p^*M$$

3.8 Affine Connection

At the start of this chapter we stated that we will only deal with the theory of calculus on manifolds. The basic operation in calculus is differentiation and integration. Till now we have defined the objects of which we will take the derivative but did not talked anything about taking derivative or what is

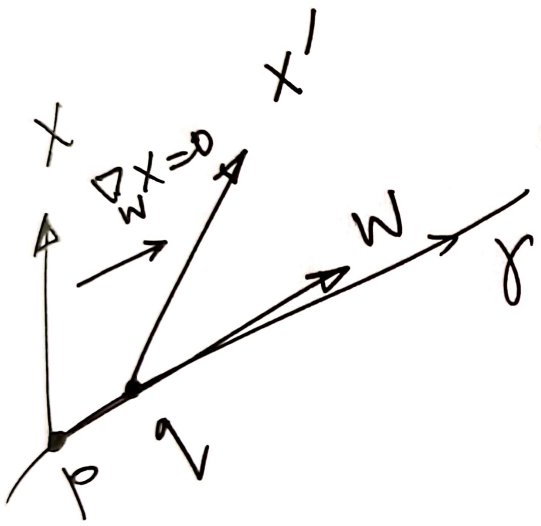

Figure 3.7: A diagram showing parallel transport of a vector X along a curve with tangent vector W

a derivative on a manifold in first place. We will first ask a very basis question, "Why is it so difficult to compute a derivative on a manifold ?" and answer to that question is that vectors or covectors or tensors live in vector spaces that are defined on a point and there is no way to relate them to corresponding vector spaces on some other point as they are totally different spaces altogether. We will now define a very important way of connecting two tangent spaces at two different points which we will call parallel transport. Intuitively for each point $p \in M$ we will map every vector $V \in T_pM$ to it's **parallel transported** vector on a nearby point $q \in M$ at T_qM. Now consider a point p with coordinates $x^\mu(p)$ and a vector $V^\alpha(p) \in T_pM$ in coordinate basis $\frac{\partial}{\partial x^\alpha}\Big|_p$. We also consider a curve $\gamma(\lambda)$ such that $\gamma(0) = p$ and $q = \gamma(\Delta\lambda)$ such that vector $W^\beta \in T_pM$ is tangent to this curve at point p. We

define a parallel transported version $V'^{\alpha}(q)$ of V^{α} from point $p \in M$ to point $q \in M$ which is linear in both V^{α} and W^{β} as

$$V'^{\alpha}(q) = V^{\alpha}(p) - \Gamma^{\alpha}_{\beta\gamma}(p) W^{\gamma} V^{\beta} \Delta\lambda$$

where $\Gamma^{\alpha}_{\beta\gamma}$ are christoffel symbols which stores the information of how much the space is curved because if the space would have been flat then the parallel transported version would have been exactly the same $V'^{\alpha}(q) = V^{\alpha}$. We can define a covariant derivative as

$$\nabla_W V \coloneqq \lim_{\Delta\lambda \to 0} \frac{V^{\alpha}(q) - V'^{\alpha}(q)}{\Delta\lambda} \frac{\partial}{\partial x^{\alpha}}\Big|_q$$

which can further we simplified

$$\begin{aligned}
\nabla_W V &= \lim_{\Delta\lambda \to 0} \frac{V^{\alpha}(q) - V^{\alpha}(p) - \Gamma^{\alpha}_{\beta\gamma}(p) W^{\gamma} V^{\beta} \Delta\lambda}{\Delta\lambda} \\
&= \lim_{\Delta\lambda \to 0} \frac{V^{\alpha} \circ x^{-1} \circ x(\lambda + \Delta\lambda) - V^{\alpha} \circ x^{-1} \circ x(\lambda) + \Gamma^{\alpha}_{\beta\gamma}(p) W^{\gamma} V^{\beta} \Delta\lambda}{\Delta\lambda} \\
&= \frac{\partial V^{\alpha}}{\partial \lambda} + \Gamma^{\alpha}_{\beta\gamma}(p) W^{\gamma} V^{\beta}
\end{aligned}$$

We can also express christoffel symbols in terms of covariant derivative as

$$\nabla_{\partial_\beta}(\partial_\alpha) = \Gamma^{\gamma}_{\alpha\beta} \partial_\gamma$$

3.8.1 Metric Compatible Connections

We now focus our attention towards a very special kind of parallel transport which is defined by invariance of angle between parallel transported vectors. This statement is the following condition

$$X[g(Y, Z)] = 0$$

for any $X, Y, Z \in T_pM$. To get from this abstract condition to a useful condition on christoffel symbols we further consider

$$\begin{aligned} X[g(Y,Z)] &= \nabla_X g(Y,Z) \\ &= (\nabla_X g)(Y,Z) + g(\nabla_X Y, Z) + g(Y, \nabla_Y Z) \\ &= (\nabla_X g)(Y,Z) = 0 \end{aligned}$$

and so finally we have

$$\nabla_X g = 0$$

The above condition in component forms is as follows

$$\begin{aligned} \nabla_\alpha(g_{\beta\gamma} dx^\beta \otimes dx^\gamma) &= \partial_\alpha g_{\beta\gamma} dx^\beta \otimes dx^\gamma + g_{\beta\gamma} \nabla_\alpha dx^\beta \otimes dx^\gamma + g_{\beta\gamma} dx^\beta \otimes \nabla_\alpha dx^\gamma \\ &= (\partial_\alpha g_{\beta\gamma} - g_{\lambda\gamma}\Gamma^\lambda_{\beta\alpha} - g_{\beta\lambda}\Gamma^\lambda_{\gamma\alpha}) dx^\beta \otimes dx^\gamma \end{aligned}$$

and so if

$$\partial_\alpha g_{\beta\gamma} - g_{\lambda\gamma}\Gamma^\lambda_{\beta\alpha} - g_{\beta\lambda}\Gamma^\lambda_{\gamma\alpha} = 0 \tag{3.2}$$

is satisfied we say that ∇ is metric compatible. Cyclic permutations of (α, β, γ) gives us two more equations

$$\partial_\gamma g_{\alpha\beta} - g_{\lambda\alpha}\Gamma^\lambda_{\gamma\beta} - g_{\alpha\lambda}\Gamma^\lambda_{\beta\gamma} = 0 \tag{3.3}$$

$$\partial_\beta g_{\gamma\alpha} - g_{\lambda\beta}\Gamma^\lambda_{\alpha\gamma} - g_{\gamma\lambda}\Gamma^\lambda_{\alpha\beta} = 0 \tag{3.4}$$

. We computer -(4.2) + (3.3) + (3.4)

$$-\partial_\alpha g_{\beta\gamma} + g_{\lambda\gamma}\Gamma^\lambda_{\beta\alpha} + g_{\beta\lambda}\Gamma^\lambda_{\gamma\alpha} + \partial_\gamma g_{\alpha\beta} - g_{\lambda\alpha}\Gamma^\lambda_{\gamma\beta} - g_{\alpha\lambda}\Gamma^\lambda_{\beta\gamma} + \partial_\beta g_{\gamma\alpha} - g_{\lambda\beta}\Gamma^\lambda_{\alpha\gamma} - g_{\gamma\lambda}\Gamma^\lambda_{\alpha\beta} = 0$$

which after some rearrangements and definitions becomes

$$(-\partial_\alpha g_{\beta\gamma} + \partial_\gamma g_{\alpha\beta} + \partial_\beta g_{\gamma\alpha}) + (g_{\lambda\gamma}\Gamma^\lambda_{\beta\alpha} - g_{\gamma\lambda}\Gamma^\lambda_{\alpha\beta}) + (g_{\beta\lambda}\Gamma^\lambda_{\gamma\alpha} - g_{\lambda\beta}\Gamma^\lambda_{\alpha\gamma}) - (g_{\lambda\alpha}\Gamma^\lambda_{\gamma\beta} + g_{\alpha\lambda}\Gamma^\lambda_{\beta\gamma}) = 0$$

$$(-\partial_\alpha g_{\beta\gamma} + \partial_\gamma g_{\alpha\beta} + \partial_\beta g_{\gamma\alpha}) + g_{\lambda\gamma}T^\lambda_{\alpha\beta} + g_{\beta\lambda}T^\lambda_{\gamma\alpha} - 2g_{\lambda\alpha}\Gamma^\lambda_{(\gamma\beta)} = 0$$

$$(-\partial_\alpha g_{\beta\gamma} + \partial_\gamma g_{\alpha\beta} + \partial_\beta g_{\gamma\alpha}) + T_{\gamma\alpha\beta} + T_{\beta\gamma\alpha} - 2g_{\lambda\alpha}\Gamma^\lambda_{(\gamma\beta)} = 0$$

where T is torsion tensor. We finally get the following expression which is satisfied by any metric compatible connection

$$\Gamma^\alpha_{(\gamma\beta)} = \frac{1}{2}g^{\alpha\lambda}(-\partial_\lambda g_{\beta\gamma} + \partial_\gamma g_{\lambda\beta} + \partial_\beta g_{\gamma\lambda}) + \frac{1}{2}(T_{\gamma\ \beta}^{\ \alpha} + T_{\beta\gamma}^{\ \ \alpha})$$

and we also know $T^\alpha_{\beta\gamma} = 2\Gamma^\alpha_{[\beta\gamma]}$, combining the two terms $\Gamma^\alpha_{(\beta\gamma)} + \Gamma^\alpha_{[\beta\gamma]} = \Gamma^\alpha_{\beta\gamma}$, we have

$$\Gamma^\alpha_{\beta\gamma} = \frac{1}{2}g^{\alpha\lambda}(-\partial_\lambda g_{\beta\gamma} + \partial_\gamma g_{\lambda\beta} + \partial_\beta g_{\gamma\lambda}) + \frac{1}{2}(T_{\gamma\ \beta}^{\ \alpha} + T_{\beta\gamma}^{\ \ \alpha} + T^\alpha_{\beta\gamma})$$

If we consider the case when connection is torsion free , that is $T = 0$ then we have a unique metric compatible torsion free connection called as Levi-Civita connection ${}^{LC}\nabla$

$${}^{LC}\Gamma^\alpha_{\beta\gamma} = \frac{1}{2}g^{\alpha\lambda}(-\partial_\lambda g_{\beta\gamma} + \partial_\gamma g_{\lambda\beta} + \partial_\beta g_{\gamma\lambda})$$

3.9 Information Manifolds

Information manifolds $(M, g, \nabla, \nabla^\star)$ are spaces of allowed parameters for a particular statistical decision-making problem. Unlike Riemannian manifolds, information manifolds have a dual-structure formed by conjugate connection pair $(\nabla, \nabla^\star)$ and thus manifolds in information geometry is called Conjugate Connection Manifolds (CCMs).

3.10 Conjugate-Connection structure

A complete treatment of the theory of connections is discussed in the next chapter and in this chapter we will only consider some very basic notions related to connections on CCMs. Given any arbitrary connection ∇ on the manifold we can define a dual connection $\nabla^\star$ as follows

$$X[g(Y,Z)] = g(\nabla_X Y, Z) + g(Y, \nabla^\star_X Z)$$

for any arbitrary vector fields X, Y and Z.

Note: $(\nabla^\star)^\star = \nabla$.

An equivalent definition can be

$$\langle U, V\rangle = \langle \prod_{\gamma(t)}^{\nabla} U, \prod_{\gamma(t)}^{\nabla^\star} V\rangle$$

which in component form is

$$\begin{aligned} g_{\alpha\beta}U^\alpha V^\beta &= g_{\alpha\beta}(U^\alpha + \Gamma^\alpha_{\rho\sigma}U^\rho X^\sigma \epsilon)(V^\beta + \tilde{\Gamma}^\beta_{\rho\sigma}V^\rho X^\sigma \epsilon) \\ &= g_{\alpha\beta}U^\alpha V^\beta + g_{\alpha\beta}U^\alpha \tilde{\Gamma}^\beta_{\rho\sigma}V^\rho X^\sigma \epsilon + g_{\alpha\beta}V^\beta \Gamma^\alpha_{\rho\sigma}U^\rho X^\sigma \epsilon + O(\epsilon^2) \end{aligned}$$

which can finally be stated as

$$U_\beta \tilde{\Gamma}^\beta_{\rho\sigma}V^\rho X^\sigma + V_\alpha \Gamma^\alpha_{\rho\sigma}U^\rho X^\sigma = 0 \tag{3.5}$$

We can also observe that this dual structure $(\nabla, \nabla^\star)$ preserves the metric for vectors X, Y parallelly transported according to ∇ and $\nabla^\star$ respectively. We can further define a connection $\tilde{\nabla} = \frac{1}{2}(\nabla + \nabla^\star)$ which we can see is self-dual and by the fundamental theorem of Riemannian Geometry there is a unique self dual connection which is levi-civita connection ${}^{LC}\nabla = \tilde{\nabla}$.

3.11 Statistical Manifolds

A pair $(\nabla, \nabla^\star)$ defines what is called as Amari-Chentsov tensor $C_{ijk} = \Gamma^k_{ij} - \tilde{\Gamma}^k_{ij}$ which is totally symmetric in it's indices. In coordinate-free form the Amari-Chentsov tensor can be written as

$$C(X, Y, Z) \coloneqq \langle \nabla_X Y - \nabla^\star_X Y, Z \rangle \tag{3.6}$$

. A statistical manifold (M, g, C) is a manifold equipped with a metric tensor g and a totally symmetric tensor C. We can further define a one-parameter family of conjugate pairs of connections $(\nabla^{-\alpha}, \nabla^\alpha = (\nabla^{-\alpha})^\star)$ which is discussed in the following section.

3.12 Family of Conjugate Connections - $(M, g, \nabla^{-\alpha}, \nabla^\alpha)$

Given a statistical manifold (M, g, C) we can define family of conjugate connection pairs $(\nabla^{-\alpha}, \nabla^\alpha)$ by making an observation that given C, α is also a totally symmetric tensor and should corrosponds to Amari-Chentsov tensor for some conjugate connection pair $(\nabla^{-\alpha}, \nabla^\alpha)$

$$\Gamma^\alpha_{ijk} = \Gamma^0_{ijk} - \frac{\alpha}{2} C_{ijk}$$
$$\Gamma^{-\alpha}_{ijk} = \Gamma^0_{ijk} + \frac{\alpha}{2} C_{ijk}$$

where Γ^0_{ijk} is levi-civita connection coefficients. We can also write the above connections in terms of $(\nabla, \nabla^\star)$

$$\Gamma^\alpha_{ijk} = \frac{1+\alpha}{2} \Gamma_{ijk} + \frac{1-\alpha}{2} \tilde{\Gamma}_{ijk} \tag{3.7}$$

3.12.1 Fundamental Theorem of Information Geometry

Theorem 3.12.1 *If a torsion-free affine connection ∇ has constant curvature κ then it's conjugate torsion-free $\nabla^\star$ has necessarily the same constant curvature κ. This results into a very useful corollary.*

Corollary 3.12.1.1 *A manifold $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ is ∇^{α}-flat if and only if it is $\nabla^{-\alpha}$-flat.*

This corollary is important because as we will see in subsequent sections that dually-flat connections results from very interesting implications from statistics and machine learning. The are also widely used because of their ease with pratical computations.

3.13 Canonical CCM structures

Till now we have discussed a general theory of information geometry but never discussed how to get g and $(\nabla, \nabla^\star)$. We will try to impose as less extra structure as possible to obtain the above quantities and this will result into a natural canonical way of forming an information manifold for any general problem in machine learning and deep learning. We can proceed in two direction

- From divergence considered in the problem to capture dissimilarity between probability distributions.
- From probability distribution itself by using max-log likelihood principle.

We start by discussing the first method for which we define formally what a divergence is in the next section.

3.13.1 Conjugate connection from divergences

Before defining what a divergence is, we wil define some important notations which will be with us in the whole chapter:

$$\partial_{i,.} f(x,y) = \frac{\partial f(x,y)}{\partial x^i}$$

$$\partial_{.,j} f(x,y) = \frac{\partial f(x,y)}{\partial y^j}$$

$$\partial_{ij,k} f(x,y) = \frac{\partial^2}{\partial x^i \partial x^j} \frac{\partial f(x,y)}{\partial y^k}$$

Now we can give formal definition of a divergence as follows.

Divergence:

$D\colon M \times M \rightarrow [0,\infty)$ on a manifold M with a local chart $\theta \subset \mathbb{R}^D$ with following properties

- $D(\theta : \theta') \geq 0 \; \forall \; \theta, \theta' \in \Theta$ where equality holds if and only if $\theta = \theta'$.
- $\partial_{i,.} D(\theta, \theta')\Big|_{\theta=\theta'} = \partial_{.,j} D(\theta, \theta'\Big|_{\theta=\theta'} = 0 \; \forall i, j$
- $-\partial_{.,i}\partial_{.,j} D(\theta, \theta')\Big|_{\theta=\theta'}$ is positive-definite.

We can also define a dual-divergence by swapping the arguments

$$D^\star(\theta, \theta') = D(\theta, \theta')$$

We have stated many times in previous chapters that divergences are a measure of dissimilarity between probability distributions. This is in a way finite

counter-part of what metric tensor captures. So we can define metric tensor as the divergence between parameters θ and $\theta + \delta\theta$

$$\begin{aligned} D(\theta, \theta + \delta\theta) &= D(\theta, \theta) + \frac{\partial}{\partial \theta^i} D\Big|_\theta \delta\theta + \frac{\partial^2}{\partial \theta^i \partial \theta^j} D\Big|_\theta \delta\theta^i \delta\theta^j \\ &= \frac{\partial^2}{\partial \theta^i \partial \theta^j} D\Big|_\theta \delta\theta^i \delta\theta^j \end{aligned}$$

and so using this we define the conjugate connection structure naturally induced by divergence $D(\theta, \theta')$

- $g \coloneqq -\partial_{i,j} D(\theta, \theta')\Big|_{\theta=\theta'}$
- $\Gamma_{ijk} \coloneqq -\partial_{ij,k} D(\theta, \theta')\Big|_{\theta=\theta'}$
- $\tilde{\Gamma}_{ijk} \coloneqq -\partial_{k,ij} D(\theta, \theta')\Big|_{\theta=\theta'}$

We can further form Amari-Chentsov tensor and further form one parameter family of conjugate connection pairs and we thus we get a very rich structure of information manifold canonically derived from just using a divergence measure. Now we move on to the second method of defining conjugate connection structure using probability distribution.

3.13.2 Conjugate connection from parametric probability distribution

Let $\mathfrak{P}$ be a parametric family of probability distribution

$$\mathfrak{P} = \{p_\theta(X)\}_{\theta \in \Theta}$$

where Θ is parameter space. Order of parameter space is dimension of parameter space. We use the familiar log-likelihood function

$$l(\theta; x) = \log p_\theta(x)$$

and define score vector

$$s_\theta = \nabla_i l = \frac{\partial}{\partial \theta^i} l(\theta; x)$$

and subsequently define a metric g first defined by ..... known as Fisher-Rao Matrix (FIM)

$$I(\theta) = g_{ij} = \mathbb{E}_\theta[\partial_i l \partial_j l]$$

One immediate glimpse of how important this distance measure is can be observed from the following Cramer-Rao lower bound on variance of estimator

$$Var_\theta[\hat{\theta}_n(x)] \geq \frac{1}{n} I^{-1}(\theta)$$

and so minimizing variance between the model and the optimal $\theta^\star$ we tend to minimize actually the Fisher-Rao metric on the information manifold. Observe that the above metric involves averaging over all the data points and thus the quantities we will further define are structures on an expected - α information manifold[3]. Now we define two types of connections which is naturally defined in terms of the above probability distribution

- Exponential connection - ${}^e\Gamma_{ijk} := \mathbb{E}_\theta[(\partial_i \partial_j l)\partial_k l]$
- Mixing connection - ${}^m\Gamma_{ijk} := \mathbb{E}_\theta[(\partial_i \partial_j l + \partial_i l \partial_j l)\partial_k l]$

[3]It is interesting to study fluctuations about this average and observe some non-trivial effects on the information manifold

and if we consider them the conjugate connections then we get Amari-Chentsov tensor also known as skewness tensor in this context

$$C_{ijk} = \mathbb{E}_\theta[\partial_i l \partial_j l \partial_k l]$$

and using this we can form one-parameter family of conjugate connections $(\nabla^\alpha, \nabla^{-\alpha})$

$$\Gamma^\alpha_{ijk} := -\frac{1+\alpha}{2} C_{ijk} = \mathbb{E}_\theta[(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l)(\partial_k l)]$$

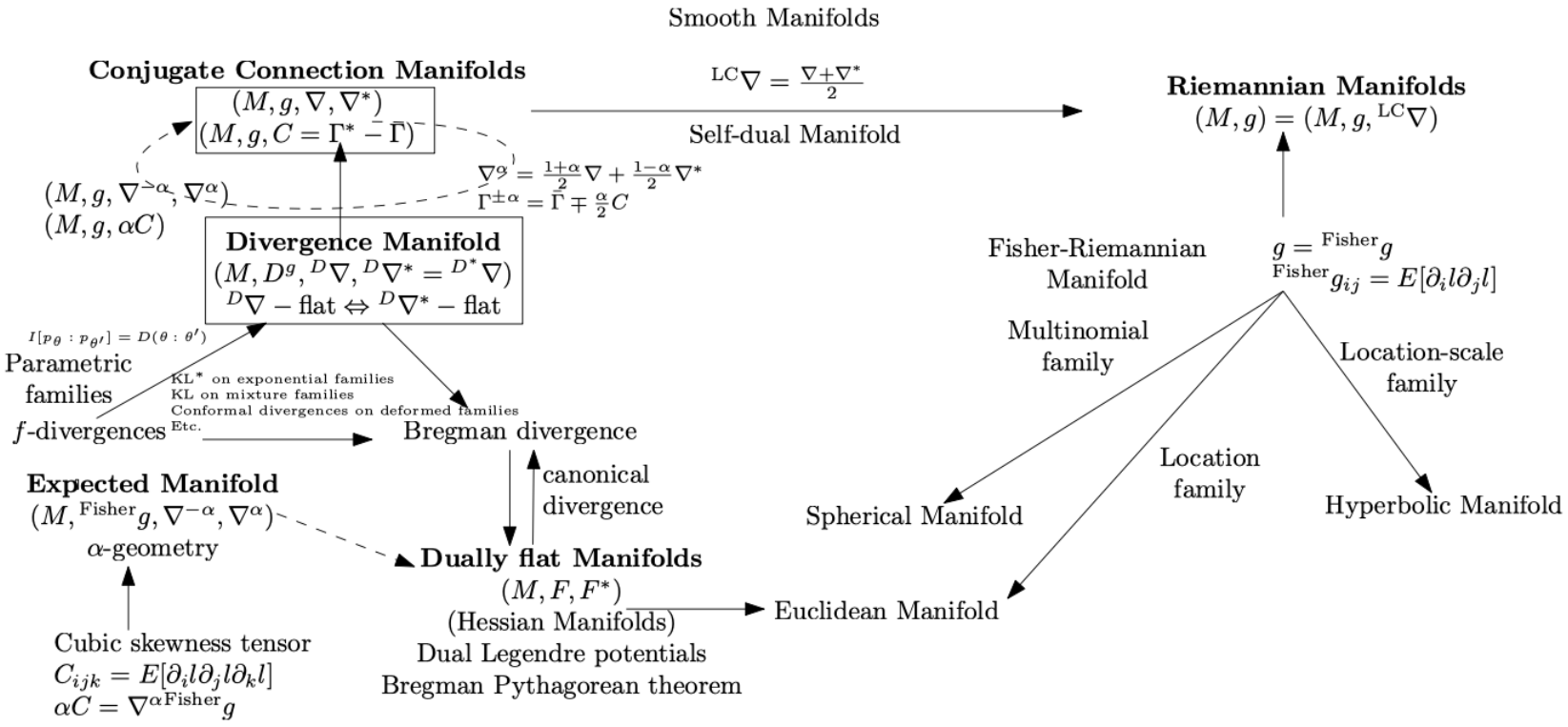

Figure 7: Overview of the main types of information manifolds with their relationships in information geometry.

Theory of Connections

4.1 Introduction

This chapter gives an extensive and exhaustive introduction to the structure of principal bundle and on them a structure called as "connections". Intuitively connections as the name suggests provide an unambiguous way to connect points in the nearby fibre space. These then further provides us with a definition of parallel transport of points in the principal bundle. The axioms of connections are realized through a lie-algebra valued one-form on the principal bundle. We will see that these Yang-Mills fields follow a very particular transformation law under the right action of the structure group. These transformation laws are a result of compatibility conditions which will be discussed in subsequent sections. After that we move on to define covariant derivative on principal bundle and induced covariant derivative on associated bundle and discuss in complete detail all the necessary machinery to do practical calculations in the subject. Using then developed formalism we will define more interesting structure of curvature forms and holonomy. We then briefly discuss the case of frame bundle in which we derive all the above quantities and show their equivalence with the commonly known quantities in differential geometry.

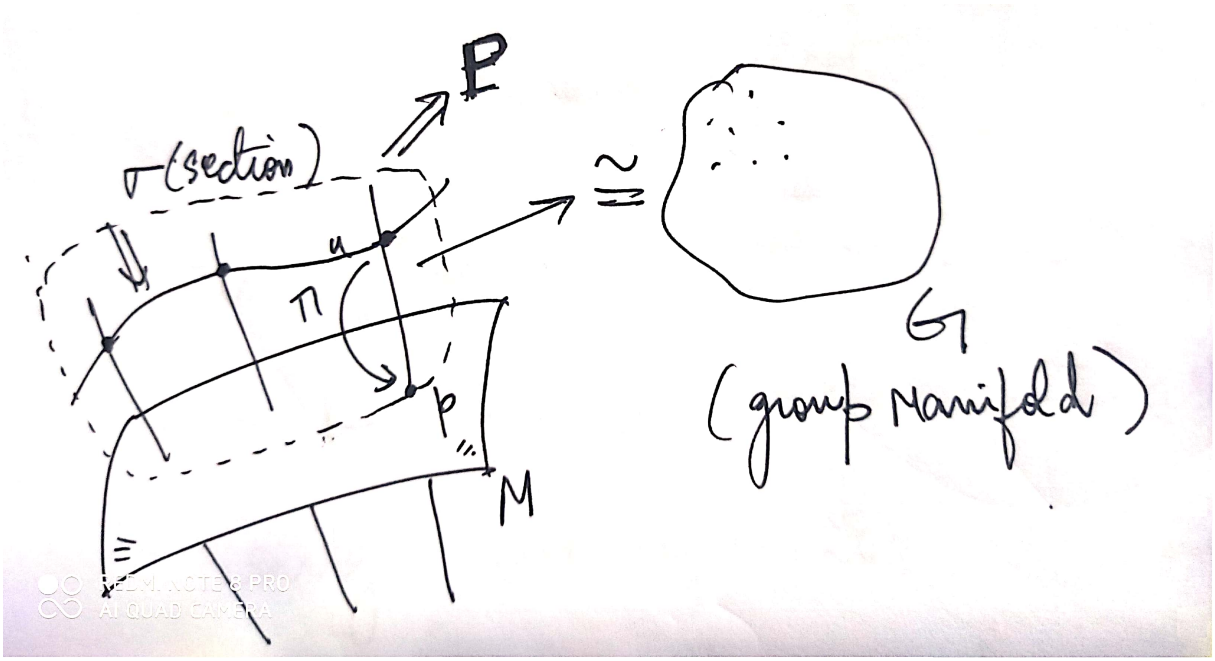

Figure 4.8: Illustration depicting a typical principal bundle showing the fibre space isomorphism with structure group manifold

4.2 Mathematical Foundations

We have principal bundle P and base manifold M with structure lie-group G with lie-algebra $\mathfrak{g} \cong T_eG$ with projection map $\pi\colon P \to M$. As we know $M \cong {}^{P}\!/_{G}$ from the definition of principal bundle and so $\pi^{-1}(p) \cong G$ where $p \in M$ because the right action is free. Thus we have a natural right action of the group on the principal bundle

$$\Psi\colon P \times G \to P$$

$$\Psi\colon (u, g) \mapsto u.g$$

where $u \in P$ and $g \in G$. We can also define the function

$$\Psi_g \colon P \to P$$

defined by $\Psi_g(u) = \Psi(u, g)$. In practical calculations we always work along with a local section $\sigma \colon U \to \pi^{-1}(U)$ defined on chart U with the obvious property of $\pi \circ \sigma = id_M$. [4] We will use $\Psi(u, g)$ and $u.g$ interchangeably whenever there is no scope of confusion.

4.2.1 Section induced local trivialization

We can use section to define local trivialization on chart U. Let $u \in P$ such that $\exists\ g \in G$ such that $\Psi(\sigma(\pi(u)), g) = u$ and so we can define local trivialization as

$$\chi \colon P \to U \times G$$

$$\chi \colon u \mapsto (\pi(u), g)$$

In fact given a local trivialization we can define a section.
Consider we have been given local trivialization $\chi^{-1} \colon U \times G \to P$ and we want a natural section $\sigma \colon U \to P$ corresponding to this local trivialization. We can choose

$$\sigma \colon p \mapsto \chi^{-1}(p, e)$$

which is a section as it satisfies $\pi \circ \sigma = id_M$. This proves that local trivialization and local section are in one-to-one correspondence.

[4] Global sections does not always exists eg. No-hair theorem for sphere !

4.2.2 Maurer-Cartan one-form

A Maurer-Cartan form $\Xi_g : T_gG \to T_eG \cong \mathfrak{g}$ takes a left invariant vector $V \in T_gG$ and maps it to it's generator $A \in \mathfrak{g}$ such that $(L_g)_*A = V$. To explain in detail, consider the curve $\gamma(t) = e.\exp(tA)$ such that $\gamma(t_o) = g$ and so we have a natural curve originating from g given by $c(t) = g.\exp(tA)$ such that V is tangent vector to this curve and so $\Xi_g(V) = A$. Fundamental theorem of ODE's ensures us that such a mapping is well-defined for all $V \in T_gG$

4.2.3 Connections

Connections is an abstract structure on the principal bundle which provides unambiguous partitioning of tangent space (T_uP) into vertical subspace (V_uP) and horizontal subspace (H_uP)[5]. A brief introduction to vertical subspace V_uP is as follows:
For any $A \in T_eG \cong \mathfrak{g}$, we can induce a left-invariant vector field at each point of the group manifold G by

$$(L_g)_*(A) = V\Big|_g$$

Now given initial point e and vector $A \in T_eG$ we can form integral curve of the above induced vector field which from the fundamental theorem of ODE's is unique given by

$$\gamma(t) = \exp(tA) \in G$$

[5]Readers are assumed to be familiar with these terms.

Now we can induce a curve on the principal bundle originating at point $u \in P$ by the right action of the above group

$$c(t) = u.\exp(tA) \in P$$

which further gives us an induced vector $X \in T_uP$at point u defined as

$$X[f] = \frac{\mathrm{d}f(u.\exp(tA))}{\mathrm{d}t}\Big|_{t=0}$$

where $f \in C^\infty(P)$ and we say $X \in V_uP$ and $A \in \mathfrak{g}$ is called the generator of the induced vector X. The induced curve at u is completely inside the fibre space at the point $p = \pi(u)$ on the base manifold and hence the name vertical subspace.
The partitioning demands the following properties:

1. $T_uP = V_uP \bigoplus H_uP$.
2. A smooth vector field X on P is separated into smooth vector fields $X = X^V + X^H$.
3. $H_{u.g}P = (R_g)_* H_uP$

These abstract specifications can be practically obtain using a lie-algebra value connection one-form on the principal bundle at each point $u \in P$.

4.2.4 Connection one-form

Connection one-form $\omega\colon T_uP \to \mathfrak{g}$ where $u \in P$ have the following defining properties

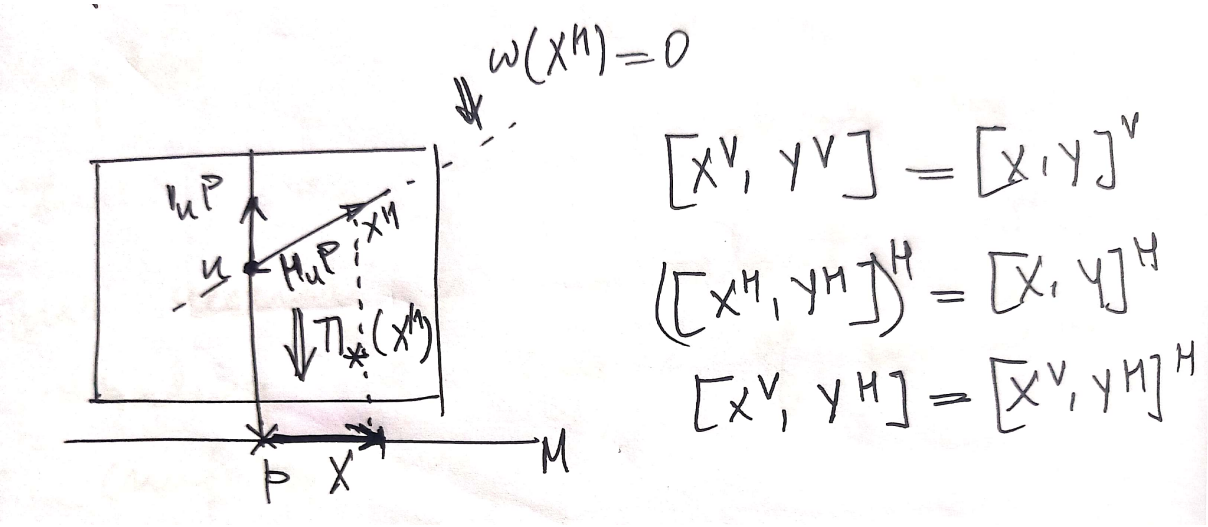

Figure 4.9: Diagram showing intuitively the projections defined by connection one-form along with some important results on right hand side

1. $\omega(A) = \mathfrak{A}$, where $\mathfrak{A} \in \mathfrak{g}$ and defined to be the generator of induced vector $A \in T_u P$.

2. $(\Psi_g)^* \omega_{u.g}(V) = (Ad_{g^{-1}})_*(\omega_u(V))$.

for any V, where $V \in T_u P$. The above properties are not completely independent but rather 2 can be derived from 1.

Proof:

$$(\Psi_g)^* \omega_{u.g}(V) = \omega_{u.g}((\Psi_g)_*(V))$$

Now consider the curve $\gamma(t) = u.\exp(tA)$ to which V is tangent at point u and so by definition $\omega_u(V) = A$. Now consider the curve for which $(\Psi_g)_*(V)$ is

tangent at point $u.g$ which is

$$\begin{aligned} c(t) &= u.\exp(tA).g \\ &= u.g.g^{-1}.\exp(tA).g \\ &= (u.g).\exp(tg^{-1}.A.g) \end{aligned}$$

and so this curve has generator $g^{-1}.A.g$. Further

$$\omega_{u.g}((\Psi_g)_*(V)) = g^{-1}.A.g = g^{-1}\omega_u(V)g = (Ad_{g^{-1}})_*(\omega_u(V))$$

We just proved that property 2 follows from property 1 and are not independent but it is good to remember both as they come handy in practical calculations. We can define H_uP as

$$H_uP = \{X \in T_uP | \omega(X) = 0\} = Ker(\omega)$$

This section is not exhaustive by any means and some standard text [1] is recommended which should be referred parallel to these notes so that any mathematical lacking here can be covered in a much more complete manner.

4.3 Local connection one-form - Yang-Mills field

One can check that the above defined connection one-form with the mentioned properties provides us with a successful realization of axiomatic connections on the principal bundles. We can further define parallel transport and even covariant derivative (only if associate bundle is a vector bundle) using the above purely coordinate free connection one-form on the principal bundle. However we are interested in how these forms behave on the base manifold and so in this section we develop the formalism for achieving this

task. We assumed before that along with principal bundle we also have a local section σ. So using σ we can obtain one-forms on M as

$$A = \sigma^* \omega$$

which in literature is known as Yang-Mills field. One can ask the question whether we can reconstruct the connection one-form on principal bundle from the Yang-Mills form. The answer is yes! and now we will derive how one can express connection one-form ω on principal bundle in terms of Yang-Mills local forms A. We will first throw the beast at you so that you can feel less scared during the derivation. The expression is

$$\omega(X) = (Ad_{g^{-1}})_*(\pi^* A(X)) + \kappa^* \Xi_g(X) \tag{4.8}$$

where $g = \kappa(u)$ and $\kappa \colon \pi^{-1}(U) \to G$ is defined as follows

$$\Psi(\sigma(\pi(u)), \kappa(u)) = u$$

which is always well defined because of definition of principal bundle. One obvious property is $\kappa(u.g) = \kappa(u).g$ where care has to be taken about the notation - $u.g$ denotes right action of g on u and $\kappa(u).g$ denotes group composition. We will only use the properties of connection one-form given in section 4.2.3 and decomposition of $T_u P$ into $T_p M \bigoplus T_g G$ where $p = \pi(u)$ and g is such that $\sigma(\pi(u)).g = u$ which is just local trivialization of point $u \in P$ in $U \times G$

4.3.1 $T_u P \cong T_p M \bigoplus T_g G$

Let $u \in P$ and have a local trivialization in $U \times G$ to be $(\pi(u), \kappa(u))) = (p, g)$. Now consider a curve $\gamma(t)$ in P originating at u whose coordinates in

$U \times G$ is $(\pi(\gamma(t)), \kappa(\gamma(t))))$. Both U and G being manifolds have coordinate representations

$$f(\pi(\gamma(t)), \kappa(\gamma(t)))) = (x^{\alpha}(\pi(\gamma(t)), g_{\rho}^{\sigma}(\kappa(\gamma(t))))$$

where f is combined coordinate map. In these coordinate we can write the vector X tangent at $u = \gamma(0)$ as

$$X = \frac{\mathrm{d}x^{\alpha}(\pi(\gamma(t))}{\mathrm{d}t}\Big|_{t=0}(\frac{\partial}{\partial x^{\alpha}}\Big|_{\pi(u)}) + \frac{\mathrm{d}g_{\rho}^{\sigma}(\kappa(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0}(\frac{\partial}{\partial g_{\rho}^{\sigma}}\Big|_{\kappa(u)})$$

You can see that the first part of X is a vector in T_pM and the second part of X is a vector in T_gG. But we can do better and obtain this decomposition in a more elegant and coordinate free approach (mathematicians all ears !) by considering the following natural mappings

- $\pi \colon \pi^{-1}(U) \to U$
- $\kappa \colon \pi^{-1}(U) \to G$

which has induced coordinate mappings

$$(x^{\alpha}(\pi(u)), g_{\rho}^{\sigma}(\kappa(u)) \mapsto x^{\alpha}(\pi(u))$$

$$(x^{\alpha}(\pi(u)), g_{\rho}^{\sigma}(\kappa(u)) \mapsto g_{\rho}^{\sigma}(\kappa(u))$$

and now we can write

$$\begin{aligned} X &= \pi_*(X) \bigoplus \kappa_*(X) \\ &= \nu \bigoplus \eta \end{aligned}$$

where $\nu = \pi_*(X) \in T_{\pi(u)}U$ and $\eta = \kappa_*(X) \in T_{\kappa(u)}G$ is introduced for notational ease. As promised earlier we will only use the properties of connection one-form and the above decomposition to get a representation in terms of local Yang-Mills form.

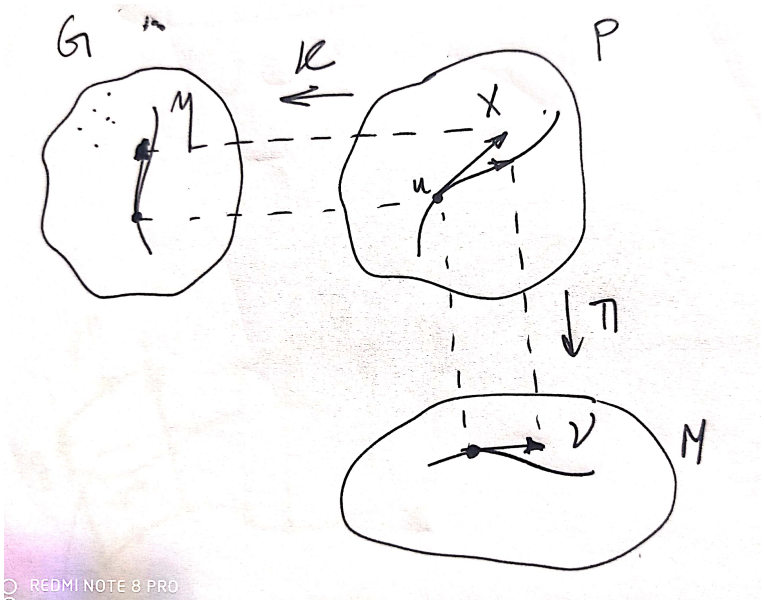

Figure 4.10: Illustration showing an arbitrary vector in the principal bundle and it's decomposition on group manifold and base manifold

4.3.2 Derivation

First we derive a very specialized case where the vector $X \in T_uP$ on which connection one-form ω operates is in vertical subspace V_uP. This case gives an intuition of how the two components (ν and η) of any general X contributes to the value of $\omega(X)$. Consider once again the defining relation of local trivialization which is

$$\Psi(\sigma(\pi(u)), \kappa(u)) = u = \Psi(\sigma \circ x^{-1} \circ x^{\alpha}(\pi(u)), g^{-1} \circ g^{\sigma}_{\rho}(\kappa(u))) \tag{4.9}$$

where we have just inserted coordinate maps x^{α} and g^{σ}_{ρ} of U and G respectively. Now consider a curve $\gamma(t)$ originating from u such that X is tangent to the curve at u. Let us calculate action of X on general smooth function $f \in C^{\infty}(P)$

$$
\begin{aligned}
X[f] &= \frac{\mathrm{d}f(\gamma(t))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\mathrm{d}f(\Psi(\sigma(\pi(\gamma(t))), \kappa(\gamma(t))))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\mathrm{d}f(\Psi(\sigma \circ x^{-1} \circ x^{\alpha}(\pi(\gamma(t))), g^{-1} \circ g^{\sigma}_{\rho}(\kappa(\gamma(t)))))}{\mathrm{d}t}\Big|_{t=0}
\end{aligned}
$$

The above expression has two set of variables $\{x^{\alpha}\}$ and $\{g^{\sigma}_{\rho}\}$, so we can decompose the above expression into two terms which we denote by V_{σ} and V_{κ} where:

$$
\begin{aligned}
V_{\sigma}[f] &= \frac{\partial f(\Psi(\sigma \circ x^{-1} \circ x^{\alpha}(\pi(\gamma(t))), g^{-1} \circ g^{\sigma}_{\rho}(\kappa(\gamma(t)))))}{\partial x^{\alpha}}\Big|_{\kappa(u)} \frac{\mathrm{d}x^{\alpha}(\pi(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\partial f(\Psi_{\kappa(u)}(\sigma \circ x^{-1} \circ x^{\alpha}(\pi(\gamma(t)))))}{\partial x^{\alpha}}\Big|_{\kappa(u)} \frac{\mathrm{d}x^{\alpha}(\pi(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\partial f \circ (\Psi_{\kappa(u)}(\sigma)) \circ x^{-1}(x^{\alpha})}{\partial x^{\alpha}}\Big|_{\kappa(u)} \frac{\mathrm{d}x^{\alpha}(\pi(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\mathrm{d}x^{\alpha}(\pi(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \frac{\partial}{\partial x^{\alpha}}(f \circ (\Psi_{\kappa(u)}(\sigma))) \\
&= \nu[f \circ (\Psi_{\kappa(u)}(\sigma))] = (\Psi_{\kappa(u)}(\sigma))_{*}\nu[f]
\end{aligned}
$$

$$
\begin{aligned}
V_{\kappa}[f] &= \frac{\partial f(\Psi(\sigma \circ x^{-1} \circ x^{\alpha}(\pi(\gamma(t))), g^{-1} \circ g^{\sigma}_{\rho}(\kappa(\gamma(t)))))}{\partial g^{\sigma}_{\rho}}\Big|_{\sigma(\pi(u))} \frac{\mathrm{d}g^{\sigma}_{\rho}(\kappa(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\partial f \circ \Psi_{\sigma(\pi(u))} \circ g^{-1}(g^{\sigma}_{\rho})}{\partial g^{\sigma}_{\rho}}\Big|_{\sigma(\pi(u))} \frac{\mathrm{d}g^{\sigma}_{\rho}(\kappa(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \\
&= \frac{\mathrm{d}g^{\sigma}_{\rho}(\kappa(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} \frac{\partial}{\partial g^{\sigma}_{\rho}}(f \circ \Psi_{\sigma(\pi(u))}) \\
&= \eta[f \circ \Psi_{\sigma(\pi(u))}] \\
&= (\Psi_{\sigma(\pi(u))})_{*}\eta[f]
\end{aligned}
$$

Combining the above two pieces we get

$$X = (\Psi_{\kappa(u)}(\sigma))_* \nu + (\Psi_{\sigma(\pi(u))})_* \eta \tag{4.10}$$

4.3.2.1 Case 1: $X \in V_u P$

If the X is in vertical subspace $V_u P$ then according to the definition of vertical subspace there always exists a natural curve $\gamma(t) = u.\exp(tA)$ such that X is tangent vector this curve at point u. The above curve completely lies in the fibre space of a single base point $p = \pi(\gamma(t))$. The curve $\gamma(t)$ in $U \times G$ is $\chi(\gamma(t)) = (\pi(\gamma(t)), \kappa(\gamma(t))) = (p, \kappa(\gamma(t)))$. Now $\kappa(\gamma(t)) = \kappa(u.\exp(tA)) = \kappa(u).\exp(tA)$ is a curve on group manifold G whose tangent vector at $\kappa(u)$ is η and is generated by $A \in \mathfrak{g}$ which can be given by $A = \Xi_{\kappa(u)}(\eta)$. Also we have ν vector to be 0 as the curve in M is just a constant point p. So the whole contribution of $\omega(X)$ comes from η and can be stated as

$$\begin{aligned}
\omega(X) &= \omega\Big|_{\sigma} \\
&= A \\
&= \Xi_{\kappa(u)}(\eta) \\
&= \Xi_{\kappa(u)}(\kappa_* X) \\
&= \kappa^* \Xi_{\kappa(u)}(X)
\end{aligned}$$

which completes the proof. This case illustrates that η component of X is purely vertical and contributes to the $\omega(X)$ through the Maurer-Cartan form of η.

4.3.2.2 Case 2: $X = X_{Ver} + X_{Hor} \in T_uP$

Now we have all the ingredients to calculate $\omega_u(X)$

$$\begin{aligned}\omega_u(X) &= \omega_u((\Psi_{\kappa(u)}(\sigma))_*\nu + (\Psi_{\sigma(\pi(u))})_*\eta) \\ &= \omega_u((\Psi_{\kappa(u)}(\sigma))_*\nu) + \omega_u((\Psi_{\sigma(\pi(u))})_*\eta)\end{aligned}$$

where the first term can be further simplified using properties of connection one-form

$$\begin{aligned}\omega_u((\Psi_{\kappa(u)}(\sigma))_*\nu) &= \omega_u((\Psi_{\kappa(u)})_*(\sigma)_*\nu) \\ &= (\Psi_{\kappa(u)})^*\omega_u(\sigma_*\nu) \\ &= (Ad_{\kappa(u)^{-1}})_*(\omega_\sigma(\sigma_*\nu)) \\ &= (Ad_{\kappa(u)^{-1}})_*(\sigma^*\omega_\sigma(\nu)) \\ &= (Ad_{\kappa(u)^{-1}})_*(A(\nu)) \\ &= (Ad_{\kappa(u)^{-1}})_*(A(\pi_*X)) \\ &= (Ad_{\kappa(u)^{-1}})_*(\pi^*A(X))\end{aligned}$$

where in third step we have used the property of connection one-form given in section 4.2.4, in fifth step we used the definition of $A = \sigma^*\omega_\sigma$ and in sixth step we used the definition of $\nu = \pi_*X$. Now the second term is simplified as

$$\begin{aligned}\omega_u((\Psi_{\sigma(\pi(u))})_*\eta) &= \Xi_{\kappa(u)}(\eta) \\ &= \Xi_{\kappa(u)}(\kappa_*X) \\ &= \kappa^*\Xi_{\kappa(u)}(X)\end{aligned}$$

combining the two terms we recover the equation

$$\omega(X) = (Ad_{\kappa(u)^{-1}})_*(\pi^* A(X)) + \kappa^* \Xi_{\kappa(u)}(X) \tag{4.11}$$

4.4 Compatibility of local connections

In the start of the above analysis we assumed the availability of a section σ on the local chart $U \subset M$. Now in the cases in which we are interested, we always have a natural choice of section on the principal bundle (frame bundle) which is set of coordinate basis. Now it happens that on the overlap of two charts U_i and U_j we have two different section σ_i and σ_j. We have a unique connection one-form ω in the principal bundle but now we have two different ways in which we can pull-back this connection one-form to the base manifold to get a local Yang-Mills field namely A_i and A_j. Now the compatibility conditions states that for every $u \in P$ the connection one-form ω_u constructed from A_i and A_j should be same and this gives us a condition on A_i and A_j. We will consider a more general situation in which we have been given two sections σ_1 and σ_2 defined on U_1 and U_2 and related as

$$\sigma_2(p) = \Psi(\sigma_1(p), \kappa(p))$$

such that p belongs to the overlapping region on which the two local sections are defined simultaneously. To find the compatibility condition we again consider a curve $\gamma(t) \in U_1 \cap U_2$ originating at point p such that vector $X \in$

$T_p(U_1 \cap U_2)$ is tangent to the curve at point p

$$\begin{aligned}\frac{\mathrm{d}f(\sigma_2(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} &= \frac{\mathrm{d}f(\sigma_2 \circ x^{-1}(x^\alpha(\gamma(t))))}{\mathrm{d}t}\Big|_{t=0} \\ &= \frac{\mathrm{d}x^\alpha(\gamma(t))}{\mathrm{d}t}\Big|_{t=0}\frac{\partial}{\partial x^\alpha}(f \circ \sigma_2) \\ &= X[f \circ \sigma_2] \\ &= (\sigma_2)_* X[f]\end{aligned}$$

However we can calculate the same quantity as

$$\frac{\mathrm{d}f(\sigma_2(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0} = \frac{\mathrm{d}f \circ \Psi(\sigma_1(\gamma(t)), \kappa(\gamma(t)))}{\mathrm{d}t}\Big|_{t=0}$$

If we observe the above expression, it is similar to the one considered in the section 4.3.2 and so we will directly write the expression

$$(\sigma_2)_* X = (\Psi_{\kappa(p)})_*(\sigma_1)_* X + (\Psi_{\sigma_1(p)})_* \eta$$

Now we act on the expression with connection one-form $\omega_{\sigma_2(p)}$

$$\begin{aligned}\omega_{\sigma_2(p)}((\sigma_2)_* X) &= \omega_{\sigma_2(p)}((\Psi_{\kappa(p)})_*(\sigma_1)_* X + (\Psi_{\sigma_1(p)})_* \eta) \\ \sigma_2^* \omega_{\sigma_2(p)}(X) &= (Ad_{\kappa(p)^{-1}})_*(\sigma_1^* \omega_{\sigma_1}(X)) + \kappa^* \Xi_{\kappa(p)}(X)\end{aligned}$$

$$A_2(X) = (Ad_{\kappa(p)^{-1}})_*(A_1(X)) + \kappa^* \Xi_{\kappa(p)}(X) \tag{4.12}$$

We have derived the compatibility condition which local Yang-Mills forms defined with respect to two different sections should satisfy on the overlapping region. We can also say that this is the transformation law of Yang-Mills

field under local gauge transformation by action of $\kappa(p)$ (local means that the κ is a function of the point p on the base manifold). We now look into an important case of sections σ_i and σ_j corresponding to local trivialization χ_i and χ_j defined on charts U_i and U_j. We show that in this case $\kappa(p) = t_{ij}(p)$, where $t_{ij}\colon U_i \cap U_j \to G$ is the transition function on the principal bundle.

$$\begin{aligned}\sigma_j(p) &= \chi_i^{-1}(p, e) = \chi_i^{-1}(p, t_{ij}.e) \\ &= \Psi(\chi_i^{-1}(p, e), t_{ij}(p)) \\ &= \Psi(\sigma_i(p), t_{ij}(p))\end{aligned}$$

so the compatibility conditions becomes

$$A_j(X) = (Ad_{t_{ij}(p)^{-1}})_*(A_i(X)) + t_{ij}^*\Xi_{t_{ij}(p)}(X) \tag{4.13}$$

We can now state the reverse argument which says that given any two local form A_i and A_j in the charts U_i and U_j respectively and following the compatibility condition for $p \in U_i \cap U_j$ can be used to reconstruct a unique connection one-form globally on the principal bundle.

4.5 Example : Frame Bundle - $G \cong GL(m, \mathbb{R})$

Till now we have discussed a general purely geometric formalism of the theory of connections and their local Yang-Mills representations without considering any special structure on the topological group manifold G. Most of the above expressions simplify if we consider some more structure on the group manifold and so in this section we will be considering the structure group to be $GL(m, \mathbb{R})$ which is group of general linear invertible matrices

under composition to be standard matrix multiplication. Here m is the dimension of the vector space on which they operate and we choose to work over the $\mathbb{R}$ field[6].

The coordinates of $a \in GL(m, \mathbb{R})$ is just entries of the matrix which we will denote by $g^\sigma_\rho(a)$ and we have an obvious identity

$$g^\sigma_\rho(a.g) = g^\sigma_\mu(a) g^\mu_\rho(g)$$

for all $a, g \in GL(m, \mathbb{R})$. Let $V \in T_eG \cong \mathfrak{g}$ and we wish to find $(L_a)_*(V)$ for $GL(m, \mathbb{R})$ case, expressing it in terms of coordinates

$$\begin{aligned}
[(L_a)_*(V)]^\sigma_\rho &= V^\mu_\nu \frac{\partial g^\sigma_\lambda(a) g^\lambda_\rho(e)}{\partial g^\mu_\nu(e)} \\
&= V^\mu_\nu g^\sigma_\lambda(a) \frac{\partial g^\lambda_\rho(e)}{\partial g^\mu_\nu(e)} \\
&= V^\mu_\nu g^\sigma_\lambda(a) \delta^\lambda_\mu \delta^\nu_\rho \\
&= V^\lambda_\rho g^\sigma_\lambda(a) \\
(L_a)_*(V) &= a.V
\end{aligned}$$

The maurer-cartan form also takes a very simple form

$$\Xi_a = g^\alpha_\mu(a^{-1}) dg^\mu_\beta(a)$$

Proof: Consider $X \in T_aG$ to be left-translated vector of some vector $V \in$

[6]We can also extend this choice to $\mathbb{C}$ or $\mathbb{H}$.

$T_eG \cong \mathfrak{g}$ and so it should be true that $\Xi_a(X) = V$. Let us check for that

$$\begin{aligned}
[\Xi_a(X)]^\alpha_\beta &= g^\alpha_\mu(a^{-1})\langle dg^\mu_\beta(a), X\rangle \\
&= g^\alpha_\mu(a^{-1})\langle dg^\mu_\beta(a), V^\lambda_\rho g^\sigma_\lambda(a)(\frac{\partial}{\partial g^\sigma_\rho(a)})\rangle \\
&= g^\alpha_\mu(a^{-1})V^\lambda_\rho g^\sigma_\lambda(a)\langle dg^\mu_\beta(a), \frac{\partial}{\partial g^\sigma_\rho(a)}\rangle \\
&= g^\alpha_\mu(a^{-1})V^\lambda_\rho g^\sigma_\lambda(a)\delta^\mu_\sigma\delta^\rho_\beta \\
&= g^\alpha_\sigma(a^{-1})V^\lambda_\beta g^\sigma_\lambda(a) \\
&= V^\alpha_\beta
\end{aligned}$$

where we used $g^\alpha_\sigma(a^{-1})g^\sigma_\lambda(a) = \delta^\alpha_\lambda$ and this completes the proof. Similarly one can show

$$(Ad_{a^{-1}})_* V = a^{-1}Va$$

where $V \in T_eG \cong \mathfrak{g}$.

We can also calculate $\kappa^*\Xi_{\kappa(p)}$ where $\kappa\colon U \to G$ and $p \in U$

$$\begin{aligned}
[\kappa^*\Xi_{\kappa(p)}]^\alpha_\beta &= g^\alpha_\mu(\kappa(p)^{-1})\frac{\partial g^\mu_\beta(\kappa(p))}{\partial x^\nu(p)}dx^\nu(p) \\
&= g^\alpha_\mu(\kappa(p)^{-1})dg^\mu_\beta(\kappa(p)) \\
\kappa^*\Xi_{\kappa(p)} &= \boldsymbol{\kappa(p)^{-1}}d\boldsymbol{\kappa(p)}
\end{aligned}$$

putting these pieces together into the compatibility condition equation (4.12)

$$A_2(X) = \boldsymbol{\kappa(p)^{-1}}A_1(X)\boldsymbol{\kappa(p)} + \boldsymbol{\kappa(p)}^{-1}d\boldsymbol{\kappa(p)}(X) \tag{4.14}$$

which is a much more familiar expression for gauge transformation given in standard texts.

Let us take an example of $G \cong U(1)$ where $\kappa(p) = \exp i\Lambda(p)$ and the above compatibility condition becomes

$$A_2 = A_1 + d\Lambda(p)$$
$$(A_2)_\mu = (A_1)_\mu + \partial_\mu \Lambda(p)$$

which is the familiar gauge transformation relation of electromagnetic potential.

We can do even more in this case by considering natural sections induced by coordinate charts on the base manifold.

4.5.1 Chart induced sections on Frame bundle

Let us take coordinates on chart $U \in M$ to be $\{x^\mu\}$ and then we have a natural section

$$\sigma \colon p \mapsto \{\frac{\partial}{\partial x^1} \dots \frac{\partial}{\partial x^\mu}\Big|_p\}$$

where $\mu = dim(M)$. The frame bundle LM has free right-action

$$\Psi \colon (\{\hat{e}_\alpha\}, a) \mapsto \{\hat{e}_\alpha g_1^\alpha(a) \dots \hat{e}_\alpha g_\mu^\alpha(a)\}$$

where $\{\hat{e}_\alpha\} \in LM$ and $a \in GL(m, \mathbb{R})$.

The change of coordinates in the above language of frame bundle can be regarded as change of natural section from one coordinate chart to the other. Let x^μ be coordinates on chart $U_i \subset M$ and y^ν be coordinates on chart $U_j \subset M$ and so the above right-action can be written as

$$\Psi \colon (\{\frac{\partial}{\partial x^\alpha}\Big|_p\}, \frac{\partial x^\sigma}{\partial y^\rho}(p)) \mapsto \{\frac{\partial x^\alpha}{\partial y^1}\frac{\partial}{\partial x^\alpha}\Big|_p \dots \frac{\partial x^\alpha}{\partial y^\mu}\frac{\partial}{\partial x^\alpha}\Big|_p\} = \{\frac{\partial}{\partial y^1}\Big|_p \dots \frac{\partial}{\partial y^\mu}\Big|_p\}$$

where a used here is just coordinate transformation matrix which is in accordance with the fact that the transformation between two sections which also provides local trivialization is equal to the transition function. Given the above setup we have local connection $\mathfrak{g}$-valued one-form which most of you is familiar with

$$\Gamma^{\alpha}_{\beta\mu} dx^{\mu}$$

where μ is the index attached with the one-form basis whereas α and β are the matrix indices ($\mathfrak{g}$ valued - $dim(M) \times dim(M)$ $matrices \in \mathfrak{gl}(m, \mathbb{R})$ in case of $GL(m, \mathbb{R})$). We boldly say that the connection one-form defined above is our good old friend "christoffel symbols" from the previous chapter. Let us strengthen our claim by calculating it's transformation law (compatibility condition)

$$\begin{aligned}
\Gamma'^{\alpha}_{\beta\sigma} dy^{\sigma} &= (g^{\alpha}_{\nu}(a^{-1})\Gamma^{\nu}_{\lambda\mu} g^{\lambda}_{\beta}(a) + g^{\alpha}_{\nu}(a^{-1})\partial_{\mu} g^{\nu}_{\beta}(a)) dx^{\mu} \\
&= (\frac{\partial y^{\alpha}}{\partial x^{\nu}} \frac{\partial x^{\lambda}}{\partial y^{\beta}} \Gamma^{\nu}_{\lambda\mu} + \frac{\partial y^{\alpha}}{\partial x^{\nu}} \frac{\partial}{\partial x^{\mu}} \frac{\partial x^{\nu}}{\partial y^{\beta}}) \frac{\partial x^{\mu}}{\partial y^{\sigma}} dy^{\sigma} \\
\Gamma'^{\alpha}_{\beta\sigma} &= \frac{\partial y^{\alpha}}{\partial x^{\nu}} (\frac{\partial x^{\lambda}}{\partial y^{\beta}} \frac{\partial x^{\mu}}{\partial y^{\sigma}} \Gamma^{\nu}_{\lambda\mu} + \frac{\partial}{\partial y^{\sigma}} \frac{\partial x^{\nu}}{\partial y^{\beta}})
\end{aligned}$$

which the familiar transformation law of christoffel symbols. In any elementary courses, one thing is stressed upon very frequently and that is "christoffel symbols are not tensors ! they do not follow tensor transformation law". This statement is not at all restrictive to just the case of christoffel symbols and now we know that christoffel symbols are just one example of local connection one-form for a particular choice of section and in general local connection-one forms do not transform like a tensor[7].

[7]Note that at a more fundamental level, transformation laws of tensors are directed by the underlying principal bundle to which tensors just lie in the associated bundle.

4.6 Covariant Derivative on Principal Bundle

An arbitrary Δ valued $k-$form $\alpha \in \Omega^k(M) \bigotimes \Delta$ where M is the manifold on which it is defined and $\Delta(+, .)$ is a general vector space can be written as follows

$$\alpha = \alpha^a_\mu dx^\mu \otimes \delta_a$$

where $\delta_a \in \Delta$ is the basis of Δ. One example of this newly defined arbitrary valud form is our old friend connection one-form ω which is $\mathfrak{g}$ valued and is defined on principal bundle.

Now we can define arbitrary valued covariant derivative with respect to connection one-form ω as follows

$$D_\omega \alpha(X_1, \dots X_{k+1}) \coloneqq d\alpha(X_1^H, \dots X_{k+1}^H) \tag{4.15}$$

where $X_i \in T_pM$ and X_i^H is the horizontal component of X_i.

4.6.1 Curvature 2-form

We can immediately find covariant derivative of connection one-form ω which is called as curvature two-form Ω and will be useful in defining familiar definition of curvature tensor on the base manifold

$$\Omega(X, Y) \coloneqq D\omega(X, Y) = d\omega(X^H, Y^H)$$

Expressing $X^H = X - X^V$ and $Y^H = Y - Y^H$ and putting it in the above expression

$$\begin{aligned} D\omega(X, Y) &= d\omega(X - X^V, Y - Y^H) \\ &= d\omega(X, Y) - d\omega(X^V, Y) - d\omega(X, Y^V) + d\omega(X^V, Y^V) \end{aligned}$$

Let $\omega(X) = \omega(X^V) = A$ and $\omega(Y) = \omega(Y^V) = B$.

We will need the following propositions

- $d\xi(X, Y) = X[\xi(Y)] - Y[\xi(X)] + \xi([X, Y])$ holds true for any 2-form ξ and all X and Y.
- if $V \in V_uP$ and $W \in H_uP$ then $[V, W] \in H_uP$ for any V and W.
- For any $V, W \in V_uP$, $\omega([V, W]) = [\omega(V), \omega(W)]$ holds true.

Consider the term

$$\begin{aligned} d\omega(X^V, Y) &= X^V[\omega(Y)] - Y[\omega(X^V)] - \omega([X^V, Y]) \\ &= X^V[B] - Y[A] - \omega([X^V, Y^V + Y^H]) \\ &= X^V[B] - Y[A] - \omega([X^V, Y^V]) \\ &= X^V[B] - Y[A] - [\omega(X^V), \omega(Y^V)] \\ &= X^V[B] - Y[A] - [A, B] \end{aligned}$$

and the next term

$$\begin{aligned} d\omega(X, Y^V) &= X[\omega(Y^V)] - Y^V[\omega(X)] - \omega([X, Y^V]) \\ &= X[B] - Y^V[A] - [A, B] \end{aligned}$$

and the last term

$$\begin{aligned} d\omega(X^V, Y^V) &= X^V[\omega(Y^V)] - Y^V[\omega(X^V)] - \omega([X^V, Y^V]) \\ &= X^V[B] - Y^V[A] - [A, B] \end{aligned}$$

Putting this all together we get

$$\begin{aligned} d\omega(X,Y) - d\omega(X^V,Y) - d\omega(X,Y^V) + d\omega(X^V,Y^V) &= d\omega(X,Y) - X^V[B] + Y[A] + [A,B] \\ &\quad - X[B] + Y^V[A] + [A,B] \\ &\quad + X^V[B] - Y^V[A] - [A,B] \\ &= d\omega(X,Y) + Y[A] - X[B] + [A,B] \\ &= d\omega(X,Y) + [\omega(X),\omega(Y)] \end{aligned}$$

and finally we have proved an extremely important equation known as Cartan struture equation

$$D\omega(X,Y) = d\omega(X,Y) + [\omega(X),\omega(Y)] = \Omega(X,Y) \tag{4.16}$$

or we can write it as

$$D\omega = d\omega + \omega \wedge \omega$$

This quantity is useful as we can prove that the above curvature two-form is generator of holonomy group and so we have exactly same information capture by holonomy group as that of by curvature tensor. This provides a new more algebraic viewpoint to the curvature properties of the bundle. Some brief introduction to this will be given in the next section. Careful readers will we skeptical about the notation used in the above equation as wedge product of two same forms is zero, but that was the case when the forms were real valued however in the equation above we have $\mathfrak{g}-$ valued form which do not commute and hence is not zero on anti-symmetrization. Let us make this point clear by considering a $\mathfrak{g}-$ valued $r-$ form α and another $\mathfrak{g}-$ valued $p-$ form β which can be written as

$$\xi = \xi^\alpha \otimes T_\alpha$$

$$\eta = \eta^{\beta} \otimes T_{\beta}$$

where T_{α} is basis of lie-algebra $\mathfrak{g}$. We now define lie-bracket of $\mathfrak{g}-$ value forms

$$[\xi, \eta] \coloneqq \xi \wedge \eta - (-1)^{rp} \eta \wedge \xi$$

If forms were real valued then the above expression will be zero, but it is not the case here as we will see

$$\begin{aligned}
\xi \wedge \eta(X_1, \ldots X_{r+p}) &= \sum_{P \in S_{r+p}} \frac{1}{r!p!} (-1)^{sgn(P)} \xi(X_{P(1)}, \ldots X_{P(r)}) \eta(X_{P(r+1)}, \ldots X_{P(r+p)})) \\
&= \sum_{P \in S_{r+p}} \frac{1}{r!p!} (-1)^{sgn(P)} \xi^{\alpha}(X_{P(1)}, \ldots X_{P(r)}) T_{\alpha} \eta^{\beta}(X_{P(r+1)}, \ldots X_{P(r+p)})) T_{\beta} \\
&= \sum_{P \in S_{r+p}} \frac{1}{r!p!} (-1)^{sgn(P)} \xi^{\alpha}(X_{P(1)}, \ldots X_{P(r)}) \eta^{\beta}(X_{P(r+1)}, \ldots X_{P(r+p)})) T_{\alpha} T_{\beta} \\
&= (\xi^{\alpha} \wedge \eta^{\beta}) \otimes T_{\alpha} T_{\beta}(X_1, \ldots X_{r+p})
\end{aligned}$$

Using the above expression in 4.6.1 we get

$$\begin{aligned}
\xi \wedge \eta - (-1)^{rp} \eta \wedge \xi &= (\xi^{\alpha} \wedge \eta^{\beta}) \otimes T_{\alpha} T_{\beta} - (-1)^{rp} (\eta^{\beta} \wedge \xi^{\alpha}) \otimes T_{\beta} T_{\alpha} \\
&= (\xi^{\alpha} \wedge \eta^{\beta}) \otimes (T_{\alpha} T_{\beta} - T_{\beta} T_{\alpha}) \\
&= (\xi^{\alpha} \wedge \eta^{\beta}) \otimes [T_{\alpha}, T_{\beta}] \\
&= (\xi^{\alpha} \wedge \eta^{\beta}) \otimes f_{\alpha\beta}^{\gamma} T_{\gamma}
\end{aligned}$$

where $f_{\alpha\beta}^{\gamma}$ are structure constant of the lie-algebra of the structure group. We have proved that

$$[\xi, \eta] = (\xi^{\alpha} \wedge \eta^{\beta}) \otimes f_{\alpha\beta}^{\gamma} T_{\gamma}$$

4.6.2 Physical meaning of Curvature 2-form

Let us take an example where principal bundle is the frame bundle (where each point in the bundle is a set of basis for the tangent space on the base manifold) where we have an structure group which is subgroup of $GL(m, \mathbb{R})$. The right action of structure group on frame bundle is free and so fibre space is isomorphic to the structure group itself. Now vertical vectors lie in the fibre space and is responsible for changing the basis at the same point on the base manifold and horizontal vectors are responsible for connecting two different basis (frames) at different point on the base manifold. A general curve in the frame bundle is a combination of these two effects. Let us consider $p \in M$ and vectors $\partial_\alpha, \partial_\beta \in T_pM$ such that there horizontal lift on principal bundle at point u is $V, W \in H_uP$. We also know the following proposition

- For any two vectors $V, W \in H_uP$ such that $\pi_*(V) = X \in T_pM$ and $\pi_*(W) = Y \in T_pM$ then $\pi_*([V, W]^H) = [X, Y]$ which is the statement that the horizontal component of $[V, W]$ is the horizontal lift of the vector $[X, Y]$.

Using this we have

$$\begin{aligned}\pi_*([V, W]^H) &= [\pi_*(V), \pi_*(W)] \\ &= [\partial_\alpha, \partial_\beta] \\ &= 0\end{aligned}$$

which means that vector $[V, W]$ is vertical and we have $[V, W]^V = [V, W]$. We also know that lie-bracket captures the failure of flow trajectories being closed along two different vectors and if it is zero then the two flows form a closed loop with two flows leading to the exactly same point. Consider

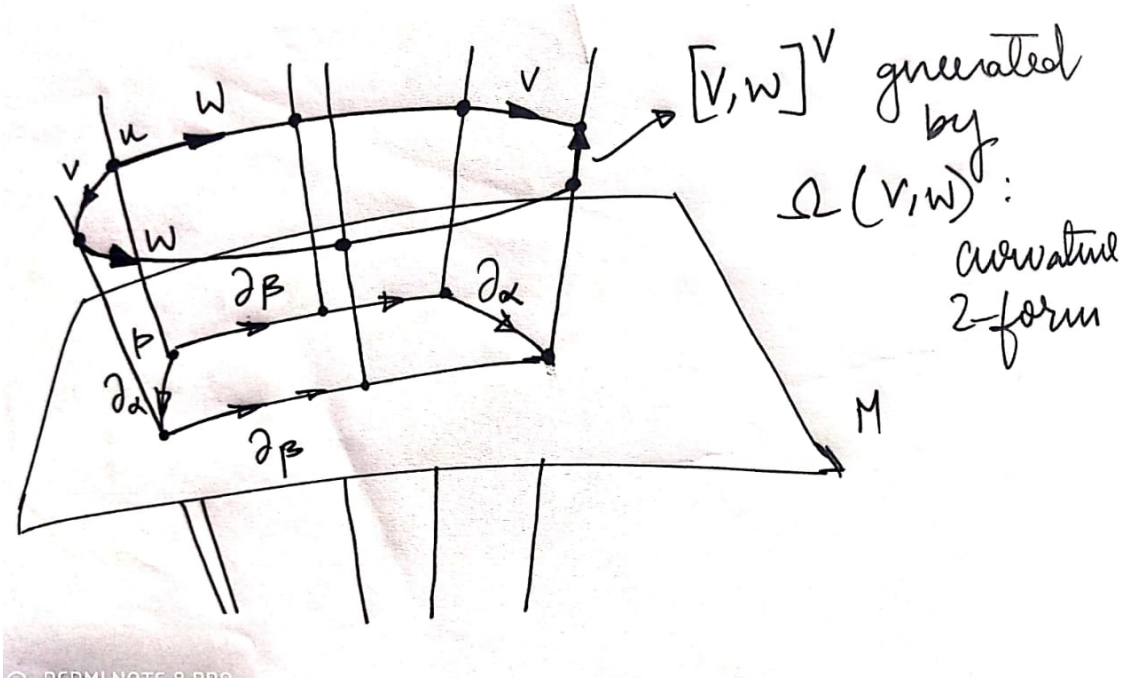

Figure 4.11: Illustration describing the failure of closure in horizontal lifted path in principal bundle for a closed path on base manifold

the expression for curvature two-form 4.16 and evaluate this with horizontal vectors $V, W \in H_u P$

$$\begin{aligned} D\omega(V,W) &= d\omega(V,W) + [\omega(V), \omega(W)] \\ &= d\omega(V,W) \\ &= V[\omega(W)] - W[\omega(V)] - \omega([V,W]) \\ &= -\omega([V,W]) \\ \Omega(V,W) &= -\omega([V,W]^V) \end{aligned}$$

The above quantity $\omega([V,W]^V) \in \mathfrak{g}$ is the generator of induced vector $[V,W]$ and so the curvature 2-form $\Omega(V,W)$ is the generator of the transformation between the frames horizontally reached by two different curves on the base manifold with tangent vectors ∂_α and ∂_β. The set of transformations corresponding to all possible closed loops at a point has a group structure (with

composition defined by the composition of loops) called as Holonomy group. The closing statement is that curvature 2-form physically captures the information of how a frame will change along a horizontal lifted closed curve.

4.6.3 Covariant derivative of $\sigma-$ horizontal type forms

Note that we can calculate covariant derivative for any arbitrary valued k-form without any special properties whatsoever but for our purposes we will now consider a special type of $F-$valued $k-$form known as $\sigma-$ horizontal type form which has the following properties

- It is annihilated by any vertical vector.
- $(\Psi_g)^* \tilde{\alpha} = \sigma_{g^{-1}} \circ \tilde{\alpha}$

where F is a vector space manifold which has a finite-dimensional representation of group action σ_g. The importance of these kinds of forms will be clear in subsequent sections when we will discuss covariant derivative on associated bundle. For $\sigma-$ horizontal type $F-$valued $k-$form we can express covariant derivative in a much more elegant way

$$D\tilde{\alpha} = d\tilde{\alpha} + \sigma'(\omega) \wedge \tilde{\alpha} \tag{4.17}$$

Proof: Before proving the above expression, we first quote an extremely important identity:

For any $k-$ form ξ the following identity holds true

$$\begin{aligned} d\xi(X_1, \dots X_{k+1}) = &\sum_{i=1}^{k} (-1)^{i+1} X_i \xi(X_1, \dots, \hat{X}_i, \dots X_{r+1}) \\ &+ \sum_{i<j} (-1)^{i+j} \xi([X_i, X_j], X_1, \dots, \hat{X}_i, \dots, \hat{X}_j, \dots X_{r+1}) \end{aligned} \tag{4.18}$$

where hat represents that the argument is not included. Now by definition

$$D\tilde{\alpha}(X_1 \dots X_{k+1}) = d\tilde{\alpha}(X_1^H \dots X_{k+1}^H)$$

Writing $X_i^H = X_i - X_i^V$ and putting it in the above definition and using 4.18

$$\begin{aligned} D\tilde{\alpha}(X_1 \dots X_{k+1}) &= d\tilde{\alpha}(X_1 - X_1^V, \dots X_{k+1} - X_{k+1}^V) \\ &= d\tilde{\alpha}(X_1, \dots X_{k+1}) - \sum_{i=1}^{k} (-1)^i X_i^V [\tilde{\alpha}(X_1, \dots, \hat{X}_i, \dots X_{r+1})] \end{aligned}$$

All other terms in the expression vanishes because of the property of $\sigma-$ horizontal type form which annihilates the vertical vectors.

To prove further we have to workout another important identity

$$\begin{aligned} X_0[\tilde{\alpha}(X_1, \dots X_k)] &= \frac{\mathrm{d}\tilde{\alpha}_{\Psi_{\exp(tA)}(u)}((\Psi_{\exp(tA)})_* X_1, \dots (\Psi_{\exp(tA)})_* X_k)}{\mathrm{d}t}\Big|_{t=0} \\ &= \frac{\mathrm{d}(\Psi_{\exp(tA)})^* \tilde{\alpha}_{\Psi_{\exp(tA)}(u)}(X_1, \dots X_k)}{\mathrm{d}t}\Big|_{t=0} \\ &= \frac{\mathrm{d}\sigma_{\exp(-tA)} \circ \tilde{\alpha}_u(X_1, \dots X_k)}{\mathrm{d}t}\Big|_{t=0} \\ &= -\sigma'(A)(\tilde{\alpha}(X_1, \dots X_k)) \\ &= -\sigma'(\omega(X_0))(\tilde{\alpha}(X_1, \dots X_k)) \end{aligned}$$

and finally we have a very important identity

$$X_0[\tilde{\alpha}(X_1, \dots X_k)] = -\sigma'(\omega(X_0))(\tilde{\alpha}(X_1, \dots X_k))$$

Here $X_0 \in V_uP$ and $\gamma(t) = \Psi_{\exp(tA)}(u)$ is the curve to which X_0 is tangent at the point $u \in P$. We put this in the above expression we have,

$$-\sum_{i=1}^{k}(-1)^i X_i^V[\tilde{\alpha}(X_1,\ldots,\hat{X}_i,\ldots X_{r+1})] = \sum_{i=1}^{k}(-1)^i\sigma'(\omega(X_i))(\tilde{\alpha}(X_1,\ldots,\hat{X}_i,\ldots X_k))$$

defining the above expression as

$$(\sigma'(\omega)\wedge\tilde{\alpha})(X_1,\ldots X_{k+1}) := \sum_{i=1}^{k}(-1)^i\sigma'(\omega(X_i))(\tilde{\alpha}(X_1,\ldots,\hat{X}_i,\ldots X_k))$$

Putting all the pieces together we recover the expression 4.17. For a moment consider 0-form case of the identity 4.6.3, we have

$$X_0[\tilde{\alpha}] = -\sigma'(\omega(X_0))(\tilde{\alpha}) \tag{4.19}$$

Few things to observe about this identity is as follows

- Right hand side is a vector $X_0 \in V_uP$ being operated on a $F-$ valued function $\tilde{\alpha}$ and so the result is not a real number (which is, in standard case) but rather an element in F.
- Left hand side is a vector in $T_fF \cong F$ generated by $A = \omega(X_0)$ at point $f = \tilde{\alpha}(u) \in F$.

We can also prove that covariant derivative of a $\sigma-$ horizontal type $k-$ form is again a $\sigma-$ horizontal type $(k+1)-$ form.

Proof:

The first property is immediate from the definition of covariant derivative as even if at least one argument is vertical vector then it's horizontal component will be zero and thus from 4.15 it will be zero and so it annihilates any verti-

cal vector in the argument. Second property can be proved by considering

$$\begin{aligned}
(\Psi_g)^* D\tilde{\alpha}(X_1, \ldots X_{k+1}) &= D\tilde{\alpha}((\Psi_g)_* X_1, \ldots (\Psi_g)_* X_{k+1}) \\
&= d\tilde{\alpha}([(\Psi_g)_* X_1]^H, \ldots [(\Psi_g)_* X_{k+1}]^H) \\
&= d\tilde{\alpha}((\Psi_g)_* X_1^H, \ldots (\Psi_g)_* X_{k+1}^H) \\
&= (\Psi_g)^* d\tilde{\alpha}(X_1^H, \ldots X_{k+1}^H) \\
&= d(\Psi_g)^* \tilde{\alpha}(X_1^H, \ldots X_{k+1}^H) \\
&= d\sigma_{g^{-1}} \circ \tilde{\alpha}(X_1^H, \ldots X_{k+1}^H) \\
&= \sigma_{g^{-1}} \circ d\tilde{\alpha}(X_1^H, \ldots X_{k+1}^H) \\
&= \sigma_{g^{-1}} \circ D\tilde{\alpha}(X_1, \ldots X_{k+1})
\end{aligned}$$

and finally we can write $(\Psi_g)^* D\tilde{\alpha} = \sigma_{g^{-1}} \circ D\tilde{\alpha}$ and hence we can say that the covariant derivative is a $\sigma-$ horizontal type $(k+1)-$ form.

4.7 Assosiated Bundles

We can defining right-action Θ_g on $P \times F$ as:

$$\Theta_g(u, f) \coloneqq (\Psi(u, g), \sigma_{g^{-1}}(f))$$

where $u \in P$ and $f \in F$. Note that here σ_g can be any finite-dimensional representation of the structure group which can suitably act on the vector space F and so we can have natural outer product representations as well which will act as above on outer product vector space. Once we have defined a right-action on the $P \times F$, we can naturally form a fibre bundle $E = P \times_G F$ known as associated bundle. Any point on the associated bundle will be an equivalence class $[(u, f)] = \{(\Psi(u, g), \sigma_{g^{-1}}(f)) | \forall g \in G\}$. We have a natural

projection mapping $\iota\colon P \times F \to E$

$$\iota\colon (u, f) \mapsto [(u, f)]$$

which in turn gives us to mappings

$$\iota_u\colon f \mapsto [(u, f)]$$

from which we can immediately derive

$$\begin{aligned}\iota_{\Psi(u,g)}(f) &= [(\Psi(u, g), f)] \\ &= [(u, \sigma_g(f))] \\ &= \iota_u(\sigma_g(f)) \\ &= \iota_u \circ \sigma_g(f)\end{aligned}$$

and finally we get an important property $\iota_{\Psi(u,g)} = \iota_u \circ \sigma_g$.

$$\iota_f\colon u \mapsto [(u, f)]$$

We can also define projection map of this fibre bundle by $\pi_E([(u, f)]) \coloneqq \pi(u)$ which is independent of the representation of the equivalence class as $\pi(u) = \pi(\Psi(u, g))$. Now to represent a point on E we need to fix a particular value of (u, f) which will serve as a representative for the equivalence class $[(u, f)]$ corresponding to that point. The fixing of (u, f) is called gauge-fixing. We saw previously that defining a section provided us with local trivialization on the principal bundle, we can follow a similar procedure and can get induced local trivialization on the associated bundle as follows. First we consider a section $s\colon M \to P$ on principal bundle and then we can define a local

trivialization as

$$\begin{aligned}[(u, f)] &= [(\Psi(s(p), g), f)] \\ &= [(s(p), \sigma_g(f))] \to (p, \sigma_g(f))\end{aligned}$$

where $p = \pi(u) = \pi_E([(u, f)])$. Since the linear finite dimensional group action on F is free and so we get a local trivialization $\xi\colon E \to M \times F$

$$\xi\colon [(u, f)] \mapsto (\pi_E([(u, f)]), \sigma_g(f))$$

and so we can say that the fibre space over the associated bundle is F. We can also see an interesting fact that the transition function on the associated bundle is same as that on the principal bundle

$$\begin{aligned}[(s_j(p), f_j)] &= [(\Psi(s_j(p), t_{ij}(p)), f_j)] \\ &= [(s_i(p), \sigma_{t_{ij}(p)}(f_j))] \\ &= [(s_i(p), f_i)]\end{aligned}$$

from which we can observe that the section s_i induced coordinate f_i is related by finite-dimensional group action $\sigma_{t_{ij}(p)}$ to section s_j induced coordinate f_j which is the same transition function t_{ij} but just expressed in finite-dimensional representation.

4.7.1 Isomorphism between $\Lambda^k(M, E)$ and $\Omega^k_{\sigma,Hor}(P, F)$

Now we move on to define some important relations between $F-$ valued k-forms on principal bundle and $E-$ valued k-forms on base manifold which will be of great use while defining covariant derivative on associated bundle.

- Let $\Lambda^k(M, E)$ be the space of $E-$ valued $k-$ form on base manifold M and let some arbitrary $\alpha \in \Lambda^k(M, E)$.

- Let $\Omega^k_{\sigma,Hor}(P,F)$ be the space of $\sigma-$ horizontal type $F-$ valued $k-$ form on principal bundle P and let some arbitrary $\tilde{\alpha} \in \Omega^k_{\sigma,Hor}(P,F)$.

Let $u \in P$ and $p \in M$ such that $\pi(u) = p$. Also consider arbitrary k vectors $\{Y_i \in T_uP\}_{i=1\ldots k}$ and $\{X_i \in T_pM\}_{i=1\ldots k}$such that $\pi_*(Y_i) = X_i$. We define

$$\alpha_p(X_1, \ldots X_k) \coloneqq \iota_u \circ \tilde{\alpha_u}(Y_1, \ldots Y_k)$$

We need to prove that the above definition is independent of the choice of u and Y_i. So we take another point $u' \in P$ such that $\pi(u') = \pi(u) = p$ and we can write $u' = \Psi(u,g)$. Also a generalized vector $Y'_i \in T_{u'}P$ can be expressed as $Y'_i = (\Psi_g)_*(Y_i) + Z_i$ such that $Z_i \in V_{u'}P \; \forall i = 1 \ldots k$ are vertical vectors and follows $\pi_*(Y'_i) = X_i$.
Proof:

$$\begin{aligned}
\iota_{\Psi(u,g)} \circ \tilde{\alpha}_{\Psi(u,g)}(Y'_1, \ldots Y'_k) &= \iota_{\Psi(u,g)} \circ \tilde{\alpha}_{\Psi(u,g)}((\Psi_g)_*(Y_1) + Z_1, \ldots (\Psi_g)_*(Y_k) + Z_k) \\
&= \iota_{\Psi(u,g)} \circ \tilde{\alpha}_{\Psi(u,g)}((\Psi_g)_*(Y_1), \ldots (\Psi_g)_*(Y_k)) \\
&= \iota_{\Psi(u,g)} \circ (\Psi_g)^* \tilde{\alpha}_{\Psi(u,g)}(Y_1, \ldots Y_k) \\
&= \iota_u \circ \sigma_g \circ (\Psi_g)^* \tilde{\alpha_u}(Y_1, \ldots Y_k) \\
&= \iota_u \circ \sigma_g \circ (\Psi_g)^* \tilde{\alpha_u}(Y_1, \ldots Y_k) \\
&= \iota_u \circ \sigma_g \circ \sigma_{g^{-1}} \circ \tilde{\alpha_u}(Y_1, \ldots Y_k) \\
&= \iota_u \circ \tilde{\alpha_u}(Y_1 \ldots Y_k)
\end{aligned}$$

where in second step we used the property of $\sigma-$ horizontal type form that vertical vectors Z_i are annihilated, in fourth step we used the property of natural projection map proved in previous section, in sixth we again used the

property of $\sigma-$ horizontal type form. We can also obtain reverse mapping

$$\begin{aligned}
\alpha_p(X_1, \dots X_k) &= \iota_u \circ \tilde{\alpha_u}(Y_1, \dots Y_k) \\
\alpha_p(\pi_*(Y_1), \dots \pi_*(Y_k)) &= \iota_u \circ \tilde{\alpha_u}(Y_1, \dots Y_k) \\
\pi^* \alpha_p(Y_1 \dots Y_k) &= \iota_u \circ \tilde{\alpha_u}(Y_1, \dots Y_k) \\
\iota_u^{-1} \circ \pi^* \alpha_p(Y_1 \dots Y_k) &= \tilde{\alpha_u}(Y_1 \dots Y_k)
\end{aligned}$$

and hence we obtain the reverse mapping

$$\iota_u^{-1} \circ \pi^* \alpha_p = \tilde{\alpha_u}$$

and finally we have established the isomorphism between $\Lambda^k(M, E)$ and $\Omega^k_{\sigma,Hor}(P, F)$

- $\alpha_p(X_1, \dots X_k) = \iota_u \circ \tilde{\alpha_u}(Y_1, \dots Y_k)$
- $\iota_u^{-1} \circ \pi^* \alpha_p = \tilde{\alpha_u}$

If we consider the case of $k = 0$ then $\tilde{\alpha} \in \Omega^0_{\sigma,Hor}(P, F)$ is a smooth $F-$ valued function on P and $\alpha \in \Lambda^0(M, E)$ is a section on E where the isomorphism is defined as

- $\iota^{-1} \circ \alpha(\pi(u)) \coloneqq \tilde{\alpha}(u)$ - $F-$ valued smooth function on P.
- $\alpha(p) \coloneqq [(u, \tilde{\alpha}(u))]$ - A section on E.

where $u \in P$ is any point in $\pi^{-1}(p)$ and the properties of $\sigma-$ horizontal type form ensures that the above mapping is independent of the chosen u.

4.7.2 Induced connection on Associated Bundle

Let $u \in P$ and $e = \iota_f(u) = [(u, f)] \in E$ and where $f \in F$ is fixed. We will now workout mapping between vertical vector in principal bundle and

thus induced vertical vertical vector in associated bundle. Consider a curve in principal bundle $\gamma(t) = \Psi(u, \exp(tA))$ with vector X_u^V tangent to this curve at point u which is generated by $A \in \mathfrak{g}$ and thus completely lies in fibre space at the point $p = \pi(u) \in M$. Using the natural mapping between principal bundle and associated bundle $\iota_f(u) = [(u, f)] \in E$ we have a curve in E

$$\tilde{\gamma}(t) = [(\Psi(u, \exp(tA)), f)]$$

Now consider an arbitrary smooth function $h\colon E \to \mathbb{R}$, then we have

$$\begin{aligned}
X_u^V[h \circ \iota_f] &= \left.\frac{\mathrm{d}h([\Psi(u, \exp(tA)), f])}{\mathrm{d}t}\right|_{t=0} \\
(\iota_f)_* X_u^V[h] &= \left.\frac{\mathrm{d}h([\Psi(u, \exp(tA)), f])}{\mathrm{d}t}\right|_{t=0} \\
&= \left.\frac{\mathrm{d}h([u, \sigma_{\exp(tA)}(f)])}{\mathrm{d}t}\right|_{t=0} \\
&= \left.\frac{\mathrm{d}h \circ \iota_u(\sigma_{\exp(tA)}(f))}{\mathrm{d}t}\right|_{t=0} \\
&= Y_f[h \circ \iota_u] \\
(\iota_f)_* X_u^V[h] &= (\iota_u)_* Y_f[h]
\end{aligned}$$

and we know that ι_u is a bijection and so we have

$$\sigma'(\omega(X_u)) \coloneqq (\iota_u^{-1} \circ \iota_f)_* X_u^V = Y_f \in T_f F \cong F$$

where $Y_f \in T_f F \cong F$ is tangent to the curve $\eta(t) = \sigma_{\exp(tA)}(f)$ in the fibre manifold F. We observe one important fact that the above definition provides us with a mapping $\sigma'\colon \mathfrak{g} \to T_f F \cong F$. Now we can define induced vertical subspace on associated bundle $V_e E$ as

$$Z_e^V \coloneqq (\iota_f)_* X_u^V = (\iota_u)_* Y_f$$

we can see that such a definition follows similar property

$$\begin{aligned}(\pi_E)_*(Z_e^V) &= (\pi_E \circ \iota_f)_*(X_u^V) \\ &= (\pi)_*(X_u^V) \\ &= 0\end{aligned}$$

where we have used $\pi_E \circ \iota_f(u) = \pi(u)$. In similar manner we can define the horizontal subspace H_eE

$$Z_e^H \coloneqq (\iota_f)_* X_u^H$$

4.7.3 Connection one-form on Associated Bundle

We know that ι_u where u is fixed is a bijection and so ι_{u*} provides us with vector space isomorophism between $T_fF \cong F$ (F is a vector space - vector bundle) and V_eE. That means for any $Z \in V_eE$ there exist $\mu = (\iota_u^{-1})_*(Z) \in T_fF \cong F$ which further maps to E by $\iota_u(\mu) \in E$. So we have a one-form $\omega_E \colon TE \to E$ defined as

$$\omega_E \coloneqq Z \mapsto \iota_u \circ (\iota_u^{-1})_*(Z) \in E$$

We can immediately obtain a direct relation between ω_E and ω by using 4.7.2 where we consider $Z = (\iota_f)_*(X_u^V)$, we have

$$\begin{aligned}\omega_E((\iota_f)_*(X_u^V)) &= \iota_u \circ (\iota_u^{-1})_*((\iota_f)_*(X_u^V)) \\ &= \iota_u \circ (\iota_u^{-1} \circ \iota_f)_*(X_u^V) \\ &= \iota_u \circ \sigma'(\omega(X))\end{aligned}$$

4.7.4 Covariant derivative on Associated Bundle - Koszul Calculus

We are finally in position to define covariant derivative on associated bundle by exploiting the above isomorphism between $\Lambda^k(M,E)$ and $\Omega^k_{\sigma,Hor}(P,F)$. Let $u \in P$, $\pi(u) = p \in M$ and $X_i \in T_pM$, $Y_i \in T_uP$ satisfying $\pi_*(Y_i) = X_i$, we have

$$(d_\omega \alpha)_p(X_1, \dots X_{k+1}) \coloneqq \iota_u \circ (D_\omega \tilde{\alpha})_u(Y_1, \dots Y_{k+1}) \tag{4.20}$$

First we will consider a simple and important case of $0-$ forms. In the familiar case of differential geometry where we are interested in finding covariant derivative of vector field (or tensor fields) on base manifold, which when translated in the language of bundles is equivalent to finding covariant derivative of a section α on associated bundle (tangent bundle if we are considering vector field). But we know that for a section on associated bundle there corresponds a $F-$ valued smooth function $\tilde{\alpha}$ on P and by definition 4.20 we have,

$$\begin{aligned}(d_\omega \alpha)_p(X) &= \iota_u \circ (D_\omega \tilde{\alpha})_u(Y) \\ &= \iota_u \circ d\tilde{\alpha}(Y^H) \\ &= \iota_u(Y^H[\tilde{\alpha}])\end{aligned}$$

where $\pi_*(Y) = \pi_*(Y^H) = X$ and so we can say that $Y^H \in H_uP$ is horizontal lift of $X \in T_pM$. Finally we define the familiar ∇ as

$$\nabla\alpha(X) = \nabla_X \alpha = (d_\omega \alpha)_p(X) = \iota_u(Y^H[\tilde{\alpha}])$$

We can further obtain a simpler expression of covariant derivative in terms of local connection one-form which will be important in practical calculations.

First we consider basis $\{\mathbf{e}_\alpha\}$ for the vector space manifold F. Let there be a mapping $\tilde{e}_\alpha \colon P \to F$ defined by

$$\tilde{e}_\alpha \colon s(p) \mapsto \mathbf{e}_\alpha$$

where s is a local section on P. Using the induced local trivialization on associated bundle we have an induced basis on E

$$e_\alpha(p) \coloneqq [(s(p), \tilde{e}_\alpha(s(p)))] = [(s(p), \mathbf{e}_\alpha)]$$

Now as e_α is a section on E, we can find it's covariant derivative by considering $X \in T_pM$ and $Y \in T_uP$ which satisfies $\pi_*(Y) = X$. As we know that it does not matter what Y and u we chose as far as it satisfies $\pi_*(Y) = X$ and so we can chose $Y = s_*(X)$

$$\begin{aligned}
\nabla e_\alpha(X) &= \iota_{s(p)} \circ D\tilde{e}_\alpha(s_*(X)) \\
&= \iota_{s(p)} \circ (d\tilde{e}_\alpha(s_*(X)) + \sigma'(\omega(s_*(X)))\tilde{e}_\alpha(s(p))) \\
&= \iota_{s(p)} \circ (s_*(X)[\tilde{e}_\alpha] + \sigma'(s^*\omega(X))\mathbf{e}_\alpha) \\
&= \iota_{s(p)} \circ (X[\tilde{e}_\alpha \circ s] + \sigma'(A(X))\mathbf{e}_\alpha) \\
&= \iota_{s(p)} \circ A(X)^\beta_\alpha \mathbf{e}_\alpha \\
&= A(X)^\beta_\alpha \iota_{s(p)}(\mathbf{e}_\alpha) \\
&= A(X)^\beta_\alpha e_\beta(p) \\
&= A^\beta_{\mu\alpha} X^\mu e_\beta(p)
\end{aligned}$$

where $A = s^*(\omega) = A^\beta_{\mu\alpha} dx^\mu$ (where α, β are $\mathfrak{g}-$ valued indices and μ is form index) is local connection one-form on M. Note that $X[\tilde{e}_\alpha \circ s] = 0$ because $\tilde{e}_\alpha \circ s(p) = \mathbf{e}_\alpha$ is a constant basis on F. In the fifth step we have used the fact that the finite linear group representation is in matrix form and hence the

lie-algebra will also be in matrix form and so can directly act on elements of F. We know that local connection one-form corresponding to natural coordinate basis in frame bundle are chistoffel symbols $\Gamma^{\beta}_{\mu\alpha}$ which are very common choice in differential geometry.

4.7.5 Curvature 2-form on Associated Bundle

We move on to defining curvature 2-form on associated bundle as

$$R^{\nabla}_{p}(X, Y) \coloneqq -\iota_u \circ \sigma'(\Omega_u(X^H, Y^H)) \circ \iota_u^{-1}$$

where $u \in \pi^{-1}(p)$ and $X^H, Y^H \in H_uP$ which satisfies $\pi_*(X^H) = X, \pi_*(Y^H) = Y \in T_pM$. Here we have curvature 2-form which is $\mathfrak{g}-$ valued but then use the mapping $\sigma' : \mathfrak{g} \to F$ to get an effective 2-form on principal bundle which is $F-$ valued. But we know from our discussion in the section 4.7.1 that for any $F-$ valued $k-$ form on principal bundle P has a corresponding $E-$ valued $k-$ form on base manifold M given by the mapping $\iota_u : F \to E$. Using the mapping 4.19 in this case, we have

$$\begin{aligned}
\sigma'(\Omega(X^H, Y^H))(\tilde{\alpha}(u)) &= -Ver([X^H, Y^H])_u[\tilde{\alpha}] \\
&= -([X^H, Y^H])[\tilde{\alpha}] + Hor([X^H, Y^H])[\tilde{\alpha}] \\
&= -X^H(Y^H[\tilde{\alpha}]) + Y^H(X^H[\tilde{\alpha}]) + ([X, Y]^H)[\tilde{\alpha}] \\
&= -\iota_u^{-1} \circ (\nabla_X \nabla_Y \alpha - \nabla_Y \nabla_X \alpha - \nabla_{[X,Y]}\alpha)
\end{aligned}$$

and finally we have

$$\begin{aligned}
-\iota_u \circ \sigma'(\Omega(X^H, Y^H)) \circ \iota_u^{-1}(\alpha(p)) &= (\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]})\alpha(p) \\
R^{\nabla}_{p}(X, Y)\alpha(p) &= (\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]})\alpha(p)
\end{aligned}$$

This is the standard expression which all of you have seen. When it is first introduced without the language of principal bundle it seems pretty ad-hoc but with the theory of connections we are able to derive this expression from first principles and this shows the power of theory of connections. We have developed all the necessary mathematical prerequisites required to understand research papers in this field.

Information Geometry in Deep Learning

4.1 Introduction

In this chapter we will exploit the formalism developed in the previous chapters on non-euclidean geometry and theory of connections in the learning theory of statistical models starting with generalization of gradient descent. A generalization of gradient descent on non-euclidean information manifolds called as natural gradient takes into account metric tensor while computing lengths of steps in which an smooth loss function will decrease maximally. In this chapter for introducing natural gradient we follow the classic original paper [15] and for discussion on one of the tractable applications of natural gradient in deep learning namely natural neural networks we again refer and follow the lines of the original paper [16]. This provides an excellent discussion on very first example of projective learning algorithm in non-canonical parametrization setting for neural networks in deep learning.

4.2 Natural Gradient Descent

In this section we will sketch out a derivation for natural gradient and we will show that natural gradient is a much more direct and "natural" definition for what we would call a gradient when one incorporate non-euclidean geometric framework which in this context is divergence (or probability distribution) induced metric on information manifolds. Consider the loss function $l(\theta)$ where $\theta \in \Theta$ is parameter in configuration space. Consider we change this parameter infinitesimally along a vector $d\theta = \epsilon a$ under the constraint that $||a|| = 1 = \sum_{ij} g_{ij} a^i a^j = a^\intercal G a$ and correspondingly we have change in loss function $l(\vec{\theta})$ as

$$\begin{aligned} dl(\theta) = l(\theta + d\theta) - l(\theta) &= l(\theta) + \epsilon \nabla l(\theta)^\intercal \cdot a - l(\theta) \\ &= \epsilon \nabla l(\theta)^\intercal \cdot a \end{aligned}$$

Now we want that a (direction) in which we get the extremum for $dl(\theta)$ under the constraint $||a|| = 1$ and we have the following Lagrange for the above optimization problem

$$L(a) = \nabla l(\theta)^\intercal \cdot a - \lambda(a^\intercal G a - 1)$$

and we have

$$\begin{aligned} \frac{\partial L(a)}{\partial a^i} &= \frac{\partial}{\partial a^i}(\nabla l(\theta)^\intercal \cdot a - \lambda(a^\intercal G a - 1)) \\ &= \nabla l(\theta) - 2\lambda G a \\ &= 0 \end{aligned}$$

and we get

$$a = \frac{1}{2\lambda} G^{-1} \nabla l(\theta)$$

where λ can be found from the constraint. We define the natural gradient as

$$\tilde{\nabla} l(\theta) = G^{-1} \nabla l(\theta)$$

which has a new factor of G^{-1} in it unlike standard gradient used in learning theory. If you observe carefully then you can see that even the standard gradient is just a special case where the metric is flat-euclidean and $G = G^{-1} = \mathbb{I}$. We now have a new parameter update rule

$$\theta_{t+1} = \theta_t - \eta_t \tilde{\nabla} l(\theta)$$

where η_t is the learning rate at the step t. Though the computations involved for calculating fisher information matrix is very expensive and for neural networks in deep learning in standard canonical parametrization setting it becomes nearly intractable to even compute such kind of metrics. But we can restrict the form of fisher information matrix to as simple as an identity matrix by reparametrization (change of basis) and changing the architecture of NN. One such interesting case is provided in the paper [16].

PyGlow: Python package for Information Theory of Deep Learning

In this chapter we provide a brief introduction to a python package **PyGlow: Information Theory of Deep Learning** authored by **Bhavya Bhatt**. This package provides intensive support for nearly all the information theoretic learning algorithms in deep learning. The package provides a clean easy-to-code API using which programmer can use his/her own IB measures and attach multiple dynamics evaluator with the architecture itself. We now give a brief overview of the basic module structure and their functionalities.

6.1 Data Structures and Data handling

PyGlow has a dedicated data handling and specific methods for data preparation.

6.1.1 Core layer module - glow.layers

Layer provides all major layer data structure which can be added to your custom deep neural network.

```
from glow.layers import Dense, Dropout, Conv2d, Flatten
```

6.1.2 Datasets - glow.datasets

This module provide extensive support for all major standard dataset which are used for bench-marking the performances of models.

```
from glow.datasets import mnist

# hyperparameter
batch_size = 64
num_workers = 3
validation_split = 0.2
num_epochs = 2

train_loader, val_loader, test_loader = mnist.load_data(
    batch_size=batch_size, num_workers=num_workers,
    validation_split=validation_split
)
```

6.2 PyGlow Core Package

6.2.1 Core model module - glow.models

```
from glow.models import Network, Sequential, IBSequential, HSICSequential
```

6.2.1.1 Network

This is the base class for all the neural network modules. User should also subclass this module. The module has an exceptional similarity with the famous deep learning library **Keras**.

6.2.1.2 Sequential

This sub-module is vanilla functional model class with similar methods as that of Sequential class of **Keras**.

```
model = Sequential(input_shape=(1, 28, 28), gpu=True)
model.add(Conv2d(filters=16, kernel_size=3, stride=1, padding=1,
activation='relu'))
model.add(Flatten())
```

6.2.1.3 IBSequential

This is modification on Sequential which provides additional Information Bottleneck plugins and attachers to enable training dynamics tracking. The module also provides information plane trajectory evaluators.

```
model = IBSequential(input_shape=(1, 28, 28), gpu=True, track_dynamics=True,
save_dynamics=True)
model.add(Conv2d(filters=16, kernel_size=3, stride=1, padding=1,
```

```
activation='relu'))
model.add(Flatten())
```

6.2.1.4 HSICSequential

The module is base class for all HSIC type neural network models. The module is flexible for different loss function for different layers setting and ensure maximum information preservation during training. For more information on the semantics and related algorithms refer to the paper [17].

```
model = HSICSequential(input_shape=(1, 28, 28), gpu=True)
model.add(Conv2d(filters=16, kernel_size=3, stride=1, padding=1,
activation='relu'))
model.add(Flatten())
```

6.2.2 Information Bottleneck API - glow.information_bottleneck

6.2.2.1 Estimator

This module provides different algorithms for estimation of mutual information including recent ones like **EDGE** and **MINE** estimators. These methods are relevant in IB based models for both information plane trajectory analysis and training using IB loss functions. One of the unique feature of PyGlow is this specialized module for mutual information estimator which is new in itself because there does not exist much open implementations on the subject.

```
from glow.information_bottleneck import Estimator
```

6.3 Sample Codes

6.3.1 Defining your own custom criterion for dynamics evaluation

Defining your own custom criterion can be very useful as the same instance you can either use it for analysing training dynamics or you can use it as IB-based loss objective for HSIC networks.

Following is the general structure which your custom criterion class should follow in order for the PyGlow to detect your criterion function. NOTE: In most of the cases you won't require implementation for *eval_dynamics_segment* so only criterion function is sufficient for PyGlow to accept your instance. You need *eval_dynamics_segment* implementation when you totally have to change the way you want to process a single dynamics segment.

```
# importing Estimator class from glow
from glow.information_bottleneck import Estimator

class CustomCriterion(Estimator):
    def __init__(self, function, gpu, **kwargs):
        super().__init__(gpu, **kwargs)
        self.func = function  # attributes other than parameters

    def criterion(self, x, y):
        # Define your logic for calculating criterion here
        # TODO

    def eval_dynamics_segment(self, dynamics_segment):
        segment_size = len(dynamics_segment)
        # NOTE: dynamics_segment is a list where
```

```
        # input = dynamics_segment[0]
        # label = dynamics_segment[segment_size - 1]
        # and remaining tensors are hidden layers output
        output_segment = []  # output list which contains
        calculated coordinates
        for idx in range(1, segment_size-1):
            h = dynamics_segment[idx]
            # Define your logic of using h, input and label here
            # to apply your custom criterion on them.
            output_segment.append([self.criterion(h, x),
            self.criterion(h, y)])

        return output_segment
```

6.3.2 Training MNIST data classifier in PyGlow

This code is to make you realize how much similarity PyGlow shares with Keras API structure. The purpose of this type of API is to make anyone just entering this field to quickly realize the power of these tools rather than get overwhelmed by the complexity of the theory itself.

```
from glow.layers import Dense, Dropout, Conv2d, Flatten
from glow.models import Sequential
from glow.datasets import mnist
import numpy as np

# hyperparameter
batch_size = 64
num_workers = 3
validation_split = 0.2
num_epochs = 2
```

```
# load the dataset
train_loader, val_loader, test_loader = mnist.load_data(
    batch_size=batch_size, num_workers=num_workers,
    validation_split=validation_split
)

model = Sequential(input_shape=(1, 28, 28), gpu=True)
model.add(Conv2d(filters=16, kernel_size=3, stride=1, padding=1,
activation='relu'))
model.add(Flatten())
model.add(Dropout(0.4))
model.add(Dense(500, activation='relu'))
model.add(Dropout(0.4))
model.add(Dense(200, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(10, activation='softmax'))

model.compile(optimizer='SGD', loss='cross_entropy', metrics=['accuracy'])
# training the model
model.fit_generator(train_loader, val_loader, num_epochs, show_plot=False)
```

6.3.3 Analysing training dynamics using HSIC Criterion

This code explains how we can analyse training process by using calculating HSIC criterion as measure of dependence between hidden representation and output label (generalization) and between hidden representation and input variable (complexity) and obtaining 2-D training trajectory in information plane (we will continue calling it information plane even though HSIC criterion has no counter part in information theory).

```
# importing PyGlow modules
import glow
from glow.layers import Dense, Dropout, Conv2d, Flatten
from glow.datasets import mnist, cifar10
from glow.models import IBSequential
from glow.information_bottleneck.estimator import HSIC

# hyperparameter
batch_size = 64
num_workers = 3
validation_split = 0.2
num_epochs = 7

# load the dataset
train_loader, val_loader, test_loader = mnist.load_data(
    batch_size=batch_size, num_workers=num_workers,
    validation_split=validation_split
)
model = IBSequential(input_shape=(1, 28, 28), gpu=True, track_dynamics=True,
save_dynamics=True)
model.add(Conv2d(filters=16, kernel_size=3, stride=1, padding=1,
activation='relu'))
model.add(Flatten())
model.add(Dense(1000, activation='relu'))
model.add(Dense(500, activation='relu'))
model.add(Dense(200, activation='relu'))
model.add(Dense(10, activation='softmax'))

# compile the model
model.compile(optimizer='SGD', loss='cross_entropy', metrics=['accuracy'])
model.attach_evaluator(HSIC(kernel='gaussian', gpu=True, sigma=5))
```

6.3.4 HSIC Network - Models that train without back-prop

In this notebook we explain how you can use HSIC Bottleneck paradigm for training a feed-forward neural network. For more information refer to the paper [17]

```
import glow
from glow.layers import Dense, Dropout, Conv2d, Flatten, HSICoutput
from glow.datasets import mnist, cifar10
from glow.models import IBSequential, Sequential, HSICSequential, Network
from glow.information_bottleneck.estimator import HSIC
from glow.information_bottleneck import Estimator
import torch

# hyperparameter
batch_size = 64
num_workers = 3
validation_split = 0.2
num_epochs = 3

# load the dataset
train_loader, val_loader, test_loader = mnist.load_data(
    batch_size=batch_size, num_workers=num_workers,
    validation_split=validation_split
)
model = HSICSequential(input_shape=(1, 28, 28), gpu=True)
model.add(Conv2d(filters=16, kernel_size=3, stride=1, padding=1,
activation='relu'))
model.add(Flatten())
model.add(Dense(500, activation='relu'), HSIC(kernel='gaussian', gpu=True,
sigma=5), regularize_coeff=100)
model.add(Dense(200, activation='relu'))
```

```
model.compile(loss_criterion=HSIC(kernel='gaussian', gpu=True, sigma=10),
optimizer='SGD', regularize_coeff=100)
model.pre_training_loop(num_epochs, train_loader, val_loader)
```

Bibliography

[1] Nakahara, M. *Geometry, topology and physics*, 2003.

[2] Gerd Rudolph, Matthias Schmidt. *Differential Geometry and Mathematical Physics*. Springer Netherlands, 2013.

[3] N. Tishby and N. Zaslavsky, "Deep Learning and the Information Bottleneck Principle," abs/1503.02406, 2015.

[4] R. Shwartz, N. Tishby, "Opening the Black Box of Deep Neural Networks via Information," arXiv/1703.00810, Aug 2018.

[5] S. Arora, R. Ge, B. Neyshabur, Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," arXiv/1802.05296, 2018.

[6] Z. Allen-Zhu, Y. Li, Z. Song, "A Convergence Theory for Deep Learning via Over-Parameterization," arXiv/1811.03962, 2018.

[7] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey and D. D. Cox, "On the Information Bottleneck Theory of Deep Learning," *International Conference on Learning Representations,* 2018.

[8] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, R. Wang, "On Exact Computation with an Infinitely Wide Neural Net," arXiv/1904.11955, 2019.

[9] R. Kuditipudi, X. Wang, H. Lee, Y. Zhang, Z. Li, W. Hu, S. Arora, R. Ge, "Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets," arXiv/1906.06247, 2019.

[10] M. Noshad and A. O. Hero III, "Scalable Mutual Information Estimation using Dependence Graphs," abs/1801.09125, Aug 2018.

[11] I. Belghazi, S. Rajeswar, A. Baratin, R. Devon Hjelm and A. C. Courville, "MINE: Mutual Information Neural Estimation," abs/1801.04062, 2018.

[12] R. A. Amjad and B. C. Geiger, "How (Not) To Train Your Neural Network Using the Information Bottleneck Principle," abs/1802.09766, 2018.

[13] A. A. Alemi, I. Fischer, J. V. Dillon and K. Murphy, "Deep Variational Information Bottleneck," abs/1612.00410, 2016.

[14] F. Nielsen, "An elementary introduction to information geometry," arXiv/1808.08271, 2018.

[15] Amari, Shun-ichi, "Natural Gradient Works Efficiently in Learning," 10.1162/089976698300017746, 2000.

[16] G. Desjardins, K. Simonyan, R. Pascanu and K. Kavukcuoglu, "Natural Neural Networks," arXiv/1507.00210, 2015.

[17] Wan-Duo K. Ma, J. P. Lewis, W. B. Kleijn, “The HSIC Bottleneck: Deep Learning without Back-Propagation,” arXiv/1908.01580, 2019.