THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Mathematical Foundations of Equivariant Neural Networks

JIMMY ARONSSON

Division of Algebra and Geometry
Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2023

Mathematical Foundations of Equivariant Neural Networks
Jimmy Aronsson
Gothenburg 2023
ISBN 978-91-7905-854-8

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5320
ISSN 0346-718X

Division of Algebra and Geometry
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Telephone +46 (0)31 772 1000

Author e-mail: `jimmyar@chalmers.se`

Typeset with LaTeX
Printed by Chalmers Digitaltryck
Gothenburg, Sweden 2023

Mathematical Foundations of Equivariant Neural Networks

Jimmy Aronsson

Division of Algebra and Geometry
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Deep learning has revolutionized industry and academic research. Over the past decade, neural networks have been used to solve a multitude of previously unsolved problems and to significantly improve the state of the art on other tasks. However, training a neural network typically requires large amounts of data and computational resources. This is not only costly, it also prevents deep learning from being used for applications in which data is scarce. It is therefore important to simplify the learning task by incorporating inductive biases - prior knowledge and assumptions - into the neural network design.

Geometric deep learning aims to reduce the amount of information that neural networks have to learn, by taking advantage of geometric properties in data. In particular, *equivariant neural networks* use symmetries to reduce the complexity of a learning task. Symmetries are properties that do not change under certain transformations. For example, rotation-equivariant neural networks trained to identify tumors in medical images are not sensitive to the orientation of a tumor within an image. Another example is graph neural networks, i.e., permutation-equivariant neural networks that operate on graphs, such as molecules or social networks. Permuting the ordering of vertices and edges either transforms the output of a graph neural network in a predictable way (*equivariance*), or has no effect on the output (*invariance*).

In this thesis we study a fiber bundle theoretic framework for equivariant neural networks. Fiber bundles are often used in mathematics and theoretical physics to model nontrivial geometries, and offer a geometric approach to symmetry. This framework connects to many different areas of mathematics, including Fourier analysis, representation theory, and gauge theory, thus providing a large set of tools for analyzing equivariant neural networks.

Keywords: geometric deep learning, equivariance, induced representations, convolutional neural networks, fiber bundles, gauge theory, symmetry.

List of publications

This doctoral thesis builds upon the licentiate thesis of the same author,

> **Aronsson, J.** (2021). G-equivariant Convolutional Neural Networks. *Licentiate thesis, Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg.*

and is based on the work contained in the following papers:

I. **Aronsson, J.** (2022). Homogeneous vector bundles and G-equivariant convolutional neural networks. *Sampling Theory, Signal Processing, and Data Analysis, 20*(2), 1-35.

II. Gerken, J.E., **Aronsson, J.**, Carlsson, O., Linander, H., Ohlsson, F., Petersson, C., and Persson, D. (2021). Geometric Deep Learning and Equivariant Neural Networks. *arXiv preprint arXiv:2105.13926*.

III. **Aronsson, J.**, Müller, D, and Schuh, D. (2022). Geometrical aspects of lattice gauge equivariant convolutional neural networks. *arXiv preprint arXiv:2303.11448*

Author contributions

I. My own work.

II. Responsible for the section on group equivariant layers for homogeneous spaces. Participated in the theoretical study of gauge equivariant neural networks. Wrote part of the introduction and edited the manuscript.

III. Identified the connection between L-CNNs and the fiber bundle theoretic framework for equivariant neural networks, and wrote the corresponding section. Contributed to the generalization of L-CNNs to rotoreflections. Wrote part of the introduction and conclusion, and edited the manuscript.

Acknowledgements

This thesis marks the end of my journey as a PhD student. When I first started in 2018, the research group consisted of myself, my advisor Daniel Persson, and my co-advisors Christoffer Petersson and Robert Berman. The project was something of an experiment, an opportunity for us all to explore what the machine learning hype is all about and to perhaps make a few contributions of our own. Our research group has grown steadily over the years and now includes Fredrik Ohlsson, Jan Gerken, Oscar Carlsson, Hampus Linander, Daniel Schuh, Emma Andersdotter Svensson, and Heiner Spieß. I have also enjoyed a close collaboration with David Müller, through innumerable discussions and, ultimately, a joint paper. This thesis is the result of a collaborative effort and I am deeply grateful to you all.

Thanks to all of my colleagues and friends at the Department of Mathematical Sciences here in Göteborg: Felix Held, Linnea Hietala, Erik Jansson, Anton Johansson, Jimmy Johansson, Carl-Joar Karlsson, Gustav Lindwall, Per Ljung, Malin Mosquera, Gabrijela Obradović, Edvin Wedin, Olof Zetterqvist, Linnea Österberg, and many others. Thanks to my examiner Håkan Samuelsson for your support, and to Marie Kühn, Aila Särkkä, and Elisabeth Eriksson for patiently answering my many questions.

I am fortunate to have been part of the *Wallenberg AI, Autonomous Systems and Software Program (WASP)*, through which I have made many friends: Karl Bengtsson Bernander, Georg Bökman, Lucas Brynte, Johan Edstedt, Emilio Jorge, Rita Laezza, Jakob Lindqvist, Yara Lochman, Pavlo Melnyk, and Niklas Åkerblom, to name a few.

A large number of friends have helped me throughout the years, sharing my joy on sunny days and providing shelter on stormy ones. My heartfelt thanks to Iris, Angelica, Elin Thomas, Annika, Rebecka, John, Bogdan, Zakarias, Johan, Bella, Pilo, Daniel, Kevin, Joni, Mårten, Henrik, Simon, Herman, Isabelle, and my dog Agapi.

Finally, to my family: I cannot thank you enough for giving me the freedom to choose my own path and for supporting me every step of the way.

Contents

1 Introduction

This thesis consists of three chapters of background material followed by three papers, **Papers I-III**. The background material should be read as needed, it is not necessary to read all three chapters before starting with the papers.

We begin with an introduction to deep learning that focuses on a supervised learning problem and a particular class of neural networks known as multilayer perceptrons. This introduction sets the stage for a more detailed discussion on *convolutional neural networks (CNNs)* and their *translation equivariance* property that has inspired much research under the term *geometric deep learning*.

Equivariant neural networks, including the *group equivariant convolutional neural networks* and the *gauge equivariant neural networks* studied in **Papers I-III**, make use of symmetries as an inductive bias that simplifies deep learning tasks. That is, such networks are designed to deliver high performance while using fewer parameters and less data than their non-equivariant counterparts.

This thesis centers around a mathematical framework for equivariant neural networks that uses principal bundles and associated vector bundles. Not only does this framework let us incorporate global symmetries (*group-equivariance*) and local symmetries (*gauge-equivariance*) into neural networks, the bundle-theoretic language allows it to capture relevant geometric information in a more general sense. The drawback is that it requires an advanced set of mathematical tools. For this reason, we also provide introductions to representation theory and to mathematical gauge theory.

Notational remark: In this thesis and in **Papers I-II**, we follow the mathematics convention of letting the word *vector* refer to *an element of a vector space*, or linear space. We stress this point because the data in an equivariant neural network is modeled as sections of associated vector bundles. This definition does include, for example, scalar fields and tensor fields despite the name. Since **Paper III** is aimed primarily at a physics audience, it follows the physics convention and reserves the word *vector* to mean *geometric vector*; an object with magntiude and direction and that transforms in a certain way.

2 Deep Learning

In this chapter, we give mathematical introductions to essential aspects of deep learning, convolutional neural networks, and geometric deep learning.

2.1 Essentials of deep learning

Consider the problem of estimating an unknown function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ between two spaces $\mathcal{X}$ and $\mathcal{Y}$, given a *training data set*

$$S_{\text{train}} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, N_{\text{train}}\}, \tag{2.1}$$

with *labels* $y_i = \mathcal{F}(x_i) + \epsilon_i$ that may contain random noise ϵ_i.

In a house price estimation task, for instance, $\mathcal{X}$ could be the set of all houses in some geographical region, parameterized by numerical features such as square meters, number of rooms, construction year, etc. In this example, the codomain $\mathcal{Y} = [0, M]$ would be the possible price range for some realistic upper bound $M > 0$, and the unknown function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ would associate each house with its, in some sense, true value. The training data set S_{train} would contain a list of recently sold houses x_i together with their final selling price y_i, which may differ from their true value $\mathcal{F}(x_i)$ for one reason or another.

Another example is an image classification task in which $\mathcal{X}$ consists of medical images and $\mathcal{F} : \mathcal{X} \to \{0, 1\}$ is a binary function that specifies whether a given image $x \in \mathcal{X}$ depicts a tumor. In this case, the training set S_{train} would be a relatively small set of manually labeled images and any noise $\epsilon_i \in \{0, 1\}$ would be due to misclassification.

A natural approach for approximating the unknown function $\mathcal{F}$ is to consider a space of parameterized functions $\mathcal{F}_\theta : \mathcal{X} \to \mathcal{Y}$, and use the training data set to optimize the parameter θ so to minimize the discrepancy between $\mathcal{F}_\theta$ and $\mathcal{F}$. This is what neural networks do.

Among the simplest and most basic neural networks are *multilayer perceptrons*. These are functions $\mathcal{F}_\theta : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ consisting of a sequence of *layers*

$$x^l = \sigma^l \left(W^l x^{l-1} + b^l \right), \qquad l = 1, \dots, L, \tag{2.2}$$

where $x^0 \in \mathbb{R}^{N_0}$ is the input to the first layer. The components of the *bias vector* $b^l \in \mathbb{R}^{N_l}$ and of the *weight matrix* $W^l \in \mathbb{R}^{N_l \times N_{l-1}}$ in each layer are *trainable parameters*, and the *activation functions* $\sigma^l : \mathbb{R} \to \mathbb{R}$ are non-linear functions that are applied component-wise to $W^l x^{l-1} + b^l$. Common choices for the activation functions include the rectified linear unit $\mathrm{ReLU}(x) = \max\{0, x\}$ as well as the sigmoid $\sigma(x) = (1 + e^{-x})^{-1}$. The neural network is thus the function

$$\mathcal{F}_\theta : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}, \qquad \mathcal{F}_\theta(x^0) = x^L, \tag{2.3}$$

and θ consists of all trainable parameters, i.e., components of the bias vectors b^l and weight matrices W^l. This means that the number of trainable parameters can grow extremely large if the dimensions are large ($N_\ell \gg 0$) and the network is sufficiently *deep* ($L \gg 0$). It is not uncommon for a neural network to have billions of trainable parameters.

We estimate the difference between the neural network $\mathcal{F}_\theta$ and the unknown function $\mathcal{F} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ by comparing the output $\mathcal{F}_\theta(x_i)$ to the corresponding label y_i for each training data point x_i. The comparison is made using a non-negative distance function

$$d : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to [0, \infty). \tag{2.4}$$

Common choices include the Euclidean norm $d(v, w) = \|v - w\|$ or its square $d(v, w) = \|v - w\|^2$. The function $\mathcal{F}$ is estimated by minimizing the *loss function*

$$\ell(\theta) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} d(f_\theta(x_i), y_i). \tag{2.5}$$

Minimizing the distance between $\mathcal{F}_\theta(x_i)$ and $\mathcal{F}(x_i)$ would have provided a better approximation, but we only know the labels $y_i = \mathcal{F}(x_i) + \epsilon_i$. This makes it important to have accurate labels with little noise. If $d(v, w) = \|v - w\|$, for example, we obtain the upper bound

$$\frac{1}{N_{\text{train}}} \sum_i \|\mathcal{F}_\theta(x_i) - \mathcal{F}(x_i)\| \leq \ell(\theta) + \frac{1}{N_{\text{train}}} \sum_i \|\epsilon_i\|, \tag{2.6}$$

and the same bound (with different notation) holds for any distance function $d(v, w)$ that satisfies the triangle inequality. If the labels contain a lot of noise, then we cannot guarantee a good estimate of $\mathcal{F}$ regardless of how much we are able to minimize the loss function $\ell(\theta)$.

The trainable parameters constituting θ are often initialized randomly, meaning that the network produces nonsense estimates before training. The network has no predictive power in the initial stages. However, an iterative optimization method such as gradient descent,

$$\theta \mapsto \theta - \alpha \nabla \ell(\theta), \tag{2.7}$$

with *learning rate* $\alpha > 0$, ensures that the training loss $\ell(\theta)$ always decreases in each iteration of training (called an *epoch* to distinguish it from other iterative processes). There are many different optimization methods to choose between, but most methods are based on gradient descent. One such example is *stochastic* gradient descent, which effectively allows the loss function to increase in some epochs; this helps prevent the loss function from getting stuck in a local minima.

Given a sufficiently large number of trainable parameters, and after training for a sufficient number of epochs, a well-designed neural network eventually reaches a small training loss. What this means is that the network accurately predicts the correct label y_i for inputs x_i in the training data set. However, the reason why we develop and train neural network is to apply them outside of the training context. In the house price estimation task, for example, the network is trained on historic data with the hopes of using it to estimate values of houses that are not yet on the market. So it is crucial to investigate whether the network has learned to solve the actual problem at hand; whether it extracts relevant information from data and makes an educated guess, or if it has simply learned the training data by heart. A network that performs well on training data but fails to make accurate predictions outside of the training data set is said to suffer from *overfitting*.

One example would be if there is only a single training data point, $N_{\text{train}} = 1$. The network has little chance to learn general features of the data distribution if it only sees a single instance, hence it is likely to overfit. So one way to prevent overfitting is to increase the amount of training data. Another method is to add regularization terms to (2.5) that prevent the training loss from becoming too fine-tuned to the (noisy) training data.

The performance is also evaluated during each epoch by applying the network to a separate *test data set* S_{test} and computing the *test loss*. This information is not used to update the network parameters through gradient descent, it is only used for evaluation. However, performance on test data is often used as a stopping criteria: If the test loss has begun to increase while the training loss is still decreasing, then the network is likely starting to overfit. Training may thus be aborted at this point.

Model parameter such as the number of layers L, the dimensions N_l, learning rates, regularization parameters, etc., are known as *hyperparameters*. It is common to train a neural network multiple times with different combinations of hyperparameters and then choose the best-performing combination. As the test data has then indirectly influenced the training of all these network designs, causing statistical bias, a third *validation data set* $S_{\text{validation}}$ is used to evaluate the different combinations of hyperparameters. Hopefully, the resulting neural network $\mathcal{F}_\theta$ is a good approximation to the unknown function $\mathcal{F}$.

In this introduction to deep learning, we have discussed one of the most basic neural network designs - the multilayer perceptron. We have also focused on the task of approximating an unknown function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ given a training data set (2.1) of labeled data (x_i, y_i). This is known as a *supervised* learning task. There exists an ocean of neural network designs and different kinds of tasks, with different training procedures and evaluation methods [Goodfellow et al., 2016]. Nevertheless, the material we have presented here summarizes some of the core aspects that many deep learning methods have in common.

2.2 Convolutional neural networks

In this thesis, we are primarily interested in a particular type of neural network called *convolutional neural networks (CNNs)*. These networks are applied to *data points* that have a 2D or 3D grid structure, which we represent mathematically as finitely supported functions

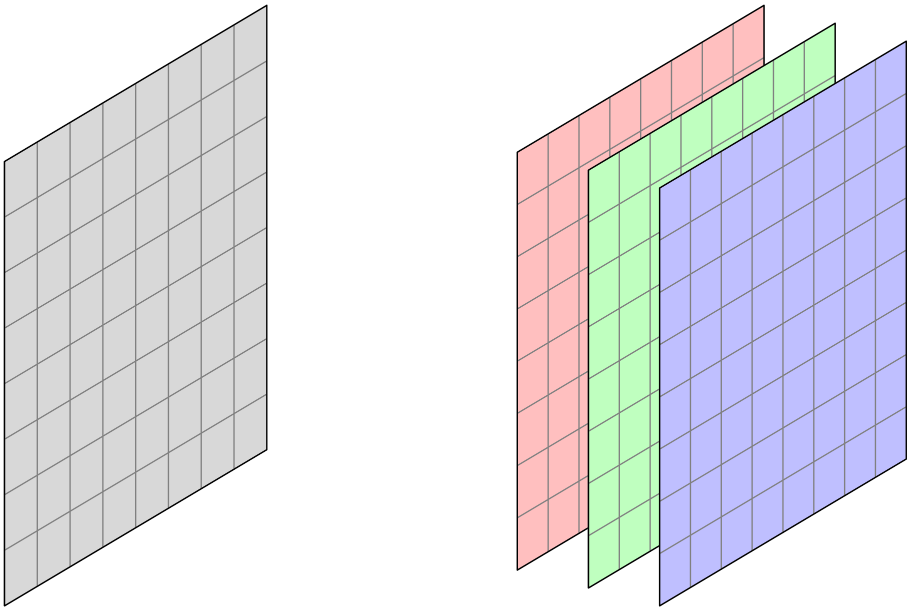
$$f : \mathbb{Z}^2 \to \mathbb{R}^m, \tag{2.8}$$

where m is the number of *channels*. For example, a 3D array of size $10 \times 20 \times 5$ is modeled as a function $f : \mathbb{Z}^2 \to \mathbb{R}^5$ supported on a 10×20 grid in $\mathbb{Z}^2$ and with $m = 5$ channels.

Digital images satisfy (2.8) if we view $\mathbb{Z}^2$ as the pixel grid: Digital images map each pixel to either a grayscale value ($m = 1$) or an RGB array ($m = 3$). Finite support is then analogous to finite image resolution. Equivalently, we can view RGB images as being 3D arrays in which the red, green, and blue color channels are stacked on top of each other (Figure 2.1).

The main building blocks of CNNs are called *convolutional layers*. The name is inspired by convolutions of real-valued functions $\kappa, f : \mathbb{R} \to \mathbb{R}$,

$$(\kappa * f)(y) = \int_{-\infty}^{\infty} \kappa(y - x) f(x) \, \mathrm{d}x, \tag{2.9}$$



(a) Grayscale image ($m = 1$).

(b) RGB image ($m = 3$).

Figure 2.1: Digital images have a 2D (grayscale) or 3D (RGB) grid structure. We can also view them as functions $f : \mathbb{Z}^2 \to \mathbb{R}^m$ with $m = 1$ (grayscale) or $m = 3$ (RGB) channels.

but have been generalized to allow vector-valued data points (2.8):

$$[\kappa \star f](y) = \sum_{x \in \mathbb{Z}^2} \kappa(x - y) f(x). \tag{2.10}$$

Here, $\kappa : \mathbb{Z}^2 \to \operatorname{Hom}(\mathbb{R}^m, \mathbb{R}^n)$ is a matrix-valued *kernel* (or *filter*) for some natural number $n \in \mathbb{N}$. Note that (2.10) differs from (2.9) not only in dimensionality and the domain of integration/summation, we have also involuted the kernel: $\kappa(x - y)$ versus $\kappa(y - x)$. This means that (2.10) is a cross-correlation operator rather than a convolution operator, but it can easily be turned into the latter by redefining the kernel. **Papers I-III** use the convention (2.10) as it makes some proofs easier to formulate. Convolutional layers are actually implemented in the form of cross-correlations in standard machine learning platforms such as `PyTorch` and `TensorFlow`.

Convolutional layers are trained by optimizing the matrix elements in $\kappa(x)$ for each $x \in \mathbb{Z}^2$. This is possible because, in practice, κ is only supported on a small number of points around the origin in $\mathbb{Z}^2$. When referring to a convolutional layer with a "3×3 kernel", for instance, we mean that

$$\operatorname{supp}(\kappa) = \big\{(x_1, x_2) \in \mathbb{Z}^2 \,\big|\, x_1, x_2 = -1, 0, 1\big\}. \tag{2.11}$$

That is, the kernel size 3×3 only refers to the support of κ, which is different

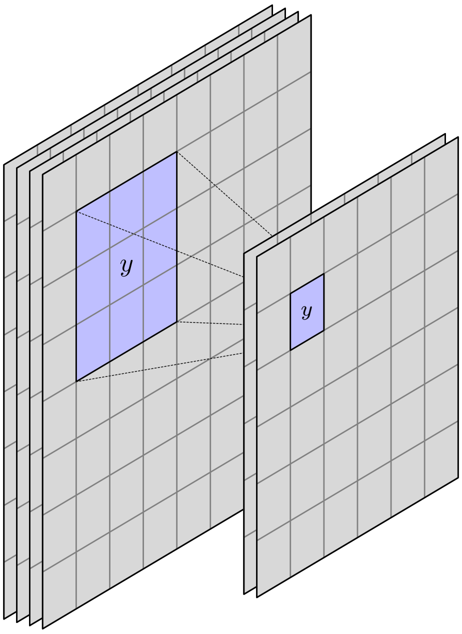


Figure 2.2: A convolutional layer that maps a 4-channel data point $f : \mathbb{Z}^2 \to \mathbb{R}^4$ into a 2-channel data point $\kappa \star f : \mathbb{Z}^2 \to \mathbb{R}^2$. This convolutional layer uses a 3×3 kernel, meaning that κ is supported on a 3×3 grid. This should not be confused with the 4×2 matrix dimension of $\kappa(x)$.

from the matrix dimensions $n \times m$ of $\kappa(x) \in \mathrm{Hom}(\mathbb{R}^m, \mathbb{R}^n)$. The most common kernel sizes are $k \times k$ where k is a small, odd integer.

To compute the output $[\kappa \star f](y)$ of a convolutional layer at a point $y \in \mathbb{Z}^2$, we first transform each m-channel input array $f(x)$ into an n-channel output array $\kappa(x - y)f(x)$ for each point $x \in \mathbb{Z}^2$, and then sum over x. This procedure can be visualized as placing the $k \times k$ kernel on top of the input data point f, with the kernel support centered at y, and computing pointwise inner products between f and each row in the kernel κ (Figure 2.2). When computing $[\kappa \star f](z)$ at another point $z \in \mathbb{Z}^2$, we simply reposition the kernel and repeat the process. Convolutional layers are thereby computed by “sliding” the kernel across the input data point f.

The “sliding kernel” interpretation illustrates that convolutional layers employ *weight sharing*, i.e., the same $k^2 * n * m$ non-zero kernel matrix elements are used to compute the output at each point. The very small number of weights in convolutional layers makes CNNs relatively efficient to train.

There is another interesting consequence of the sliding kernel: Consider the translation operator on $\mathbb{Z}^2$ that translates each grid point by the same amount,

$$L_{x_0} : \mathbb{Z}^2 \to \mathbb{Z}^2, \qquad L_{x_0}(x) = x + x_0, \qquad x_0 \in \mathbb{Z}^2. \tag{2.12}$$

This translation operator induces a translation operator on data points, moving the argument in the opposite direction:

$$(L_{x_0} f)(x) = f(x - x_0). \tag{2.13}$$

Applying a convolutional layer to the translated data point gives the relation

$$\begin{aligned} [\kappa \star L_{x_0} f](y) &= \sum_{x \in \mathbb{Z}^2} \kappa(x - y) f(x - x_0) \\ &= \sum_{x \in \mathbb{Z}^2} \kappa(x - (y - x_0)) f(x) = L_{x_0}[\kappa \star f](y). \end{aligned} \tag{2.14}$$

Convolutional layers thus commute with the translation operator. Intuitively, this means that convolutional layers preserve the global symmetry in $\mathbb{Z}^2$, which is important for applications. As an example, consider using a CNN for a facial recognition task, and suppose for argument's sake that the kernel κ has learned to recognize human eyes. Thanks to the *translation equivariance* (2.14), it does not matter where the eyes are located in any particular image, the sliding kernel will eventually find them.

We can add a bias vector $b \in \mathbb{R}^n$ after the convolutional layer and then apply a non-linear activation function $\sigma : \mathbb{R} \to \mathbb{R}$ to each component. These operations are independent of $x \in \mathbb{Z}^2$, hence the composition

$$\sigma\left(\kappa \star f + b\right) \tag{2.15}$$

is still translation equivariant; it still commutes with the translation operator. We can therefore build arbitrarily long sequences of layers (2.15) that preserve translation equivariance. That being said, there are also other layers that break equivariance. One example is *pooling layers* which are used for downsampling; essentially, throwing away (hopefully redundant) information in order to speed up computations. To give an explicit example, *max pooling layers* split $\mathbb{Z}^2$ into small components and computes the maximum value of data points f in each component, which clearly breaks equivariance (Figure 2.3).

Translation equivariance is an important property of convolutional layers that facilitates learning, but it is not the only aspect that determines the performance of a CNN. Sometimes, achieving higher performance may require using pooling layers or other layers that break equivariance. It is also common practice to use a multilayer perceptron or some other neural network for the last layers in a CNN, and these networks also break equivariance. However, the general idea is that if equivariance aids in the extraction of useful features in the first layers, it may not be needed later on.

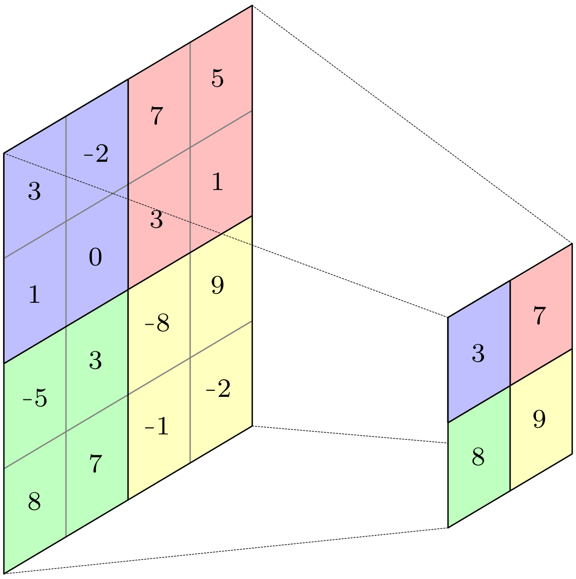


Figure 2.3: Max pooling layer.

2.3 Geometric deep learning

In many machine learning tasks, data is effectively treated as arrays of numbers that have no useful geometric properties. It is common, for example, to reshape 2D arrays into 1D arrays by stacking columns on top of each other, even if this destroys geometric information that may have been present.

Geometric deep learning is an umbrella term, introduced by Bronstein et al. [2017], for deep learning methods that make direct use of geometric information in data. CNNs are among the simplest examples of geometric deep learning methods, as convolutional layers use the translation symmetry in $\mathbb{Z}^2$ to efficiently solve learning tasks with relatively little data and few parameters. On the other hand, CNNs still assume that data points are flat arrays of numbers.

Other types of data have natural curvature, one example being meteorological data on intercontinental regions. Another example is spherical images, such as those used in `Google Street View`. Furthermore, aerial photographs may have a flat geometry and can be processed using ordinary CNNs, but there is an orientation ambiguity in aerial photographs that CNNs do not understand. When processing such images, it would be desirable to use convolutional layers that commute not only with translations but also with rotations, i.e., layers that are equivariant with respect to planar rototranslations.

Translations and rotations are *global* symmetry transformations. When people refer to symmetric objects, such as the mirror symmetry of a butterfly or the 120° rotational symmetry of an equilateral triangle, they are implicitly referring to global symmetries. Translations, rotations, and (mirror) reflections move the elements of the underlying space, such as the rotation matrix

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \theta \in \mathbb{R}, \tag{2.16}$$

which rotates elements of $\mathbb{R}^2$. If we model the circle as the set

$$S^1 = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 \,\middle|\, x_1^2 + x_2^2 = 1 \right\}, \tag{2.17}$$

then applying the rotation matrix $R(\theta)$ to each point $x \in S^1$ causes each point to move, but the set itself is unaffected. Intuitively, the circle is globally symmetric under rotations because it is rotationally invariant as a set.

Local symmetry refers to properties that are invariant to certain internal degrees of freedom. Local symmetries are also known as *gauge symmetries* and were first studied by physicists, since they arise naturally in electromagnetism, quantum chromodynamics, and other areas. However, local symmetries are also relevant in mathematics and in various applications. Anyone who has studied linear algebra knows that vectors can be represented in different bases, and this affects their numerical representation as column vectors. For example, the velocity of a moving car is a well-defined geometric vector in classical mechanics. It has direction and magnitude, but its representation as a column vector

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = a\boldsymbol{e}_1 + b\boldsymbol{e}_2 + c\boldsymbol{e}_3 \in \mathbb{R}^3, \tag{2.18}$$

depends on which basis elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3 \in \mathbb{R}^3$ have been chosen. The *choice of basis* is an internal degree of freedom in mathematical models that affects how vectors (and matrices and tensors) are represented in, say, a computer. This choice can affect computations and conclusions drawn from the model, despite being irrelevant for the underlying application. The freedom to choose a basis is a superfluous degree of freedom internal to the model; it is a local symmetry.

Taking relevant symmetries into account when designing neural networks can simplify the learning task, by making the symmetries built-in properties of the network. Such networks can use symmetry, by means of equivariance, as a tool for extracting relevant information from data. This also reduces the amount of information that networks need to learn. Local and global symmetries are geometric properties and *equivariant neural networks* are thus geometric deep learning methods.

LeCun et al. [1989] pioneered the use of convolutional layers in neural networks, and Cohen and Welling [2016a] were among the first to extend CNNs beyond translation equivariance. Their *group-equivariant convolutional neural networks (GCNNs)* allow equivariance under a swathe of global symmetries. GCNNs use convolutional layers of the form

$$[\kappa \star f](g) = \int_G \kappa(g^{-1}g')f(g')\,\mathrm{d}g', \qquad g \in G, \tag{2.19}$$

where G is the global symmetry group under consideration, $\mathrm{d}g'$ is a left Haar measure on the group, f is a vector-valued function representing input data, and κ is a matrix-valued sliding kernel. The discrete convolutions (2.10) used in ordinary CNNs correspond to the special case $G = \mathbb{Z}^2$ of discrete translation symmetry in two dimensions. Note that the Haar measure on $\mathbb{Z}^2$ is the counting measure, so the integral (2.19) does reduce to a discrete sum in this case. The introduction of GCNNs was a breakthrough not only in terms of generalization, GCNNs have a rich mathematical theory that builds upon representation theory and the theory of fiber bundles [Cohen et al., 2018b]. This connection to highly developed areas of mathematics has enabled rapid progress in the field.

Here are some of the most popular examples of GCNNs:

- *Spherical CNNs* ($G = SO(3)$) are powerful models for analyzing spherical data and solving rotation symmetric tasks in $\mathbb{R}^3$, since their convolutional layers commute with rotations [Cohen et al., 2018a, Gerken et al., 2022, Toft et al., 2022].

- As mentioned above, tasks involving digital images sometimes benefit from rotational equivariance in addition to the translation equivariance of CNNs. Such networks have been successfully used, e.g., in medical image analysis [Bekkers et al., 2018, Veeling et al., 2018, Lafarge et al., 2021].

- *Graph neural networks* are equivariant with respect to permutations, and therefore commute with elements of a symmetric group $G = S_n$ for some natural number n. Elements of this group are bijections of the set $\{1, \dots, n\}$ and they act on any enumerated set of objects $\{x_1, \dots, x_n\}$ by permuting indices. This makes graph neural networks apt for applications involving molecular data, social networks, or other types of graph data. The vertex enumeration of a graph is often not intrinsically important information, and graph neural networks produce consistent output for different vertex enumerations. See [Zhou et al., 2020] for an extensive review.

  There are many different kinds of graph neural networks and not all of them are special cases of GCNNs. However, any permutation-equivariant linear map can be expressed as a convolutional layer in a GCNN.

While GCNNs are among the most well-known equivariant neural networks, there are many other related models. These include, for instance, *steerable CNNs* [Cohen and Welling, 2016b, Weiler et al., 2018, Cesa et al., 2021], *B-spline CNNs* [Fey et al., 2018, Bekkers, 2019], and *PDE-based GCNNs* [Smets et al., 2023].

Local symmetry is arguably more niche and has therefore received less attention, but there are some interesting applications nevertheless. The applications mainly go in two directions:

1. We mentioned earlier that local symmetry originated in physics. Neural networks that are equivariant with respect to local (gauge) symmetries have been developed for applications primarily in lattice gauge theory [Luo et al., 2021, Favoni et al., 2022]. Equivariant neural networks have been shown to approximate various physical quantities and solve tasks more effectively than ordinary CNNs [Favoni et al., 2022]. They have also been suggested as alternatives to Monte Carlo methods in some situations where the latter are relatively inefficient [Kanwar et al., 2020].

2. *Learning on manifolds* refers broadly to deep learning methods designed to process data on curved spaces in geometrically consistent ways. Ordinary neural networks can process such data by working in local coordinates in different regions, but the network output can depend heavily on the choice of coordinates and the induced frame of reference. There is also no guarantee that overlapping regions produce comparable output. Equivariant networks solve this problem by imposing constraints that make the layers transform in certain ways under change of coordinates [Cheng et al., 2019, Weiler et al., 2021]. The choice of local coordinate system or, more precisely, of the induced reference frame, is a local symmetry just like the choice of basis discussed earlier.

The books by Bronstein et al. [2021] and Weiler et al. [2021] are two excellent resources for learning more about geometric deep learning.

Paper I investigates whether equivariance to global symmetries can be achieved by other means than through convolutional layers in a GCNNs. The framework used in this paper also describes equivariance to local symmetries. In **Paper II**, we review different variants of equivariant neural networks for local or global symmetries. Finally, in **Paper III** we develop neural networks for applications in lattice gauge theory. These networks are simultaneously equivariant with respect to global lattice symmetry and local gauge symmetry.

3 Representation theory

Assume that a given learning task is symmetric in some sense, for example translation invariance in image classification, or rotational symmetry in radial tasks such as depth estimation. The symmetry is encoded in a group G that acts by linear transformations on vectors (or scalars or tensors). By extension, these symmetry transformations also act on data points, which are modeled as vector-valued functions. Neural network layers also act on data points and we say, roughly speaking, that a neural network is *equivariant* if its layers commute with the symmetry transformations.

The assignment of a linear transformation $\rho(g)$ to each group element $g \in G$ is called a *representation*. Their induced action on data points make representations central to the study of equivariant neural networks. In fact, the brief description of equivariance given above can be sharpened: Layers in an equivariant neural network *intertwine* representations on the relevant input and output spaces. We will say more about this later.

Definition 1. Let G be a Lie group. A *(strongly continuous) representation* of G on a topological vector space V is a group homomorphism

$$\rho : G \to GL(V), \tag{3.1}$$

such that the following map is continuous for all $g \in G$ and all $v \in V$:

$$G \times V \to V, \qquad (g, v) \mapsto \rho(g)v. \tag{3.2}$$

Representations are denoted (ρ, V) or simply ρ. A representation is *unitary* if V is a complex Hilbert space and $\rho(g)$ is unitary, $\rho(g)^{-1} = \rho(g)^{\dagger}$, for all $g \in G$.

Even though almost everything in this chapter holds for locally compact groups, we restrict attention to Lie groups. This is because the theory introduced here is intended to be used in the next chapter, in which all spaces (including groups) are required to be manifolds.

Theorem (Hilbert's fifth problem). *Assume that G is both a topological group and a topological manifold. Then G is a Lie group.*

Remark 1. Manifolds are assumed to be Hausdorff and second countable.

Remark 2. Any finite or countably infinite group G is a 0-dimensional Lie group when equipped with the discrete topology and the smooth structure defined by the atlas

$$\mathcal{A} = \{(U_g, \varphi_g) \mid g \in G\} . \tag{3.3}$$

Here, the chart (U_g, φ_g) corresponding to each $g \in G$ is given by $U_g = \{g\}$ and

$$\varphi_g : \{g\} \to \mathbb{R}^0, \qquad g \mapsto 0. \tag{3.4}$$

As such, the representation theory of Lie groups includes that of finite groups and of countably infinite groups.

We will give some examples of real representations, and representations can be studied for vector spaces over any field. Complex representations are especially well-behaved, however, because the complex numbers are algebraically closed. Most introductory texts focus on complex representations for this reason, and so do we: Unless otherwise stated, vector spaces V are assumed to be complex.

Example 1. Let G be a topological group. For any topological vector space V, there is a *trivial representation* that sends each $g \in G$ to the identity operator,

$$\rho : G \to GL(V), \qquad \rho(g) = \mathrm{Id}_V . \tag{3.5}$$

The special case $V = \mathbb{C}$ is known as *the* trivial representation. ■

Example 2. Consider the rotation group $SO(2)$ and identify each element with its angle of rotation θ. Then the map

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{3.6}$$

defines a representation on $\mathbb{R}^2$ as well as on $\mathbb{C}^2$. Indeed, for all angles θ, ϕ,

$$\begin{aligned} R(\theta)R(\phi) &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -\cos\theta\sin\phi - \sin\theta\cos\phi \\ \sin\theta\cos\phi + \cos\theta\sin\phi & \cos\theta\cos\phi - \sin\theta\sin\phi \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{bmatrix} = R(\theta+\phi), \end{aligned}$$

where we have applied standard trigonometric identities for the sum of two angles. In particular, $R(\theta)R(-\theta) = R(0) = \mathrm{Id}$ and so $R(-\theta) = R(\theta)^{-1}$. ■

Example 3. Let S_n be the symmetric group whose elements are permutations of n objects, for $n = 1, 2, 3, \ldots$. The elements $g \in S_n$ can be modeled as bijections

$$g : \{1, \ldots, n\} \to \{1, \ldots, n\}. \tag{3.7}$$

The symmetric group can be used to relabel the vertices in a graph, for example, and is thus relevant for graph neural networks. Given any basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n \in V$ in an n-dimensional vector space, and a permutation $g \in S_n$, there is a unique linear operator $\rho(g)$ that permutes the basis vectors:

$$\rho(g)\boldsymbol{e}_i = \boldsymbol{e}_{g(i)}, \qquad i = 1, \ldots, n. \tag{3.8}$$

The mapping $g \mapsto \rho(g)$ is a representation of S_n. ■

The next definition shows how to build representations from simpler ones.

Definition 2. The *direct sum* of two G-representations (ρ, V_ρ) and (σ, V_σ) is the representation $(\rho \oplus \sigma, V_\rho \oplus V_\sigma)$ defined by

$$(\rho \oplus \sigma)(g) : v \oplus w \mapsto \rho(g)v \oplus \sigma(g)w. \tag{3.9}$$

Similarly, their *tensor product* is the representation $(\rho \otimes \sigma, V_\rho \otimes V_\sigma)$ defined by

$$(\rho \otimes \sigma)(g) : v \otimes w \mapsto \rho(g)v \otimes \sigma(g)w. \tag{3.10}$$

We can also deconstruct some representations into smaller ones:

Definition 3. Let (ρ, V) be a G-representation and suppose that $W \subseteq V$ is a linear subspace that is closed under the representation,

$$\rho(g)w \in W, \qquad w \in W, g \in G. \tag{3.11}$$

Then W is an *invariant subspace*, and (ρ, W) is a *subrepresentation* of (ρ, V).

All representations (ρ, V) have two obvious subrepresentations, obtained by letting $W \subseteq V$ be either the full space $W = V$ or the trivial subspace $W = \{0\}$. These subrepresentations are not very interesting and are not considered *proper*, because W is not a proper subspace.

Definition 4. A representation is *reducible* if it has a proper subrepresentation. Representations that are not reducible are called *irreducible*.

Example 4. A trivial representation (Id_V, V) is irreducible iff $\dim V = 1$. ■

The next lemma explains why unitary representations are especially interesting: They can be decomposed into direct sums of subrepresentations.

Lemma 1. *Let (ρ, V) be a unitary representation and suppose that $W \subseteq V$ is a closed, invariant subspace. Then the orthogonal complement $W^\perp$ is also a closed invariant subspace, and (ρ, V) decomposes into a direct sum representation on $W \oplus W^\perp$.*

Proof. Recall the definition of the orthogonal complement,

$$W^\perp = \{v \in V \mid \langle w, v\rangle = 0 \text{ for all } w \in W\}\,. \tag{3.12}$$

According to a standard result in functional analysis, if V is a Hilbert space and $W \subseteq V$ is a closed subspace, then $W^\perp$ is a closed subspace and $V = W \oplus W^\perp$. In particular, W and $W^\perp$ are Hilbert spaces in their own right. It is therefore sufficient to prove that $W^\perp$ is an invariant subspace, i.e., that $\rho(g)v \in W^\perp$ for each $v \in W^\perp$ and all $g \in G$. Because the representation ρ is unitary, it satisfies

$$\rho(g)^\dagger = \rho(g)^{-1} = \rho(g^{-1}), \tag{3.13}$$

and the lemma now follows from W being an invariant subspace:

$$\langle w, \rho(g)v\rangle = \langle \rho(g)^\dagger w, v\rangle = \langle \rho(g^{-1})w, v\rangle = 0, \tag{3.14}$$

for all $w \in W$. That is, $\rho(g)v \in W^\perp$. □

It seems likely that unitary representations can be decomposed into direct sums of *irreducible* subrepresentations, if we just keep decomposing a representation into smaller and smaller subrepresentations until these cannot be decomposed any further. This is certainly true for compact groups, which follows from the famous Peter-Weyl theorem [Deitmar and Echterhoff, 2014, Theorem 7.2.4]. It is also true for many other groups if we replace direct sums with *direct integrals*. In this sense, irreducible representations are the prime numbers of representation theory; the building blocks used to construct all other representations.

It is sometimes possible to translate one representation into another. An almost trivial example is that the real $SO(2)$-representation (3.6) can be transformed into the following real $SO(2)$-representation,

$$\tilde{R}(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & \\ \sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{3.15}$$

that performs rotations about the y-axis in $\mathbb{R}^3$. All we need to do is identify $\mathbb{R}^2$ with the xz-plane in $\mathbb{R}^3$ via the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^3$ given by

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{3.16}$$

in the standard bases. If we fix a vector $v \in \mathbb{R}^2$ and a rotation angle θ, it seems likely that the following two procedures give the same result:

1. First use $R(\theta)$ to rotate v and then map the rotated vector to the xz-plane.
2. First map v to the xz-plane and then use $\tilde{R}(\theta)$ to rotate about the y-axis.

Indeed, a short calculation shows that $T \circ R(\theta) = \tilde{R}(\theta) \circ T$. In this example, the relation between $R(\theta)$ and $\tilde{R}(\theta)$ was rather obvious, but this way of relating two representations can be applied much more generally.

Definition 5. An *intertwiner*, or *equivariant map*, between two G-representations (ρ, V_ρ), (σ, V_σ) is a bounded linear map $T : V_\rho \to V_\sigma$ satisfying, for all $g \in G$,

$$T \circ \rho(g) = \sigma(g) \circ T. \tag{3.17}$$

We let $\mathrm{Hom}_G(V_\rho, V_\sigma)$ denote the space of all such intertwiners. Moreover, if an intertwiner $T \in \mathrm{Hom}_G(V_\rho, V_\sigma)$ is a (unitary) isomorphism, then (ρ, V_ρ), (σ, V_σ) are said to be *(unitarily) equivalent*.

Example 5. When viewed as a complex representation, the $SO(2)$-representation (3.6) is unitarily equivalent to the direct sum representation $\rho(\theta) = e^{i\theta} \oplus e^{-i\theta}$:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = T \begin{bmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{bmatrix} T^\dagger, \tag{3.18}$$

where $T = \frac{1}{\sqrt{2}} \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix}$. ■

Irreducible representations have strict limitations on the possible intertwiners, as the next lemma shows. This is a famous result known as *Schur's first lemma*.

Lemma 2 [Deitmar and Echterhoff, 2014, Lemma 6.1.7]. *Let (ρ, V) be a unitary representation of a topological group G. Then the following are equivalent:*

(a) (ρ, V) is irreducible.

(b) If $T : V \to V$ is an intertwiner, there is a constant $\lambda \in \mathbb{C}$ such that $T = \lambda\,\mathrm{Id}$.

This lemma is motivated by the following observation: Let (ρ, V) be a unitary representation of G and suppose that $T : V \to V$ is an intertwiner. If λ is an eigenvalue of T, then the corresponding eigenspace E_λ is an invariant subspace of V since, for each $v \in E_\lambda$ and all $g \in G$,

$$T\rho(g)v = \rho(g)Tv = \lambda\rho(g)v, \tag{3.19}$$

hence $\rho(g)v \in E_\lambda$. The unitary representation (ρ, V) is thus reducible if E_λ is a proper subspace. Equivalently, if (ρ, V) is irreducible, we must have $E_\lambda = V$ which means that $Tv = \lambda v$ for all $v \in V$ and so $T = \lambda \operatorname{Id}$. This is not a complete proof because it assumes the existence of an eigenvalue, but it provides some intuition. The other direction is more straightforward.

The next result, *Schur's second lemma*, is even more famous than the first. It is proven by noting that for any intertwiner $T : V_\rho \to V_\sigma$, its adjoint $T^\dagger$ is also an intertwiner and the same is true of their composition $T^\dagger T : V_\rho \to V_\rho$. Schur's first lemma can then be applied to the positive semi-definite map $T^\dagger T$, which lets us extract information about T.

Corollary 3 [Deitmar and Echterhoff, 2014, Corollary 6.1.9]**.** *Let (ρ, V_ρ), (σ, V_σ) be irreducible unitary representations and assume that $T : V_\rho \to V_\sigma$ is an intertwiner. Then T is either zero or invertible with continuous inverse. In the latter case there is a scalar $c > 0$ such that cT is unitary. The space $\operatorname{Hom}_G(V_\rho, V_\sigma)$ is zero unless ρ and σ are unitarily equivalent, in which case the space has dimension 1.*

Let us end by mentioning the regular representations of unimodular Lie groups. For reasons explained in **Papers I-II**, these representations are closely connected to convolutional layers in GCNNs as well as to Fourier analysis. This enables the use of Fourier analytic methods when studying GCNNs.

Example 6. Let G be a unimodular Lie group. For each $g \in G$, define the map

$$\mathcal{L}_g : L^2(G) \to L^2(G), \qquad (\mathcal{L}_g f)(g') = f(g^{-1}g'), \tag{3.20}$$

for $f \in L^2(G)$ and $g' \in G$. The assignment $g \mapsto \mathcal{L}_g$ is a unitary representation of G called the *left regular representation*. It is unitary thanks to left-invariance of the Haar measure on G:

$$\begin{aligned} \langle \mathcal{L}_g f, \mathcal{L}_g f' \rangle &= \int_G f(g^{-1}g')\overline{f'(g^{-1}g')}\,\mathrm{d}g' \qquad [g' \mapsto gg'] \\ &= \int_G f(g')\overline{f'(g')}\,\mathrm{d}g' = \langle f, f' \rangle. \end{aligned}$$

Similarly, one can define a unitary *right regular representation* by

$$\mathcal{R}_g : L^2(G) \to L^2(G), \qquad (\mathcal{R}_g f)(g') = f(g'g), \tag{3.21}$$

for $f \in L^2(G)$ and $g, g' \in G$. ■

The involution operator $(Tf)(g) = f(g^{-1})$ intertwines the left and right regular representations and leads to the following lemma.

Lemma 4. *The left and right regular representations of a unimodular Lie group G are unitarily equivalent.*

4 Mathematical gauge theory

In this chapter we offer an introduction to mathematical gauge theory, starting with fiber bundles. This is arguably the most important background section in this thesis, as fiber bundles are foundational to the theory of equivariant neural networks studied in **Papers I-III**. For example, the data points used as inputs to equivariant neural networks are modeled as certain functions, called sections, of vector bundles.

After introducing fiber bundles, we discuss connections and parallel transport on principal bundles and associated bundles, and also Yang-Mills theory; these topics are especially relevant for the lattice gauge equivariant neural networks in **Paper III**. These sections are based on the expositions in Hamilton [2017], Schuller [2016], and Sontz [2015].

One of the main takeaways from **Paper I** is that homogeneous vector bundles form a natural setting for group equivariant convolutional neural networks. In the final section of this chapter, we analyze how the elements of such bundles transform under left-translation, and how left-translations are related to parallel transport.

4.1 Fiber bundles

As mentioned above, fiber bundles are central to the mathematical framework for equivariant neural networks studied in **Papers I-III**. Let us introduce the subject with a few motivating examples. Much more detailed expositions of fiber bundles can be found in Husemöller [1966] and Kolár et al. [2013].

Consider a smooth manifold $\mathcal{M}$ of dimension d and recall that, for each $x \in \mathcal{M}$, there is an associated tangent space $T_x\mathcal{M}$. Vector fields on $\mathcal{M}$ assign a tangent vector $X_x \in T_x\mathcal{M}$ to each point x of an open subset $U \subseteq \mathcal{M}$ and we would like to view vector fields as smooth functions. However, since the vectors X_x lie in different tangent spaces for different points $x \in U$, the concept of smooth vector fields makes little sense unless the spaces $T_x\mathcal{M}$ themselves vary smoothly in x. The *tangent bundle* $T\mathcal{M}$ solves this problem by defining a suitable topology and smooth structure on the disjoint union of all tangent spaces,

$$T\mathcal{M} := \dot{\bigcup_{x \in \mathcal{M}}} T_x\mathcal{M}. \tag{4.1}$$

The structure on $T\mathcal{M}$ is defined in such a way that the projection $\pi : T\mathcal{M} \to \mathcal{M}$, $X_x \mapsto x$, that sends a tangent vector to the point it is attached to on the manifold, is smooth. Vector fields are then defined as smooth maps

$$X : U \to T\mathcal{M}, \quad \text{satisfying} \quad \pi \circ X = \mathrm{Id}_U, \tag{4.2}$$

ensuring that $X(x) = X_x$ lies in the correct tangent space $T_x\mathcal{M}$ for each $x \in U$. We picture the tangent bundle as a copy of the manifold $\mathcal{M}$ with each tangent space $T_x\mathcal{M}$ attached at its corresponding point $x \in \mathcal{M}$.

Now choose a local coordinate chart

$$(x^1, \dots, x^d) : U \to \mathbb{R}^d, \qquad U \subseteq \mathcal{M}. \tag{4.3}$$

For $\mu = 1\dots, d$ and each point $x \in U$, the coordinate tangent vectors $(\partial_\mu)_x$ are functionals that act on smooth functions $f : \mathcal{M} \to \mathbb{R}$ by computing their partial derivative with respect to x^μ and evaluating at the point x:

$$(\partial_\mu)_x : f \mapsto \left.\frac{\partial f}{\partial x^\mu}\right|_x . \tag{4.4}$$

For each $x \in U$, these coordinate tangent vectors form a basis in $T_x\mathcal{M}$ called a *coordinate basis* and, as suggested by our notation, the mapping

$$\partial_\mu : U \to T\mathcal{M}, \qquad x \mapsto (\partial_\mu)_x, \tag{4.5}$$

is a smooth vector field for each $\mu = 1, \dots, d$.

Remark 3. In order to reduce clutter, we often write ∂_μ in lieu of $(\partial_\mu)_x$ even when referring to individual coordinate tangent vectors rather than the field.

The coordinate basis provides an isomorphism $T_x\mathcal{M} \simeq \mathbb{R}^d$ for each $x \in U$, so the tangent bundle can locally be identified with the Cartesian product $U \times \mathbb{R}^d$. One may expect this idea to extend to the entire tangent bundle, making $T\mathcal{M}$ isomorphic to $\mathcal{M} \times \mathbb{R}^d$. If so, we would call $T\mathcal{M}$ a *trivlal* bundle. In general, however, the lack of a global coordinate chart prevents this idea from coming to fruition; the tangent bundles $T\mathcal{M}$ of most manifolds $\mathcal{M}$ have more intricate geometries and are not isomorphic to $\mathcal{M} \times \mathbb{R}^d$.

Lie groups G have a similar geometric structure: If $K \leq G$ is a closed subgroup, then G is the disjoint union of cosets

$$G = \dot{\bigcup_{g \in G}} gK. \tag{4.6}$$

Each coset is considered a subset $gK \subset G$ when viewed as a collection of group elements, and it is considered a point $gK = x \in G/K$ when viewed as an object (a set) in its own right. These two roles played by cosets are closely related, of course, but may cause confusion nevertheless. The quotient map $q : G \to G/K$ sends each group element $g \in G$ to its coset $gK \in G/K$.

If we first picture the quotient manifold G/K as a collection of points $gK \in G/K$, and we then change perspective by considering cosets as subsets $gK \subset G$, then each subset $gK \subset G$ is effectively attached or glued onto the point $gK \in G/K$. This is analogous to tangent spaces $T_x\mathcal{M}$ being attached at points $x \in \mathcal{M}$.

One can also define analogues of vector fields as smooth maps

$$s : U \to G, \quad \text{satisfying} \quad q \circ s = \mathrm{Id}_U, \tag{4.7}$$

where $U \subseteq G/K$ is an open subset. It turns out that Lie groups G are locally isomorphic to Cartesian products $U \times K$ but again, this local property does not hold globally: Most groups G are not isomorphic to $G/K \times K$.

In bot of these examples, we attached a *fiber* ($T_x\mathcal{M}$ or gK) to each element of a *base space* ($\mathcal{M}$ or G/K), thereby obtaining a larger *total space* ($T\mathcal{M}$ or G). A surjective *projection* (π or q) mapped each element of the total space to the point on the base space where the corresponding fiber was attached. Furthermore, the total space could locally be viewed as a Cartesian product ($U \times \mathbb{R}^d$ or $U \times K$) involving a *characteristic fiber* ($\mathbb{R}^d$ or K), but this local property did not always extend globally. Together, these properties define a fiber bundle.

Definition 6. A *(smooth) fiber bundle* is a structure (E, π, X, F), where E, X, F are smooth manifolds and where $\pi : E \to X$ is a smooth surjective map with the following property: For each $x \in X$, there exists a neighbourhood $U \subseteq X$ containing x and a diffeomorphism

$$\phi : \pi^{-1}(U) \to U \times F, \tag{4.8}$$

called a *local trivialization*, such that the following diagram commutes:

$$\begin{tikzcd}
\pi^{-1}(U) \arrow[r, "\phi"] \arrow[dr, "\pi"'] & U \times F \arrow[d, "\pi_1"] \\
& U
\end{tikzcd}$$

Here, $\pi_1 : U \times F \to U$ is the projection onto the first coordinate. The smooth manifolds E, X, and F are respectively called the *total space*, the *base space*, and the *characteristic fiber* of the bundle.

Remark 4. It is common to denote fiber bundles by their projection $\pi : E \to X$, or simply by the total space E, leaving the other ingredients implicit.

There are many different kinds of fiber bundles, depending on the characteristic fiber F and its properties.

Definition 7. Suppose that the fibers $E_x = \pi^{-1}(\{x\})$ of a fiber bundle $\pi : E \to X$ are finite-dimensional vector spaces. If the mappings $F \to E_x$ given by

$$v \mapsto \phi^{-1}(x, v), \tag{4.9}$$

are linear isomorphisms for all local trivializations $\phi : \pi^{-1}(U) \to U \times F$ and each $x \in U$, then E is called a *(smooth) vector bundle.*

Example 7. The tangent bundle is a smooth vector bundle. Its fibers $\pi^{-1}(\{x\}) = T_x\mathcal{M}$ are d-dimensional vector spaces and the local trivializations are obtained from local coordinate charts, hence (4.9) becomes

$$\phi : \mathbb{R}^d \to T_x\mathcal{M}, \qquad \left(v^1, \ldots, v^d\right) \mapsto \sum_{\mu=1}^{d} v^\mu \partial_\mu. \tag{4.10}$$

which is an isomorphism of vector spaces. ■

Remark 5. We henceforth use Einstein notation for implied summation over pairs of matching indices. For example, the sum in (4.10) is simply written as

$$v^\mu \partial_\mu. \tag{4.11}$$

Definition 8. Let $\pi : E \to X$ and $\pi' : E' \to X'$ be fiber bundles. A smooth map $\varphi : E \to E'$ is called a *bundle map (bundle morphism)* if there is a smooth map

$f : X \to X'$ such that the following diagram commutes:

$$\begin{tikzcd} E \arrow[r, "\varphi"] \arrow[d, "\pi"'] & E' \arrow[d, "\pi'"] \\ X \arrow[r, "f"] & X' \end{tikzcd}$$

If both φ and f are diffeomorphisms, the bundle map is called an *isomorphism of bundles*, or a *bundle isomorphim*.

Remark 6. The notion of *isomorphism* introduced in Definition 8 depends on the type of bundle. An *isomorphism of vector bundles*, for example, must be linear on each fiber.

We mentioned earlier that the tangent bundle $T\mathcal{M}$ can be isomorpic to $\mathcal{M} \times \mathbb{R}^d$, even though this is rarely the case. Cartesian products such as $\mathcal{M} \times F$ trivially satisfy the definition of a fiber bundle

$$\pi : \mathcal{M} \times F \to \mathcal{M}, \qquad \pi(x, f) = x. \tag{4.12}$$

Definition 9. A fiber bundle $\pi : E \to \mathcal{M}$ is *trivial* if it is isomorphic to $\mathcal{M} \times F$.

Definition 10. Let $\pi : E \to \mathcal{M}$ be a fiber bundle and let $U \subseteq \mathcal{M}$ be an open set. A *(local) section* is a smooth map $s : U \to E$ satisfying

$$\pi \circ s = \mathrm{Id}_U\,. \tag{4.13}$$

That is, sections map each point $x \in U$ to an element of its fiber E_x.

In the equivariant neural networks studied in **Papers I-III**, each data point is modeled as a section of a vector bundle. Data points therefore include vector fields, which are sections of the tangent bundle $T\mathcal{M}$, but they also include other types of fields such as scalar fields, tensor fields and other types of sections.

The word *gauge* in mathematical gauge theory refers to sections $\sigma : U \to P$ of principal bundles. Principal bundles and their gauges are thus of fundamental importance, as we shall see.

Definition 11. Let K be a Lie group. A *(smooth) principal bundle* with *structure group K* is a fiber bundle $\pi : P \to \mathcal{M}$ equipped with a free, smooth, right action

$$P \times K \to P, \qquad (p, k) \mapsto p \triangleleft k, \tag{4.14}$$

with the following properties for each $x \in \mathcal{M}$.

(i) Let $P_x = \pi^{-1}(\{x\})$ be the fiber at x. Then

$$p \in P_x,\ k \in K \quad \Rightarrow \quad p \triangleleft k \in P_x. \tag{4.15}$$

That is, the K-action preserves fibers.

(ii) For each $p \in P_x$, the mapping $k \mapsto p \triangleleft k$ is a diffeomorphism $K \to P_x$.

Principal bundles with structure group K are also called *principal K-bundles*.

GCNNs are concerned with the following principal bundle, which we touched upon in the introduction to this part.

Proposition 5. *Let G be a Lie group and let $K \leq G$ be a closed subgroup. Then the quotient map*

$$q : G \to G/K, \qquad g \mapsto gK, \tag{4.16}$$

defines a principal bundle over $\mathcal{M} = G/K$ with structure group K.

Proof. It is known [Steenrod, 1960, pp. 31-33] that for any Lie group G and any closed subgroup K, there exists a local gauge $\sigma : U \to G$ on some open subset $U \subset G/K$. Use this gauge to define a mapping $\phi : q^{-1}(U) \to U \times K$ by

$$\phi(g) = \big(q(g), \sigma(q(g))^{-1}g\big) = (gK, k^{-1}), \tag{4.17}$$

where $k = g^{-1}\sigma(q(g)) \in K$. Then ϕ is smooth, and so is its inverse

$$\phi^{-1}(gK, k) = \sigma(gK)k, \tag{4.18}$$

hence ϕ is a diffeomorphism. Moreover, it is evident from (4.17) that $q = \pi_1 \circ \phi$, and ϕ is therefore a local trivialization around any point $gK \in U$. As for points $gK \notin U$, fix an arbitrary $\tilde{g} \in G$ and define the map

$$\tilde{\sigma} : \tilde{g}U \to G, \qquad \tilde{\sigma}(gK) = \tilde{g}\sigma(gK). \tag{4.19}$$

This map is also a gauge and thereby induces a local trivialization $\phi_{\tilde{g}}$ in the same manner as above. Since any point $gK \in G/K$ lies in some neighbourhood $\tilde{g}U$ and thus admits a local trivialization $\phi_{\tilde{g}}$, we find that $q : G \to G/K$ is a fiber bundle. Finally, observe that right-multiplication

$$G \times K \to G, \qquad (g, k) \mapsto gk, \tag{4.20}$$

is a free, smooth, right K-action that satisfies both conditions in Definition 11, implying that $q : G \to G/K$ is a principal bundle with structure group K. □

Definition 12. Consider a principal K-bundle $\pi : P \to \mathcal{M}$ and suppose that $\rho : K \to GL(V_\rho)$ is a finite-dimensional representation of the structure group. Now define the following equivalence relation $\sim$ on the product $P \times V_\rho$,

$$(p, v) \sim (p \triangleleft k, \rho(k)^{-1}v), \qquad p \in P, v \in V_\rho, k \in K. \tag{4.21}$$

Then the quotient space

$$P \times_\rho V_\rho = \{\text{equivalence classes } [p, v] \mid p \in P, v \in V_\rho\}, \tag{4.22}$$

is called an *associated (vector) bundle*.

In **Paper I**, we give an example that shows explicitly how the tangent bundle $T\mathcal{M}$ is an associated bundle. More importantly, we demonstrate that GCNNs are naturally formulated in terms of *homogeneous vector bundles*, which are also associated bundles. In other words, this concept is very useful.

Of course, one cannot simply claim that associated bundles $P \times_\rho V_\rho$ are bundles by virtue of their name, it must be proven. One proof uses a gauge $\sigma : U \to P$ to pick representatives $(\sigma(x), v) \in P \times V_\rho$ of equivalence classes $[\sigma(x), v] \in P \times_\rho V_\rho$, and these representatives are then used to construct the local trivializations. A complete proof can be found in [Kolár et al., 2013, §10.7]

Proposition 6. *Let P be a principal K-bundle and let (ρ, V_ρ) be a finite-dimensional representation of the structure group K. Then $P \times_\rho V_\rho$ is a vector bundle.*

The next proposition explains the aforementioned problem of choosing bases in tangent spaces. The family of all bases in all tangent spaces $T_x\mathcal{M}$ is a principal bundle called the *frame bundle $F\mathcal{M}$*. If it would be possible to smoothly assign a basis to each tangent space, then that assignment would define a global gauge $\sigma : \mathcal{M} \to F\mathcal{M}$. For most manifolds, however, the frame bundle is nontrivial and does not admit a global gauge.

Proposition 7. *A principal K-bundle P is trivial iff it admits a global gauge.*

Proof. First assume that P is trivial and let $\varphi : \mathcal{M} \times K \to P$ be an isomorphism of bundles. Then the following map is global gauge for any fixed $k \in K$:

$$\sigma : \mathcal{M} \to P, \qquad \sigma(x) = \varphi(x, k). \tag{4.23}$$

For the other direction, let $\sigma : \mathcal{M} \to P$ be a global gauge and define the map

$$\phi : \mathcal{M} \times K \to P, \qquad (x, k) \mapsto \sigma(x) \triangleleft k, \tag{4.24}$$

which is smooth, as the K-action and σ are both smooth. If we now fix $x \in \mathcal{M}$, then the restriction $\phi_x : K \to P_x$ given by $k \mapsto \phi(x, k)$ is a diffeomorphism by Definition 11(ii), hence (4.24) has a smooth inverse

$$\phi^{-1} : P \to \mathcal{M} \times K, \qquad p \mapsto \big(\pi(p), \phi^{-1}_{\pi(p)}(p)\big). \tag{4.25}$$

Furthermore, (4.24) preserves basepoints and satisfies

$$\phi(x, kk') = \omega(x) \triangleleft k \triangleleft k' = \phi(x, k) \triangleleft k', \tag{4.26}$$

and is therefore an isomorphism of principal bundles. That is, P is trivial. □

4.2 Connections and parallel transport

As discussed in the previous part, the theory of equivariant neural networks models the input data as sections $s : U \to E$ of vector bundles. The argument is that sections contain geometric information that equivariant neural networks can extract and learn from, thereby reducing the amount of data and/or model parameters needed to solve a given task. One difficulty arising from this way of modeling data is that we cannot compute linear combinations

$$\alpha s(x) + \beta s(y), \tag{4.27}$$

for $x, y \in U$ and scalars α, β, nor can we compute integrals such as

$$\int_U s(x)\, \mathrm{d}x. \tag{4.28}$$

In particular, we cannot naively use sections of vector bundles as input to a run-of-the-mill neural network. The reason is that $s(x) \in E_x$ and $s(y) \in E_y$ are elements of different vector spaces when $x \neq y$, so their sum and their integral are undefined.

Trivial bundles have a way around this problem: Given a global trivialization

$$\phi : E \to \mathcal{M} \times V, \tag{4.29}$$

and a local section $s : U \to E$ with domain $U \subseteq \mathcal{M}$, there is a unique function $v : U \to V$ with the same level of regularity as the section s and which satisfies

$$\phi(s(x)) = (x, v(x)), \qquad x \in U. \tag{4.30}$$

Indeed, $v(x)$ is simply the projection of $\phi(s(x))$ onto the second component. We may thus replace any section s with its induced function v, which can be summed and integrated without issue. The same method can also be used for nontrivial bundles thanks to the existence of local trivializations

$$\phi_{\tilde{U}} : E \to \tilde{U} \times V, \qquad \tilde{U} \subset \mathcal{M}, \tag{4.31}$$

but it only works for sufficiently local sections $s : U \to E$ because it requires that $U \subseteq \tilde{U}$ for some local trivialization $\phi_{\tilde{U}}$. Let us discuss two alternatives to this method that do not have the same drawback.

First suppose that $\pi : P \to \mathcal{M}$ is a principal bundle and that $E = P \times_\rho V$ is an associated vector bundle. For any local section $s : U \to E$, there exists a unique *feature map*

$$f : \pi^{-1}(U) \to V, \tag{4.32}$$

with the same regularity as the section s and which satisfies

$$f(p \triangleleft k) = \rho(k)^{-1} f(p), \qquad p \in \pi^{-1}(U), k \in K. \tag{4.33}$$

The relationship between sections s and feature maps f is summarized by

$$s(x) = [p, f(p)], \quad \text{for any} \quad p \in P_x. \tag{4.34}$$

This equation is well-defined thanks to (4.33). Indeed, a short calculation shows that $[p \triangleleft k, f(p \triangleleft k)] = [p, f(p)]$ for all $k \in K$, so the equivalence class depends only on the basepoint $x = \pi(p)$ and not on any specific element $p \in P_x$.

So the problem with adding or integrating sections vanishes if we replace them with feature maps. This is a common approach for G-equivariant convolutional neural networks, which correspond to the special case when the base manifold $\mathcal{M} \simeq G/K$ is a homogeneous space with symmetry group G, and the principal bundle is given by the quotient map $q : G \to G/K$. See Cohen et al. [2018b] for an introduction and **Paper I** for a formal treatment. Homogeneous spaces and homogeneous vector bundles are also discussed in Section 4.4.

The second method is to connect different fibers to each other and then use the connection to transport elements between fibers. This method makes it possible to define, for instance, integrals of the form

$$\int_U T_{x \to x_0}(s(x)) \, \mathrm{d}x, \tag{4.35}$$

where

$$T_{x \to x_0} : E_x \to E_{x_0}, \tag{4.36}$$

is a linear operator transporting elements of the fiber at $x \in U$ to the fiber at a fixed basepoint $x_0 \in \mathcal{M}$. The transport operator (4.36) should ideally perform transportation, along a curve from x to x_0, while modifying the transported object as little as possible. This rough idea is made precise by the concept of *parallel transport*, the details of which depend on how fibers are connected.

Connections and parallel transport can be defined on any smooth fiber bundle but we will restrict attention to *principal (Ehresmann) connections* on principal bundles $\pi : P \to \mathcal{M}$. These will allow us to perform parallel transport on the principal bundle as well as on any associated vector bundle.

One can visualize curves in P as moving in two orthogonal directions: vertically and horizontally. The vertical direction is defined by the projection $\pi : P \to \mathcal{M}$ down to the base manifold. A curve $\gamma : [0, 1] \to P$ is thus considered vertical if its projection $\pi \circ \gamma$ is constant - an idea that extends to *vertical tangent vectors*.

Definition 13. Let $\pi : P \to \mathcal{M}$ be a principal bundle over $\mathcal{M}$ and consider the differential of the projection at a point $p \in P$:

$$\mathrm{d}\pi_p : T_pP \to T_{\pi(p)}\mathcal{M}. \tag{4.37}$$

A *vertical tangent vector* at p is a tangent vector $X_p \in T_pP$ such that

$$\mathrm{d}\pi_p(X_p) = 0. \tag{4.38}$$

The vertical tangent vectors at p form a linear subspace $V_pP := \ker \mathrm{d}\pi_p \subset T_pP$ known as the *vertical tangent space*.

In contrast, the horizontal direction is thought of as moving between different fibers. Parallel transport maps transport objects along horizontal curves or, more precisely, in the direction of *horizontal tangent vectors*. Whereas the vertical direction is uniquely defined by the projection π, however, there are many different choices when it comes to the horizontal direction.

Definition 14. A *horizontal tangent space* is any subspace $H_pP \subset T_pP$ such that

$$T_pP = V_pP \oplus H_pP. \tag{4.39}$$

An *Ehresmann connection* is a smooth assignment $p \mapsto H_pP$ of horizontal tangent spaces and can be defined for any smooth fiber bundle. Since we are focusing on principal bundles, however, we would ideally want an Ehresmann connection that is compatible with the right-action

$$R_k : P \to P, \qquad p \mapsto p \triangleleft k, \tag{4.40}$$

of the structure group K. Compatibility with this action will be important for parallel transport on associated bundles to be well-defined.

Definition 15. Let $\pi : P \to \mathcal{M}$ be a principal bundle with structure group K. A *principal (Ehresmann) connection* on P is a smooth assignment $p \mapsto H_pP$ of horizontal tangent spaces such that

$$(\mathrm{d}R_k)_p\big(H_pP\big) = H_{p\triangleleft k}P, \tag{4.41}$$

for all $p \in P$, $k \in K$.

Example 8. Any principal bundle $\pi : P \to \mathcal{M}$ satisfies

$$(\mathrm{d}\pi)_{p\triangleleft k}(T_{p\triangleleft k}P) = \mathrm{d}(\pi \circ R_k)_p\,(T_pP) = (\mathrm{d}\pi)_p(T_pP), \tag{4.42}$$

as moving within a fiber does not affect the projection ($\pi \circ R_k = \pi$). In particular, a tangent vector $X_p \in T_pP$ vanishes under the projection $(\mathrm{d}\pi)_p$ if and only if $(\mathrm{d}R_k)_p(X_p) \in T_{p\triangleleft k}P$ vanishes under $(\mathrm{d}\pi)_{p\triangleleft k}$. This proves that

$$(\mathrm{d}R_k)_p(V_pP) = V_{p\triangleleft k}P. \tag{4.43}$$

Now suppose that P is equipped with a Riemannian metric g that is compatible with the fiberwise action $p \mapsto p \triangleleft k$, in the sense that the isomorphism

$$(\mathrm{d}R_k)_p : T_pP \to T_{p\triangleleft k}P, \tag{4.44}$$

is unitary. Then $(\mathrm{d}R_k)_p$ preserves orthogonality between tangent vectors and we can therefore obtain a principal connection on P by defining the horizontal tangent spaces to be the orthogonal complements

$$H_pP := (V_pP)^\perp, \tag{4.45}$$

with respect to the metric g. ■

Example 9. Let G be a connected matrix Lie group and K a compact subgroup. If we identify the Lie algebra $\mathfrak{g}$ with the tangent space T_eG at the identity, then the tangent spaces on G are given by left-translations

$$T_gG = (\mathrm{d}L_g)_e(\mathfrak{g}), \qquad g \in G. \tag{4.46}$$

The subgroup K is the characteristic fiber of the principal bundle $q : G \to G/K$ and thereby determines the vertical direction. It is not surprising, then, that the vertical tangent spaces are given by left-translations

$$V_gG = (\mathrm{d}L_g)_e(\mathfrak{k}), \qquad g \in G, \tag{4.47}$$

of its Lie algebra $\mathfrak{k}$. See Theorem 14 for details. Theorem 15 and the preceding discussion identifies a canonical principal connection on $q : G \to G/K$ induced by left-translations. ■

Parallel transport in principal bundles always follows a curve $\gamma : [0,1] \to \mathcal{M}$ in the base maifold. Elements $p \in P_{\gamma(0)}$ move in the direction of horizontal tangent vectors between the fibers $P_{\gamma(t)}$ above the curve γ until they reach $P_{\gamma(1)}$. That is, p is transported along a horizontal curve in P that lies vertically above γ.

Definition 16. Let $\pi : P \to \mathcal{M}$ be a principal bundle equipped with a principal connection. A curve $\gamma^{\uparrow} : [0,1] \to P$ is a *horizontal lift* of a curve $\gamma : [0,1] \to \mathcal{M}$ if

1. The horizontal lift $\gamma^{\uparrow}$ projects down to γ, that is, $\pi \circ \gamma^{\uparrow} = \gamma$.
2. The velocity vector $\dot{\gamma}^{\uparrow}(t)$ is a horizontal tangent vector for each $t \in [0,1]$.

Seeing as we want to use horizontally lifted curves to perform parallel transport in P, it is important to know whether any curve in $\mathcal{M}$ can be horizontally lifted.

Theorem 8 [Hamilton, 2017, Theorem 5.8.2]**.** *Let $\gamma : [0,1] \to \mathcal{M}$ be a curve such that $\gamma(0) = x$ and let $p \in P_x$. Then there exists a unique horizontal lift $\gamma_p^{\uparrow} : [0,1] \to P$ such that $\gamma_p^{\uparrow}(0) = p$.*

We now have everyting needed to define parallel transport in principal bundles.

Definition 17. Let $\pi : P \to \mathcal{M}$ be a principal bundle equipped with a principal connection, and let $\gamma : [0,1] \to \mathcal{M}$ be a curve in $\mathcal{M}$ with $\gamma(0) = x$ and $\gamma(1) = y$. Then the map

$$\Pi_\gamma : P_x \to P_y, \qquad p \mapsto \gamma_p^{\uparrow}(1), \tag{4.48}$$

is called the *parallel transport map in P* along the curve γ.

To summarize, parallel transport in principal bundles uses horizontally lifted curves to transport elements p between fibers. Different choices of principal connection result in different notions of horizontality, so the parallel transport of an element p along a curve γ is not uniquely defined. Moreover, parallel transport is path dependent. A different curve $\tilde{\gamma} : [0,1] \to \mathcal{M}$ with the same endpoints $\tilde{\gamma}(0) = x$ and $\tilde{\gamma}(1) = y$ does not, in general, give rise to the same parallel transport map as γ.

What about associated bundles? After all, our reason for discussing principal connections and parallel transport is because we want the ability to sum and integrate data points in neural network layers. Fortunately, most of the legwork has already been completed.

Definition 18. Let $\pi : P \to \mathcal{M}$ be a principal bundle equipped with a principal connection and let $E = P \times_\rho V_\rho$ be an associated bundle with fibers E_x. Further let $\gamma : [0,1] \to \mathcal{M}$ be a curve in $\mathcal{M}$ with endpoints $\gamma(0) = x$ and $\gamma(1) = y$. Then the map

$$T_\gamma : E_x \to E_y, \qquad [p, v] \mapsto [\Pi_\gamma(p), v], \tag{4.49}$$

is called the *parallel transport map in E along the curve γ*.

Parallel transport maps T_γ in an associated bundle are well-defined if and only if the parallel transport maps Π_γ in the principal bundle are *gauge equivariant*:

$$\Pi_\gamma(p \lhd k) = \Pi_\gamma(p) \lhd k, \tag{4.50}$$

for all $p \in P$, $k \in K$. This is because gauge equivariance implies that

$$\begin{aligned} T_\gamma([p, v]) = [\Pi_\gamma(p), v] &= [\Pi_\gamma(p) \lhd k, \rho(k)^{-1} v] \\ &= [\Pi_\gamma(p \lhd k), \rho(k)^{-1} v] = T_\gamma([p \lhd k, \rho(k)^{-1} v]). \end{aligned} \tag{4.51}$$

Gauge equivariance of Π_γ follows from the compatibility criteria (4.41) between horizontal subspaces and the right-action $p \mapsto p \lhd k$ of the structure group. See for example [Hamilton, 2017, Theorem 5.8.4].

4.3 Yang-Mills theory

A discussion on Yang-Mills theory serves two purposes: It offers a more explicit picture of the parallel transport maps discussed above, and it provides the link between neural networks and physical gauge theory in **Paper III**.

Connection 1-forms and Yang-Mills fields

Principal connections are well-motivated on an abstract level but their global and coordinate-free nature is not always practical to work with. Fortunately, any principal connection defines a global *connection 1-form* that can be combined with local gauges to form *Yang-Mills fields*.

Yang-Mills fields are inherently local and gauge-dependent objects that not only allow for connections to be constructed from the bottom up, they also make it possible to compute parallel transport maps by solving a differential equation. In this part, we summarize the main ideas of connection 1-forms and Yang-Mills fields.

Because the structure group K of any (smooth) principal bundle is a Lie group, it has an associated Lie algebra $\mathfrak{k}$. The two are linked, e.g., through an exponential function

$$\exp : \mathfrak{k} \to K, \tag{4.52}$$

which for matrix Lie groups coincides with the matrix exponential. Hamilton [2017] gives an excellent account of Lie algebras and of the exponential function, which are used extensively throughout the rest of this thesis.

Definition 19. For each $A \in \mathfrak{k}$, the vector field $X^A : P \to TP$ given by

$$X_p^A = \frac{\mathrm{d}}{\mathrm{d}t}\Big(p \triangleleft \exp(tA)\Big)\Big|_{t=0}, \qquad p \in P, \tag{4.53}$$

is called the *fundamental vector field* associated to A.

Observe that the curve $p \triangleleft \exp(tA)$ lives inside a single fiber $P_{\pi(p)}$ for all times t and its projection under π is therefore constant. It follows that $(\mathrm{d}\pi)_p(X_p^A) = 0$, hence X_p^A is a vertical tangent vector.

Fundamental vector fields are important objects in differential geometry but in this thesis, they only have instrumental value through the functions

$$\begin{aligned} i_p : \mathfrak{k} &\to V_pP \\ A &\mapsto X_p^A \end{aligned}, \tag{4.54}$$

that map any given element $A \in \mathfrak{k}$ to the vertical tangent vector X_p^A at $p \in P$. This function is a bijection with inverse $i_p^{-1} : V_pP \to \mathfrak{k}$. Consequently, if the principal bundle is equipped with a principal connection that decomposes each tangent space into a direct sum

$$T_pP = V_pP \oplus H_pP, \tag{4.55}$$

then we can define a function $\omega_p : T_pP \to \mathfrak{k}$ that decomposes tangent vectors $X_p \in T_pP$ into a sum of vertical and horizontal tangent vectors,

$$X_p = \mathrm{ver}(X_p) + \mathrm{hor}(X_p), \tag{4.56}$$

and applies the inverse i_p^{-1} to the vertical term. That is,

$$\omega_p(X_p) = i_p^{-1}(\mathrm{ver}(X_p)). \tag{4.57}$$

The function $\omega_p : T_pP \to \mathfrak{k}$ is called a *Lie-algebra-valued connection 1-form*.

Definition 20. Let $\pi : P \to \mathcal{M}$ be a principal bundle equipped with a principal connection. Let ω_p be the induced Lie-algebra-valued connection 1-forms (4.57). Then the mapping

$$\omega : p \mapsto \omega_p, \tag{4.58}$$

is called the *connection 1-form* of the principal connection.

Example 10. Recall from Example 9 that when G is a connected matrix Lie group and K is a compact subgroup, the vertical tangent spaces of the principal bundle $q : G \to G/K$ are the left-translations of the Lie algebra $\mathfrak{k}$,

$$V_g G = (\mathrm{d}L_g)_e(\mathfrak{k}). \tag{4.59}$$

The fundamental vector field (4.53) associated to $A \in \mathfrak{k}$ is given by

$$X_g^A = \frac{\mathrm{d}}{\mathrm{d}t}\Big(g \exp(tA)\Big)\Big|_{t=0} = (\mathrm{d}L_g)_e\,(A)\,, \tag{4.60}$$

hence $i_g = (\mathrm{d}L_g)_e$. Now consider the special case $K = G$ in which the principal bundle $q : G \to G/G$ consists of a fiber G attached at a single point $x = q(G)$. Since any curve γ in G projects down to the constant curve $(q \circ \gamma)(t) = x$, the tangent vector $\dot{\gamma}(0)$ vanishes under the differential $\mathrm{d}q$ and all tangent vectors on G are therefore vertical. Indeed, $K = G$ implies that $\mathfrak{k} = \mathfrak{g}$ and so

$$V_g G = (\mathrm{d}L_g)_e(\mathfrak{k}) = (\mathrm{d}L_g)_e(\mathfrak{g}) = T_g G. \tag{4.61}$$

In particular, the decomposition (4.56) of $X_g \in T_g G$ into a sum of vertical and horizontal terms is uniquely defined as $\mathrm{ver}(X_g) = X_g$ and $\mathrm{hor}(X_g) = 0$. This defines the unique principal connection on the principal bundle $q : G \to G/G$. The resulting connection 1-form ω given by

$$\omega_g(X_g) = i_g^{-1}(\mathrm{ver}(X_g)) = (\mathrm{d}L_{g^{-1}})_g(X_g), \tag{4.62}$$

is called the *Maurer-Cartan form* on G. ■

Remark 7. The connection 1-form on a principal bundle depends on the choice of principal connection, via the decomposition $X = \mathrm{ver}(X) + \mathrm{hor}(X)$ of tangent vectors into vertical and horizontal terms. Conversely, it is possible to define connection 1-forms ω first and then obtain a principal connection through the relation $H_p P = \ker \omega_p$. See [Hamilton, 2017, §5.2] for details.

Everything we have discussed so far regarding connections, parallel transport, and connection 1-forms has been presented without the use of local coordinates. The coordinate-free approach is powerful but does not give a complete picture. For example, the parallel transport maps (4.48) and (4.49) are highly abstract and difficult to work with in practice, without more detailed information of how horizontal lifts are constructed. Neither have we utilized the *gauge* aspect of mathematical gauge theory. This is about to change.

Let $x^\mu : U \to \mathbb{R}$, for $\mu = 1, \ldots, d$, be local coordinates on an open subset $U \subseteq \mathcal{M}$ of the d-dimensional base manifold. Further assume that $\sigma : U \to P$ is a gauge; a local section of the principal bundle.

Definition 21. Let $\pi : P \to \mathcal{M}$ be a principal bundle equipped with a principal connection and let ω be the resulting connection 1-form. Then the pullback $\sigma^*\omega$ of ω along a local gauge $\sigma : U \to P$,

$$(\sigma^*\omega)_x : T_x\mathcal{M} \to \mathfrak{k}, \qquad x \in U, \tag{4.63}$$

is known as a *Yang-Mills field*.

It should be noted that the Lie-algebra-valued connection 1-forms $\omega_p : T_pP \to \mathfrak{k}$ are linear, and the pullback preserves linearity. Consequently, if we use local coordinates x^μ to write tangent vectors on $\mathcal{M}$ in the coordinate basis,

$$X_x = X_x^\mu \partial_\mu \in T_x\mathcal{M}, \qquad x \in U, \tag{4.64}$$

then the Yang-Mills field satisfies

$$\begin{aligned}(\sigma^*\omega)_x(X_x) &= (\sigma^*\omega)_x\left(X_x^\mu \partial_\mu\right) \\ &= X_x^\mu (\sigma^*\omega)_x\left(\partial_\mu\right) = A_\mu(x)\,\mathrm{d}x^\mu(X_x).\end{aligned} \tag{4.65}$$

Here, we have used that $X_x^\mu = \mathrm{d}x^\mu(X_x)$, and defined $A_\mu(x) = (\sigma^*\omega)_x\left(\partial_\mu\right)$.

Theorem 9. *In local coordinates, the Yang-Mills field $\sigma^*\omega$ is of the form*

$$\sigma^*\omega = A_\mu\,\mathrm{d}x^\mu, \tag{4.66}$$

for a set of functions $A_\mu : U \to \mathfrak{k}$ known as gauge fields.

Yang-Mills fields and connection 1-forms can therefore be constructed locally by specifying a set of gauge fields A_μ. Given a basis for the Lie algebra, it even suffices to specify the component functions

$$A_\mu^i : U \to \mathbb{R}, \qquad \begin{array}{c} i = 1, \ldots, \dim \mathfrak{k} \\ \mu = 1, \ldots, \dim \mathcal{M} \end{array}. \tag{4.67}$$

In summary, every principal connection on $\pi : P \to \mathcal{M}$ determines a unique connection 1-form ω and vice versa. The connection 1-form is globally defined, gauge independent, and coordinate free. In contrast, the Yang-Mills field $\sigma^*\omega$ is inherently gauge-dependent and its expansion (4.66) in terms of gauge fields also depends on the choice of local coordinates.

Path-ordered exponentials and parallel transport

In this part, we discuss an explicit formula for parallel transport on principal bundles, under the additional assumption that K is a matrix Lie group.

Let $\gamma : [0,1] \to \mathcal{M}$ be a curve in the base space and recall how elements p of the initial fiber $P_{\gamma(0)}$ are parallel transported: by lifting γ to the unique horizontal curve $\gamma_p^{\uparrow} : [0,1] \to P$ satisfying $\gamma_p^{\uparrow}(0) = p$ and following this curve to the end,

$$\Pi_\gamma : P_{\gamma(0)} \to P_{\gamma(1)}, \qquad p \mapsto \gamma_p^{\uparrow}(1). \tag{4.68}$$

Instead of constructing the horizontal lift from scratch, suppose that we know how to find a curve $\gamma^* : [0,1] \to P$ such that $\gamma = \pi \circ \gamma^*$ but that is not horizontal. This is often much easier than finding $\gamma_p^{\uparrow}$ specifically. Not only because γ^* has to satisfy one less constraint than the horizontal lift, but because of the nature of that constraint: γ^* need not depend on the principal connection.

Any such curve γ^* can be modified to obtain the horizontal lift. This is due to the existence of a unique element $k(t) \in K$ for each $t \in [0,1]$ such that

$$\gamma_p^{\uparrow}(t) = \gamma^*(t) \triangleleft k(t). \tag{4.69}$$

As $\gamma_p^{\uparrow}(t)$ depends on the principal connection, so does $k(t)$. In fact, this function can be shown to satisfy the initial value problem[1]

$$\begin{cases} \dot{k}(t) = -\omega_{\gamma^*(t)}(\dot{\gamma}^*(t))k(t) \\ k(0) = k_0 \end{cases}, \tag{4.70}$$

see for example [Schuller, 2016, Corollary 23.3]. One way to arrive at this initial value problem is to observe that γ^* would have coincided with the horizontal lift $\gamma_p^{\uparrow}$ if its velocity $\dot{\gamma}^*(t)$ had been a horizontal tangent vector for each $t \in [0,1]$. This is due to the uniqueness of horizontal lifts. So why not modify the curve γ^* by decomposing its velocities into vertical and horizontal terms,

$$\dot{\gamma}^*(t) = \mathrm{ver}(\dot{\gamma}^*(t)) + \mathrm{hor}(\dot{\gamma}^*(t)), \tag{4.71}$$

and subtracting the vertical term? This is what the factor

$$-\omega_{\gamma^*(t)}(\dot{\gamma}^*(t)) = i_{\gamma^*(t)}^{-1}(-\mathrm{ver}(\dot{\gamma}^*(t))) \in \mathfrak{k}, \tag{4.72}$$

in (4.70) does. Because it also maps $-\mathrm{ver}(\dot{\gamma}^*(t))$ to the Lie algebra, this factor can be exponentiated to produce an element of K, which is more or less how the function $k(t)$ is constructed. Indeed, (4.70) is the differential equation for exponential growth and so we would expect its solution to be of the form

$$k(t) = \exp\left(-\int_0^t \omega_{\gamma^*(s)}(\dot{\gamma}^*(s))\,\mathrm{d}s\right) k_0. \tag{4.73}$$

[1] Here, k_0 is the unique element of K satisfying $p = \gamma_p^{\uparrow}(0) = \gamma^*(0) \triangleleft k_0$.

This is nearly the correct solution, and it would have been correct if the elements of $\mathfrak{k}$ had commuted with each other. They generally do not, however, so the actual solution is an infinite series that is reminiscent of an exponential operator similar to (4.73) but without actually being one. This infinite series can be found in [Hamilton, 2017, §5.10] but because it looks rather messy, it is almost always written in shorthand notation as a *path-ordered* exponential

$$k(t) = \mathcal{P} \exp\left(- \int_0^t \omega_{\gamma^*(s)}(\dot{\gamma}^*(s)) \, \mathrm{d}s \right) k_0. \tag{4.74}$$

Let us see what this function looks like in terms of the gauge fields. To this end, choose local coordinates $x^\mu : U \to \mathbb{R}$ for $\mu = 1, \ldots, d$ and a gauge $\sigma : U \to P$. If the entire curve γ is contained in U, setting $\gamma^* = \sigma \circ \gamma$ yields the formula

$$k(t) = \mathcal{P} \exp\left(- \int_0^t A_\mu(\gamma(s)) \dot{\gamma}^\mu(s) \, \mathrm{d}s \right) k_0, \tag{4.75}$$

where $\dot{\gamma}^\mu(t)$ for $\mu = 1, \ldots, d$ are the components of $\dot{\gamma}(t)$ in the coordinate basis. The parallel transport map along γ is therefore given by

$$\Pi_\gamma(p) = (\sigma \circ \gamma)(1) \triangleleft \mathcal{P} \left(- \int_0^1 A_\mu(\gamma(s)) \dot{\gamma}^\mu(s) \, \mathrm{d}s \right) k_0. \tag{4.76}$$

The Yang-Mills action functional

By this point we are well aware that parallel transport depends on the principal connection. This is sometimes considered a degree of freedom, a flexibility, but other times it raises the question how to identify an appropriate connection, e.g. when the parallel transport maps are determined by other arguments. This is the case in physical gauge theories such as quantum electrodynamics and quantum chromodynamics.

Any principal connection gives rise to curvature, and this curvature can be used to define a Lagrangian on the set of all connections. Integrating the Lagrangian defines an action functional. Thus, one option is to follow the principle of least action and choose the connection that minimizes the action.

Throughout this part, let $\pi : P \to \mathcal{M}$ be a principal K-bundle equipped with a principal connection and associated connection 1-form ω. Furthermore, $[\ ,\ ]_{\mathfrak{k}}$ refers to the Lie bracket in $\mathfrak{k}$.

Definition 22 [Sontz, 2015, Theorem 11.2]**.** The *curvature*, or the *field strength*, of a principal connection is the 2-form acting on vector fields $X, Y : P \to TP$ by

$$F(X, Y) = \mathrm{d}\omega(X, Y) + [\omega(X), \omega(Y)]_{\mathfrak{k}}. \tag{4.77}$$

In contrast to the previous sections, we will primarily work in local coordinates here in order to simplify the comparison with **Paper III**. To this end, choose local coordinates $x^\mu : U \to \mathbb{R}$ for $\mu = 1, \ldots, d$ and a gauge $\sigma : U \to P$ on an open set $U \subseteq \mathcal{M}$. Recall that the connection 1-form ω defines a Yang-Mills field $\sigma^*\omega$ given by

$$(\sigma^*\omega)_x = \sum_\mu A_\mu(x)\,\mathrm{d}x^\mu, \quad x \in U, \tag{4.78}$$

where we defined the gauge fields $A_\mu(x)$ by the action of the Yang-Mills field on the coordinate basis vectors in $T_x\mathcal{M}$. By precisely the same token, the pullback $(\sigma^*F)_x$ acts on pairs of tangent vectors in $T_x\mathcal{M}$ and a similar argument as for Yang-Mills fields shows that

$$(\sigma^*F)_x = F_{\mu\nu}(x)\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu, \tag{4.79}$$

where $F_{\mu\nu}(x) = (\sigma^*F)_x\,(\partial_\mu, \partial_\nu)$.

Proposition 10 [Hamilton, 2017, Proposition 5.6.2]**.** *The components $F_{\mu\nu}$ satisfy*

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]_\mathfrak{k}. \tag{4.80}$$

Now suppose that the base manifold $\mathcal{M}$ is a pseudo-Riemannian manifold with respect to a metric g, and consider the $d \times d$ matrix function with components

$$g_{\mu\nu} = g(\partial_\mu, \partial_\nu). \tag{4.81}$$

The matrix inverse $g^{\mu\nu}$ of $g_{\mu\nu}$ is used to raise the indices of $F_{\mu\nu}$,

$$F^{\mu\nu} = g^{\mu\rho} g^{\nu\sigma} F_{\rho\sigma}. \tag{4.82}$$

In the next definition, $A = (A_1, \ldots, A_d)$ is shorthand notation for a connection, in terms of the gauge fields A_μ that determine the local curvature $F_{\mu\nu}$.

Definition 23. Let (M, g) be a pseudo-Riemannian manifold and let dvol_g be the induced volume measure on $\mathcal{M}$. The *Yang-Mills Lagrangian* with respect to the connection A is the scalar function

$$\mathcal{L}_{\mathrm{YM}}[A] = -\frac{1}{4}\,\mathrm{tr}\left(\sum_{\mu,\nu} F_{\mu\nu} F^{\mu\nu}\right). \tag{4.83}$$

The integral

$$S_{\mathrm{YM}}[A] = \int_\mathcal{M} \mathcal{L}_{\mathrm{YM}}[A]\,\mathrm{dvol}_g \tag{4.84}$$

is called the *Yang-Mills action functional* with respect to A.

Whereas the gauge fields A_μ and the local curvature $F_{\mu\nu}$ are inherently gauge dependent, the Yang-Mills action functional does not depend on the chosen gauge; it is *gauge invariant*. This is an important property, particularly so in the context of physics because gauge dependent objects cannot be observable. Not that the action functional can be experimentally measured in a laboratory, but it determines the equations of motion for certain physical systems. These are called the *Yang-Mills equations* and give a necessary and sufficient criteria for A to be a critical point of $S_{\mathrm{YM}}[A]$, in the sense that the directional derivative

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} S_{\mathrm{YM}}[A + t\alpha], \tag{4.85}$$

vanishes for each so-called *variation* α. This is a somewhat technical matter that we will not go into further detail on, partly because this thesis has a maximum page limit and there are other things to discuss. More to the point, the Yang-Mills equations are not needed for a full understanding of **Paper III**. We instead refer the curious reader to Hamilton [2017].

4.4 Homogeneous vector bundles

We have previously discussed the problem that sections $s : \mathcal{M} \to E$ of vector bundles take values in different fibers E_x at different points x, which makes them difficult to integrate in convolutional layers. The reason this is a problem is because we use sections to model data points in equivariant neural networks, and convolutional layers are central to equivariance.

As we have discussed, parallel transport offers one solution, another solution being to isomorphically replace data points with feature maps. For GCNNs there exists what at first glance seems like a third solution: *Homogeneous vector bundles* are equipped with a G-action that transports vectors linearly between different fibers. Homogeneous vector bundles are claimed in **Paper I** to form the natural setting for GCNNs, but the bundles themselves are not given much attention. **Paper I** only presents a coordinate-free approach and few concrete examples. Here we explore homogeneous vector bundles in more detail.

Definition 24. Let G be a Lie group and suppose that $\mathcal{M}$ is a smooth manifold equipped with a smooth, transitive[2] left-action

$$G \times \mathcal{M} \to \mathcal{M}, \qquad (g, x) \mapsto gx. \tag{4.86}$$

We then say that $\mathcal{M}$ is a *homogeneous (G-)space* with *(global) symmetry group G*.

[2] A group action is *transitive* if, for each pair of points $x, y \in \mathcal{M}$, there exists at least one group element $g \in G$ such that $y = gx$.

Homogeneous spaces $\mathcal{M}$ with symmetry group G are always diffeomorphic to a quotient space G/K for some closed subgroup K. This can be seen by arbitrarily choosing an *origin* $o \in \mathcal{M}$ and defining K_o as its set of stabilizers

$$K_o = \{k \in G \mid ko = o\}. \tag{4.87}$$

The homogeneous space characterization theorem [Lee, 2013] states that K_o is a closed subgroup of G and that the mapping

$$F : G/K_o \to \mathcal{M}, \qquad F(gK_o) = go, \tag{4.88}$$

is a diffeomorphism for any choice of origin $o \in \mathcal{M}$. For this reason, we make no distinction between homogeneous spaces $\mathcal{M}$ and quotient spaces G/K.

Remark 8. We restrict attention to homogeneous spaces G/K with compact K.

Definition 25. Let $\mathcal{M}$ be a homogeneous space with global symmetry group G. A *homogeneous vector bundle* over $\mathcal{M}$ is a smooth vector bundle $\pi : E \to \mathcal{M}$ that is equipped with a smooth left G-action

$$G \times E \to E, \qquad (g, v) \mapsto g \cdot v, \tag{4.89}$$

satisfying $g \cdot E_x = E_{gx}$ and such that the induced map

$$L_{g,x} : E_x \to E_{gx}, \tag{4.90}$$

is linear for all $x \in \mathcal{M}$, $g \in G$.

Example 11. Let G be a Lie group and let K be a compact subgroup. If (ρ, V) is a representation of the subgroup, then the associated bundle $\pi_\rho : G \times_\rho V \to G/K$ is a homogeneous vector bundle with respect to the action

$$g \cdot [g', v] = [gg', v], \tag{4.91}$$

for all $g, g' \in G$, $v \in V$. ■

Example 12. Let $\mathcal{M}$ be a homogeneous space and let $L_g : \mathcal{M} \to \mathcal{M}$ denote the left-translation operator $L_g(x) = gx$ for each $g \in G$. Because the group action on $\mathcal{M}$ is smooth and transitive, the left-translation operator is a diffeomorphism with inverse $L_{g^{-1}}$. It follows that its differential

$$(\mathrm{d}L_g)_x : T_x\mathcal{M} \to T_{gx}\mathcal{M}, \tag{4.92}$$

is a linear isomorphism for each $x \in \mathcal{M}$ and each $g \in G$. The tangent bundle is thus a homogeneous vector bundle with respect to the action

$$G \times T\mathcal{M} \to T\mathcal{M}, \qquad (g, X_x) \mapsto (\mathrm{d}L_g)_x(X_x). \tag{4.93}$$

The *cotangent bundle* $T^*\mathcal{M}$ is dual to the tangent bundle, in the sense that each fiber $T_x^*\mathcal{M}$ is the dual space of $T_x\mathcal{M}$ for each $x \in \mathcal{M}$. In particular, any local coordinate chart yields a dual coordinate basis

$$(\mathrm{d}x^1)_x, \dots, (\mathrm{d}x^d)_x \in T_x^*\mathcal{M} \tag{4.94}$$

of dual vectors $(\mathrm{d}x^\mu)_x : T_x\mathcal{M} \to \mathbb{R}$ defined by

$$(\mathrm{d}x^\mu)_x\Big((\partial_\nu)_x\Big) = \delta^\mu_\nu, \qquad \mu, \nu = 1, \dots, d, \tag{4.95}$$

where δ^μ_ν is the Kronecker delta. We mention this because taking the adjoint of each linear isomorphism $(\mathrm{d}L_g)_x$ defines an action on $T^*\mathcal{M}$ that makes the cotangent bundle into homogeneous vector bundle when $\mathcal{M}$ is a homogeneous space. By extension, the same is true of any *type* (m, n) *tensor bundle*

$$T^{m,n}(\mathcal{M}) = \underbrace{T\mathcal{M} \otimes \cdots \otimes T\mathcal{M}}_{m} \otimes \underbrace{T^*\mathcal{M} \otimes \cdots \otimes T^*\mathcal{M}}_{n}. \tag{4.96}$$

■

Homogeneous vector bundles for $\mathcal{M} \simeq G/K_o$ are in one-to-one correspondence with bundles $G \times_\rho V$ associated to the principal bundle $q : G \to G/K_o$. To see this, choose any origin $o \in \mathcal{M}$ and any homogeneous vector bundle $\pi : E \to \mathcal{M}$. Observe that (4.90) restricts to a linear operator $L_{k,o} : E_o \to E_o$ satisfying

$$L_{k,o} \circ L_{k',o} = L_{kk',o}. \tag{4.97}$$

for all $k, k' \in K_o$. In particular, $L_{k,o}$ is invertible with inverse

$$L_{k,o}^{-1} = L_{k^{-1},o}, \tag{4.98}$$

which lets us conclude that $L_{k,o}$ is a representation of K_o on the vector space E_o. If we simplify the notation by writing the representation as $\rho(k) = L_{k,o}$, then we have the following lemma.

Lemma 11. *Let E be a homogeneous vector bundle. Then the well-defined mapping*

$$\xi : G \times_\rho E_o \to E, \qquad [g, v] \mapsto L_{g,o}(v), \tag{4.99}$$

is an isomorphism of homogeneous vector bundles.

A proof of this lemma can be found in Wallach [2018].

Transformation properties of tensors

In this part we explore how elements of homogeneous vector bundles transform under left-translations. Recall that we use the word *vector* to mean an element of a vector space, and this definition includes tensors of all types.

Definition 26. Let $\pi : E \to \mathcal{M}$ be a homogeneous vector bundle and denote by $L^2(E)$ the space of square-integrable[3] sections $s : \mathcal{M} \to E$. For $g \in G$, the map

$$\mathcal{L}_g : L^2(E) \to L^2(E), \qquad (\mathcal{L}_g s)(x) = g \cdot s(g^{-1}x), \tag{4.100}$$

is a unitary representation of G known as the *induced representation* on $L^2(E)$.

Remark 9. The induced representation is denoted $\mathrm{ind}_K^G \rho(g)$ in **Papers I-II** and in most standard texts. We have opted for the notation $\mathcal{L}_g$ here to reduce clutter.

Given two homogeneous vector bundles E_1, E_2 over a homogeneous space $\mathcal{M}$, **Papers I-II** define *G-equivariant layers* as bounded linear maps

$$\Phi : L^2(E_1) \to L^2(E_2), \tag{4.101}$$

that intertwine the induced representations:

$$\mathcal{L}_g \circ \Phi = \Phi \circ \mathcal{L}_g. \tag{4.102}$$

Of course, our mission here is not to reiterate the contents of **Papers I-II**. Rather, we aim to complement these papers by writing down transformation properties of tensors and tensor fields.

As a first step, choose local coordinates $x^\mu : U \to \mathbb{R}$ for $\mu = 1, \ldots, d$ on an open subset $U \subseteq \mathcal{M}$. Taking tensor products of the coordinate vector fields

$$\partial_\mu : U \to T\mathcal{M}, \tag{4.103}$$

and of the *covector fields* (or *differential 1-forms*)

$$\mathrm{d}x^\mu : U \to T^*\mathcal{M}, \tag{4.104}$$

defined by (4.95), produces sections

$$\underbrace{\partial_{\mu_1} \otimes \cdots \otimes \partial_{\mu_m} \otimes \mathrm{d}x^{\nu_1} \otimes \cdots \mathrm{d}x^{\nu_n}}_{\boldsymbol{e}_{\boldsymbol{\mu}}^{\boldsymbol{\nu}}} : U \to T^{m,n}(\mathcal{M}), \tag{4.105}$$

[3] See **Paper I** for details.

of the tensor bundle $T^{m,n}(\mathcal{M})$. Letting the multi-indices $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ run over all possible combinations of values $\mu_i, \nu_j = 1, \ldots, d$, the tensors $\boldsymbol{e}^{\boldsymbol{\nu}}_{\boldsymbol{\mu}}(x)$ form a basis in $T^{m,n}_x(\mathcal{M})$ for each $x \in U$.

Any type (m, n) tensor field

$$s : \mathcal{M} \to T^{m,n}(\mathcal{M}), \tag{4.106}$$

can thus be written locally in terms of component functions $s^{\boldsymbol{\mu}}_{\boldsymbol{\nu}} : U \to \mathbb{R}$ as[4]

$$\begin{aligned} s &= s^{\boldsymbol{\mu}}_{\boldsymbol{\nu}} \boldsymbol{e}^{\boldsymbol{\nu}}_{\boldsymbol{\mu}} \\ &= s^{\mu_1,\ldots,\mu_m}_{\nu_1,\ldots,\nu_n}\, \partial_{\mu_1} \otimes \cdots \otimes \partial_{\mu_m} \otimes \mathrm{d}x^{\nu_1} \otimes \cdots \otimes \mathrm{d}x^{\nu_n}. \end{aligned} \tag{4.107}$$

The tensor bundles $T^{m,n}(\mathcal{M})$ are homogeneous vector bundles whenever $\mathcal{M}$ is a homogeneous space, as discussed in Example 12. The linear map

$$L_{g,x} : T^{m,n}_x(\mathcal{M}) \to T^{m,n}_{gx}(\mathcal{M}), \tag{4.108}$$

can be constructed for each $g \in G$ and every $x \in \mathcal{M}$ as the tensor product of m copies of the differential $(\mathrm{d}L_g)_x$ and n copies of its adjoint. Alternatively, one can turn a tensor bundle into an associated bundle and use the left-translation specified by (4.91). Using the notation $\rho(k) = (\mathrm{d}L_k)_o$ for the representation

$$(\mathrm{d}L_k)_o : T_o\mathcal{M} \to T_o\mathcal{M}, \qquad k \in K_o, \tag{4.109}$$

its *dual representation* on $T^*_o\mathcal{M}$ is defined by $\rho^*(k) = \rho(k^{-1})^T$. Direct application of Lemma 11 yields the following result.

Corollary 12. *Let $\mathcal{M}$ be a homogeneous G-space with any choice of origin $o \in \mathcal{M}$. Then the tensor bundle $T^{m,n}(\mathcal{M})$ is isomorphic to the associated bundle*

$$G \times_{\rho^m \otimes (\rho^*)^n} T^{m,n}_o(\mathcal{M}), \tag{4.110}$$

where $\rho^m = \underbrace{\rho \otimes \ldots \otimes \rho}_{m \text{ times}}$ and analogously for $(\rho^*)^n$.

Now set $E = T^{m,n}(\mathcal{M})$. The induced representation $\mathcal{L}_g : L^2(E) \to L^2(E)$ acts on (square-integrable) tensor fields according to (4.100), hence

$$(\mathcal{L}_g s)(gx) = g \cdot s(x) = L_{g,x}\big(s(x)\big). \tag{4.111}$$

[4] Recall that we use the Einstein summation convention whereby the expression $T^{\boldsymbol{\mu}}_{\boldsymbol{\nu}} \boldsymbol{e}^{\boldsymbol{\nu}}_{\boldsymbol{\mu}}$ is summed over all possible combinations of $\boldsymbol{\mu}, \boldsymbol{\nu}$.

Suppose that both $x \in U$ and $gx \in U$, in which case we may expand both $s(x)$ and $(\mathcal{L}_g s)(gx)$ in components according to (4.107). Then,

$$\begin{aligned}(\mathcal{L}_g s)^{\boldsymbol{\mu}}_{\boldsymbol{\nu}}(gx)\, \boldsymbol{e}^{\boldsymbol{\nu}}_{\boldsymbol{\mu}}(gx) = (\mathcal{L}_g s)(gx) &= L_{g,x}(s(x)) \\ &= L_{g,x}\Big(s^{\boldsymbol{\mu}'}_{\boldsymbol{\nu}'}(x)\boldsymbol{e}^{\boldsymbol{\nu}'}_{\boldsymbol{\mu}'}(x)\Big) \\ &= s^{\boldsymbol{\mu}'}_{\boldsymbol{\nu}'}(x) L_{g,x}\Big(\boldsymbol{e}^{\boldsymbol{\nu}'}_{\boldsymbol{\mu}'}(x)\Big).\end{aligned} \tag{4.112}$$

The differential $(\mathrm{d}L_g)_x$ acts on coordinate basis vectors in $T_x\mathcal{M}$ by

$$(\mathrm{d}L_g)_x : (\partial_\mu)_x \mapsto R^\nu_\mu (\partial_\nu)_{gx} \tag{4.113}$$

where R^ν_μ are the components of $(\mathrm{d}L_g)_x\big((\partial_\mu)_x\big)$ in the coordinate basis $(\partial_\nu)_{gx}$. These components form the standard matrix $R = R(g,x)$ for $L_{g,x}$ with respect to the coordinate bases in $T_x\mathcal{M}$ and $T_{gx}\mathcal{M}$. Similarly, the adjoint of $(\mathrm{d}L_g)_x$ acts on dual basis vectors in $T^*_x\mathcal{M}$ by

$$(\mathrm{d}x^\mu)_x \mapsto (R^{-1})^\mu_\nu (\mathrm{d}x^\nu)_{gx}. \tag{4.114}$$

It follows that left-translations act on the tensors $\boldsymbol{e}^{\boldsymbol{\nu}'}_{\boldsymbol{\mu}'}(x) \in T^{m,n}_x(\mathcal{M})$ by

$$L_{g,x}\left(\boldsymbol{e}^{\boldsymbol{\nu}'}_{\boldsymbol{\mu}'}(x)\right) = R^{\boldsymbol{\mu}\boldsymbol{\nu}'}_{\boldsymbol{\mu}'\boldsymbol{\nu}}(g,x)\boldsymbol{e}^{\boldsymbol{\nu}}_{\boldsymbol{\mu}}(gx), \tag{4.115}$$

where

$$R^{\boldsymbol{\mu}\boldsymbol{\nu}'}_{\boldsymbol{\mu}'\boldsymbol{\nu}} = R^{\mu_1}_{\mu'_1} \otimes \cdots \otimes R^{\mu_m}_{\mu'_m} \otimes (R^{-1})^{\nu'_1}_{\nu_1} \otimes \cdots \otimes (R^{-1})^{\nu'_n}_{\nu_n}. \tag{4.116}$$

Combining this with (4.112) yields the relation

$$(\mathcal{L}_g s)^{\boldsymbol{\nu}}_{\boldsymbol{\mu}}(gx)\boldsymbol{e}^{\boldsymbol{\mu}}_{\boldsymbol{\nu}}(gx) = s^{\boldsymbol{\mu}'}_{\boldsymbol{\nu}'}(x) R^{\boldsymbol{\mu}\boldsymbol{\nu}'}_{\boldsymbol{\mu}'\boldsymbol{\nu}}(g,x)\boldsymbol{e}^{\boldsymbol{\mu}}_{\boldsymbol{\nu}}(x). \tag{4.117}$$

In particular, stabilizers $k \in K_o$ cause tensors at the origin to rotate in place:

$$(\mathcal{L}_k s)(o) = (\mathcal{L}_k s)^{\boldsymbol{\nu}}_{\boldsymbol{\mu}}(o)\boldsymbol{e}^{\boldsymbol{\mu}}_{\boldsymbol{\nu}}(o). = s^{\boldsymbol{\mu}'}_{\boldsymbol{\nu}'}(o) R^{\boldsymbol{\mu}\boldsymbol{\nu}'}_{\boldsymbol{\mu}'\boldsymbol{\nu}}(k,o)\boldsymbol{e}^{\boldsymbol{\mu}}_{\boldsymbol{\nu}}(o). \tag{4.118}$$

The canonical connection

In this part we investigate the relationship between left-translations in groups G and parallel transport in principal bundles $q : G \to G/K$. For simplicity, we restrict attention to connected matrix Lie groups G and, as always in this thesis, the subgroup K is assumed compact. The following lemma is standard.

Lemma 13. *The Lie algebra of a matrix Lie group G is given by*

$$\mathfrak{g} = \{\textit{matrices } A \textit{ such that } \exp(At) \in G \textit{ for all times } t \in \mathbb{R}\}, \tag{4.119}$$

and the exponential map $\exp : \mathfrak{g} \to G$ *coincides with the matrix exponential*

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}. \tag{4.120}$$

Remark 10. Seeing as K is a subgroup of G, its Lie algebra can be identified with the closed subspace $\mathfrak{k} \subset \mathfrak{g}$ of matrices $A \in \mathfrak{g}$ such that $\exp(At) \in K$ for all times.

Left-translations

$$L_g : G \to G, \qquad L_g(g') = gg', \tag{4.121}$$

move elements between different fibers of the principal bundle $q : G \to G/K$. It is therefore not surprising that there exists a canonical principal connection relating left-translations to parallel transport. This is not to say, however, that left-translations generally *coincide* with parallel transport under the canonical connection. Parallel transport makes direct use of the bundle structure and, in particular, depends on the choice of subgroup K. It is also path-dependent in general. Left-translations (4.121), on the other hand, only use the Lie group structure in G and do not require a path. The exact relationship between these two concepts is evidently rather subtle. Nevertheless, the concepts *do* coincide in some cases as this discussion will show.

A first step towards identifying the canonical connection on $q : G \to G/K$ is to determine its vertical tangent spaces. Differentiating (4.121) at $g' = e$ yields an isomorphism

$$(\mathrm{d}L_g)_e : \mathfrak{g} \to T_g G, \tag{4.122}$$

that, in particular, maps the Lie algebra $\mathfrak{k}$ to a closed subspace $(\mathrm{d}L_g)_e(\mathfrak{k}) \subset T_g G$.

Theorem 14. *Let G be a connected matrix Lie group and let K be a compact subgroup. The vertical tangent spaces of the principal bundle $q : G \to G/K$ are given by*

$$V_g G = \ker{(\mathrm{d}q)_g} = (\mathrm{d}L_g)_e(\mathfrak{k}). \tag{4.123}$$

Proof. Fix $g \in G$ and select an arbitrary $A \in \mathfrak{g}$ to define the curve

$$\gamma_A(t) = g\exp(At), \qquad t \in \mathbb{R}. \tag{4.124}$$

Then $\gamma_A(0) = g$ and, because $\gamma_A(t)$ is the left-translation of an exponential map, the chain rule states that $\dot{\gamma}_A(0) = (\mathrm{d}L_g)_e(A)$. The isomorphism (4.122) thereby implies that all tangent vectors in $T_g G$ are given as the velocity

$$X_A = (\mathrm{d}L_g)_e(A) = \dot{\gamma}_A(0), \tag{4.125}$$

at time $t = 0$ of a curve γ_A for some $A \in \mathfrak{g}$.

The differential $\mathrm{d}q$ is defined by its action

$$(\mathrm{d}q)_g(X_A)f = \frac{\mathrm{d}}{\mathrm{d}t}(f \circ q \circ \gamma_A)\big|_{t=0}, \tag{4.126}$$

on any smooth function $f : G/K \to \mathbb{R}$. For a tangent vector $X_A \in T_gG$ to lie in the kernel of this differential, (4.126) must vanish for all such functions and this can only happen if $q \circ \gamma_A$ is constant for small times $t \approx 0$. As

$$(q \circ \gamma_A)(t) = q(g \exp(At)) = gq(\exp(At)), \tag{4.127}$$

it follows that $q(\exp(At))$ is constant for small times and $\exp(At)$ therefore lives inside the coset $q(\exp(0)) = q(e) = K$ for $t \approx 0$. But then $\exp(At) \in K$ for *all* times, since K is a subgroup and closed under multiplication. To be precise, let $t \in \mathbb{R}$ be arbitrary and choose a sufficiently large n so that $\exp(A(t/n)) \in K$. Then

$$\exp(At) = \exp(A(t/n))^n \in K. \tag{4.128}$$

To summarize, if $X_A = (\mathrm{d}L_g)_e(A)$ lies in the vertical tangent space $V_gG = \ker \mathrm{d}q_g$ then $\exp(At) \in K$ for all times and so $A \in \mathfrak{k}$. The converse direction is proven in [Schuller, 2016, Lemma 21.1] and we therefore conclude that

$$V_gG = \ker \mathrm{d}q_g = (\mathrm{d}L_g)_e(\mathfrak{k}), \tag{4.129}$$

which was to be proven. □

It is very interesting that the vertical tangent spaces of $q : G \to G/K$ are simply left-translations of the Lie algebra $\mathfrak{k}$. It indicates that if we can find a suitable subspace $\mathfrak{m} \subset \mathfrak{g}$ such that

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}, \tag{4.130}$$

then we can define the horizontal subspaces as left-translations

$$H_gG = (\mathrm{d}L_g)_e(\mathfrak{m}). \tag{4.131}$$

This definition would automatically form an Ehresmann connection since the assignment $g \mapsto H_gG$ is smooth, but it would not necessarily be a principal connection. This is because not all possible choices of $\mathfrak{m}$ are compatible with right-translations in the sense of (4.41). Left- and right-translations commute, so on the one hand we have

$$\begin{aligned}(\mathrm{d}R_k)_g(H_gG) &= (\mathrm{d}R_k)_g \circ (\mathrm{d}L_g)_e(\mathfrak{m}) \\ &= \mathrm{d}\left(R_k \circ L_g\right)_e(\mathfrak{m}) \\ &= \mathrm{d}\left(L_g \circ R_k\right)_e(\mathfrak{m}) \\ &= (\mathrm{d}L_g)_k \circ (\mathrm{d}R_k)_e(\mathfrak{m}),\end{aligned} \tag{4.132}$$

but this is generally not the same thing as

$$H_{gk}G = (\mathrm{d}L_{gk})_e(\mathfrak{m}) = (\mathrm{d}L_g)_k \circ (\mathrm{d}L_k)_e(\mathfrak{m}). \tag{4.133}$$

For (4.131) to define a principal connection we evidently need that

$$\mathfrak{m} = (\mathrm{d}R_{k^{-1}})_k \circ (\mathrm{d}L_k)_e(\mathfrak{m}) = \mathrm{d}(R_{k^{-1}} \circ L_k)_e\,(\mathfrak{m}). \tag{4.134}$$

This can be seen by comparing the right-hand sides of (4.132) and (4.133), and using that all differentials involved are isomorphisms. The function

$$R_{k^{-1}} \circ L_k : G \to G, \qquad h \mapsto khk^{-1}, \tag{4.135}$$

is nothing but the conjugation function

$$\psi_g : G \to G, \qquad h \mapsto ghg^{-1}, \tag{4.136}$$

restricted to elements $g = k$ of the subgroup. Its differential

$$(\mathrm{d}\psi_g)_e : \mathfrak{g} \to \mathfrak{g}, \qquad A \mapsto gAg^{-1}, \tag{4.137}$$

defines the *Adjoint representation* $\mathrm{Ad}_G(g) = (\mathrm{d}\psi_g)_e$ of G on its Lie algebra, and the restriction (4.135) to conjugation with respect to elements $k \in K$ produces a representation $\mathrm{Ad}_G(K)$ of the subgroup.

The subgrup K also has an Adjoint representation Ad_K on its Lie algebra $\mathfrak{k}$ and because this is a closed subspace, $\mathfrak{k} \subset \mathfrak{g}$, and both representations are defined by conjugation, we find that Ad_K is a subrepresentation of $\mathrm{Ad}_G(K)$. Now, K is assumed compact, so there exists an inner product on $\mathfrak{g}$ such that $\mathrm{Ad}_G(K)$ is unitary. The subrepresentation Ad_K is also unitary with this inner product and $\mathrm{Ad}_G(K)$ therefore decomposes into a direct sum of subrepresentations,

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{k}^{\perp}, \tag{4.138}$$

on $\mathfrak{k}$ and its orthogonal complement $\mathfrak{k}^{\perp}$ with respect to the inner product. The orthogonal complement $\mathfrak{m} = \mathfrak{k}^{\perp}$ is then an $\mathrm{Ad}_G(K)$-invariant subspace of $\mathfrak{g}$. That is,

$$\mathfrak{m} = \mathrm{Ad}_G(k)(\mathfrak{m}) = (\mathrm{d}\psi_k)_e(\mathfrak{m}) = \mathrm{d}(R_{k^{-1}} \circ L_k)_e\,(\mathfrak{m}). \tag{4.139}$$

for all $k \in K$. This is precisely the criterion (4.134) needed for (4.131) to define a principal connection on the bundle $q : G \to G/K$.

Remark 11. A homogeneous space G/K is *reductive* if $\mathfrak{g}$ can be decomposed into a direct sum $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$ such that $\mathfrak{m}$ is $\mathrm{Ad}_G(K)$-invariant. We have thus proven that G/K is reductive whenever K is compact.

We summarize our conclusions in the form of a theorem.

Theorem 15. *Let G be a connected matrix Lie group and let K be a compact subgroup. Further let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$ be the reductive decomposition of the Lie algebra $\mathfrak{g}$. Then the horizontal tangent spaces*

$$H_gG = (\mathrm{d}L_g)_e(\mathfrak{m}), \tag{4.140}$$

define a principal connection on $q : G \to G/K$ known as its canonical connection.

The following example illustrates how parallel transport under the canonical connection is related to left-translation.

Example 13. Consider the left-invariant vector field

$$X_A : G \to TG, \qquad X_A(g) = (\mathrm{d}L_g)_e(A), \tag{4.141}$$

generated by an element $A \in \mathfrak{g}$. Let $\gamma^{\uparrow} : \mathbb{R} \to G$ be the *maximal integral curve* of this vector field starting at $\gamma^{\uparrow}(0) = e$. This means that $\gamma^{\uparrow}$ moves in the direction $\dot{\gamma}^{\uparrow}(t) = X_A(\gamma^{\uparrow}(t))$ of the vector field at all times. The exponential map on Lie groups is defined using maximal integral curves [Hamilton, 2017, §1.7] so it is almost by definition that

$$\gamma^{\uparrow}(t) = \exp(At), \qquad t \in \mathbb{R}. \tag{4.142}$$

According to the canonical connection, if $A \in \mathfrak{m}$, the velocities $\dot{\gamma}^{\uparrow}(t) = (\mathrm{d}L_g)_e(A)$ are horizontal tangent vectors for all times and $\gamma^{\uparrow}$ is then a horizontal lift of its projection $q \circ \gamma^{\uparrow} : \mathbb{R} \to G/K$.

More generally, each exponential curve

$$\gamma^{\uparrow}_{g_0}(t) = \exp(At)g_0, \qquad A \in \mathfrak{m}, g_0 \in G, t \in \mathbb{R}, \tag{4.143}$$

is the unique horizontal lift starting at $\gamma^{\uparrow}_{g_0}(0) = g_0$ of its projection $\gamma = q \circ \gamma^{\uparrow}_{g_0}$. To clarify, $\gamma^{\uparrow}_{g_0}$ is always a *lift* of its projection but it is only a *horizontal lift* with respect to the canonical connection. It follows that exponential curves (4.143) perform parallel transport in G if and only if the principal bundle $q : G \to G/K$ is equipped with its canonical connection. Moreover, any two elements g_0, g_0' of the same fiber

$$G_{\gamma(0)} = q^{-1}(\{g_0\}) = q^{-1}(\{g_0'\}), \tag{4.144}$$

yield the same projected curve $\gamma = q \circ \gamma^{\uparrow}_{g_0} = q \circ \gamma^{\uparrow}_{g_0'}$ in the homogeneous space. Parallel transport along this projected curve,

$$\Pi_\gamma : G_{\gamma(0)} \to G_{\gamma(1)}, \qquad g_0 \mapsto \gamma^{\uparrow}_{g_0}(1), \tag{4.145}$$

coincides with left-translation by $\exp(A)$,

$$\Pi_\gamma(g_0) = \exp(A)g_0 = L_{\exp(A)}(g_0). \tag{4.146}$$

■

Let us now consider the consequences of this relation between parallel transport and left-translation for homogeneous vector bundles. Definition 18 states that for any curve $\gamma : \mathbb{R} \to G/K$, the mapping

$$T_\gamma : [g, v] \mapsto [\Pi_\gamma(g), v], \tag{4.147}$$

performs parallel transport along γ in associated bundles $G \times_\rho V$. We also know from Lemma 11 that any homogeneous vector bundle $\pi : E \to \mathcal{M}$ is isomorphic to an associated bundle $G \times_\rho E_o$, where o is an arbitrarily chosen origin in $\mathcal{M}$ and $\rho(k) = L_{k,o}$ is defined through the linear maps $L_{g,x}$ between fibers in E. The inverse of the isomorphism

$$\xi : G \times_\rho E_o \to E, \qquad [g, v] \mapsto L_{g,o}(v), \tag{4.148}$$

is given by

$$\xi^{-1} : E \to G \times_\rho E_o, \qquad v \mapsto [g, L_{g^{-1},\pi(v)}(v)], \tag{4.149}$$

where $g \in G$ is any group element such that $\pi(v) = go \in \mathcal{M}$. We can use this isomorphism to define parallel transport $\mathcal{T}_\gamma : E_{\gamma(0)} \to E_{\gamma(1)}$ in homogeneous vector bundles as maps $\mathcal{T}_\gamma = \xi \circ T_\gamma \circ \xi^{-1}$, see the diagram below.

$$\begin{tikzcd}
E_{\gamma(0)} \arrow[r, dashed, "\mathcal{T}_\gamma"] \arrow[d, "\xi^{-1}"'] & E_{\gamma(1)} \\
G \times_\rho E_o \arrow[r, "T_\gamma"] & G \times_\rho E_o \arrow[u, "\xi"']
\end{tikzcd}$$

For any curve γ, this map acts on elements $v \in E_{\gamma(0)}$ by

$$\begin{aligned}
\mathcal{T}_\gamma(v) = \big(\xi \circ T_\gamma \circ \xi^{-1}\big)(v) &= \big(\xi \circ T_\gamma\big)\big([g, L_{g^{-1},\pi(v)}(v)]\big) \\
&= \xi\big([\Pi_\gamma(g), L_{g^{-1},\pi(v)}(v)]\big) \\
&= L_{\Pi_\gamma(g),o}\left(L_{g^{-1},\pi(v)}(v)\right) .
\end{aligned} \tag{4.150}$$

The subscripts make the final expression look more complicated than it actually is. Any $v \in E_{\gamma(0)}$ is first mapped to the fiber E_o through left-translation by g^{-1}, and is then mapped to the fiber $E_{\gamma(1)}$ through left-translation by $\Pi_\gamma(g)$. Indeed, as $L_{g,x}$ is induced from the action (4.89), we can write (4.150) in a more compact way as

$$\mathcal{T}_\gamma(v) = \Pi_\gamma(g) \cdot g^{-1} \cdot v = \Big(\Pi_\gamma(g) g^{-1}\Big) \cdot v. \tag{4.151}$$

This definition of parallel transport in homogeneous vector bundles works for any principal connection. It depends on the connection through Π_γ.

Returning to the canonical connection in the setting of Example 13, the parallel transport maps $\mathcal{T}_\gamma$ induced from (4.146) reduce to

$$\mathcal{T}_\gamma(v) = \left(\Pi_\gamma(g_0)g_0^{-1}\right) \cdot v = \left(\exp(A)g_0g_0^{-1}\right) \cdot v = \exp(A) \cdot v, \tag{4.152}$$

or in terms of the linear maps $L_{g,x}$,

$$\mathcal{T}_\gamma(v) = L_{\exp(A),\gamma(0)}(v) = L_{\exp(A),\pi(v)}(v). \tag{4.153}$$

The curve $\gamma(t) = (q \circ \gamma_{g_0}^{\uparrow})(t) = q(\exp(At)g_0)$ used in (4.146) was not chosen beforehand, it was defined through a choice of horizontal tangent vector $A \in \mathfrak{m}$ and the starting point g_0 for its horizontal lift $\gamma_{g_0}^{\uparrow}$. Furthermore, the absence of g_0 on the right-hand side of (4.152) means that the starting point is relevant only because it determines the fiber $E_{\gamma(0)} = E_{q(g_0)}$ on which $\mathcal{T}_\gamma$ is defined.

Consequently, the domain of the parallel transport map (4.152) can be extended from the fiber $E_{\gamma(0)}$ to the entire bundle E by varying the starting point g_0. This can also be seen from the right-hand side of (4.153), which is a well-defined linear operator on any fiber E_x and thus also on the entire bundle E.

Theorem 16. *Let G be a connected matrix Lie group and let K be a compact subgroup. Further let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$ be the reductive decomposition of the Lie algebra $\mathfrak{g}$. Given any homogeneous vector bundle $\pi : E \to G/K$ and any horizontal tangent vector $A \in \mathfrak{m}$, the mapping*

$$\mathcal{T}_A : E \to E, \qquad v \mapsto \exp(A) \cdot v, \tag{4.154}$$

performs parallel transport in E with respect to the canonical connection.

It is well-known that the exponential map is surjective for compact, connected Lie groups G. Choosing the trivial subgroup $K = \{e\}$ effectively allows us to identify G with the trivial bundle $q : G \to G/\{e\}$ and its canonical connection can then be considered a connection on G itself. The reductive decomposition reduces to $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m} = \mathfrak{m}$ since $\mathfrak{k} = 0$, implying that any tangent vector on G is a horizontal tangent vector under the canonical connection. The map

$$\exp : \mathfrak{m} \to G, \qquad A \mapsto \exp(A), \tag{4.155}$$

is therefore surjective. When combining this fact with Theorem 16 and (4.146), we obtain the following corollary with which we conclude.

Corollary 17. *Let G be a compact, connected matrix Lie group. Then, left-translations*

$$L_g : G \to G, \qquad g' \mapsto gg', \tag{4.156}$$

perform parallel transport in G under the canonical connection, and so does the action

$$\mathcal{T}_g : E \to E, \qquad v \mapsto g \cdot v, \tag{4.157}$$

in any homogeneous vector bundle $\pi : E \to G$.

References

Erik J Bekkers. B-spline cnns on lie groups. *arXiv preprint arXiv:1909.12057*, 2019.

Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 2017.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build e(n)-equivariant steerable cnns. In *International Conference on Learning Representations*, 2021.

M. Cheng, V. Anagiannis, M. Weiler, P. de Haan, T. Cohen, and M. Welling. Covariance in physics and convolutional neural networks. *arXiv preprint arXiv:1906.02481*, 2019.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*. PMLR, 2016a.

T. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018a.

Taco Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. *arXiv preprint arXiv:1906.02481*, 2018b.

Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

Anton Deitmar and Siegfried Echterhoff. *Principles of harmonic analysis*. Springer, 2014.

Matteo Favoni, Andreas Ipp, David I Müller, and Daniel Schuh. Lattice gauge equivariant convolutional neural networks. *Physical Review Letters*, 2022.

Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 869–877, 2018.

Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images. In *International Conference on Machine Learning*. PMLR, 2022.

Ian Goodfellow, Yoshua. Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016. URL `http://www.deeplearningbook.org`.

Mark J. D. Hamilton. *Mathematical gauge theory*. Springer, 2017.

Dale Husemöller. *Fibre bundles*. Springer, 1966.

Gurtej Kanwar, Michael S Albergo, Denis Boyda, Kyle Cranmer, Daniel C Hackett, Sébastien Racaniere, Danilo Jimenez Rezende, and Phiala E Shanahan. Equivariant flow-based sampling for lattice gauge theory. *Physical Review Letters*, 2020.

Ivan Kolár, Peter W. Michor, and Jan Slovák. *Natural operations in differential geometry*. Springer Science & Business Media, 2013.

Maxime W Lafarge, Erik J Bekkers, Josien PW Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 2021.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 1989.

John M. Lee. *Introduction to smooth manifolds*. Springer, 2013.

D. Luo, G. Carleo, B. Clark, and J. Stokes. Gauge equivariant neural networks for quantum lattice gauge theories. *Bulletin of the American Physical Society*, 2021.

F. Schuller. Lectures on geometrical anatomy of theoretical physics, 2016. URL `https://github.com/sreahw/schuller-geometric`.

Bart MN Smets, Jim Portegies, Erik J Bekkers, and Remco Duits. Pde-based group equivariant convolutional neural networks. *Journal of Mathematical Imaging and Vision*, 2023.

Stephen Bruce Sontz. *Principal Bundles: The Classical Case*. Springer, 2015.

Norman Steenrod. *The topology of fibre bundles*. Princeton University Press, 1960.

Carl Toft, Georg Bökman, and Fredrik Kahl. Azimuthal rotational equivariance in spherical convolutional neural networks. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022.

B.S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2018.

Nolan R. Wallach. *Harmonic analysis on homogeneous spaces*. M. Dekker, 2018.

M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *arXiv preprint arXiv:1807.02547*, 2018.

Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. Coordinate independent convolutional networks–isometry and gauge equivariant convolutions on riemannian manifolds. *arXiv preprint arXiv:2106.06020*, 2021.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 2020.