Continuous normalizing flows on manifolds

by

Luca Falorsi

11426322

2020

36 ECTS

Supervisor: Patrick Forrè, PhD

Assessor: Max Welling
Acknowledgements

I would first like to thank my thesis supervisor Patrick Forré for helping me not to lose my way during these two years.

I would also like to thank Nicola de Cao, Tim R. Davidson, Fabrizio Ambrogi, Pim de Haan, Jonas Köhler, Marco Federici and Luca Simonetto, for the wonderful years spent together in Amsterdam.

Finally, I thank my parents and my brother for their love and support.
Abstract

Normalizing flows are a powerful technique for obtaining reparameterizable samples from complex multimodal distributions. Unfortunately, current approaches are only available for the most basic geometries and fall short when the underlying space has a nontrivial topology, limiting their applicability for most real-world data. Using fundamental ideas from differential geometry and geometric control theory, we describe how the recently introduced Neural ODEs and continuous normalizing flows can be extended to arbitrary smooth manifolds. We propose a general methodology for parameterizing vector fields on these spaces and demonstrate how gradient-based learning can be performed. Additionally, we provide a scalable unbiased estimator for the divergence in this generalized setting. Experiments on a diverse selection of spaces empirically showcase the defined framework’s ability to obtain reparameterizable samples from complex distributions.
Contents

Abstract i

Contents iii

1 Introduction 1
  1.1 Summary .................................................. 2

2 Preliminaries 4
  2.1 Reparameterization trick ................................. 4
  2.2 Normalizing flows ....................................... 6
  2.3 Neural ODEs and continuous normalizing flows .......... 8

3 Densities and measure pushforwards 12
  3.1 Reparameterization trick as measure pushforward .......... 16
    3.1.1 Differential geometric prospective: pullback of inverse 17
    3.1.2 Computing the local volume change .................. 18
    3.1.3 Computing the volume change: practical advice .... 20
    3.1.4 Summary ............................................. 21

4 CNF on Manifolds 23
  4.1 Integral curves and flows ............................... 23
    4.1.1 Vector fields ....................................... 23
    4.1.2 Flows .............................................. 26
CONTENTS

4.2 Measure pushforward induced by the flow .................................. 29
4.2.1 Divergence of a vector field .............................................. 29
4.2.2 Continuity equation on Manifolds ....................................... 32
4.2.3 Continuity equation on characteristic curves ......................... 33

5 Parameterizing vector fields ..................................................... 36
5.0.1 Local frames and global constraints ..................................... 37
5.0.2 Lie groups ........................................................................ 38
5.1 Generators of vector fields ..................................................... 39
5.1.1 Time dependent vector fields ............................................. 42
5.2 Divergence using a generating set .......................................... 42
5.3 Homogeneous spaces ............................................................. 44
5.4 Embedded submanifolds of $\mathbb{R}^m$ .................................... 46
5.4.1 Embedded Riemannian Submanifolds .................................. 48
5.4.2 Generators defined by gradients of laplacian eigenfunctions ........ 49
5.4.3 Isometrically embedded Submanifolds .................................. 50

6 Backpropagation through flows .................................................. 53
6.1 A short primer on symplectic geometry ................................... 53
6.2 Cotangent lift ....................................................................... 56
6.2.1 Using local coordinates .................................................... 57
6.2.2 Using a generating set ..................................................... 58
6.2.3 Using a local frame ......................................................... 59
6.2.4 Parallelizable manifolds .................................................... 61
6.2.5 Cotangent lift on Lie Groups ............................................. 62
6.2.6 Cotangent lift on embedded submanifolds ........................... 63

7 Experiments ................................................................. 66
7.1 General setup .................................................................. 66
| Section                                      | Page |
|----------------------------------------------|------|
| 7.1.1 Density matching on manifolds          | 66   |
| 7.1.2 Mixture of densities                   | 67   |
| 7.1.3 Implementation details                | 67   |
| 7.2 Manifold specific setup and results      | 68   |
| 7.2.1 Special Orthonormal matrices          | 68   |
| 7.2.2 Stiefel Manifold                      | 69   |
| 7.2.3 Hypersphere                           | 69   |
| 7.2.4 Unitary matrices                      | 72   |
| 7.2.5 Special Unitary matrices              | 73   |
| 7.2.6 Positive definite symmetric matrices  | 75   |
| 7.3 Summary of results and comment          | 77   |
| 8 Conclusion                                | 79   |
| A Densities and induced measure             | 81   |
| A.1 volume forms                            | 81   |
| A.2 Densities                               | 83   |
| A.2.1 Induced measure                       | 86   |
| Bibliography                                | 89   |
Chapter 1

Introduction

Normalizing flows (NF) are a family of methods for defining flexible reparameterizable probability distributions on high dimensional data. They accomplish this by mapping a sample from a simple base distribution into a complex one, using a series of invertible mappings that are usually parameterized by invertible neural networks. The probability density of the final distribution is then given by the change of variable formula. As such, normalizing flows allow for fast and efficient sampling as well as density evaluation.

Due to their desirable features, flow based model have been subject to an ever growing interest in the machine learning community since their introduction (Rezende and Mohamed, 2015), and have been successfully used in the context of generative modelling, variational inference and density estimation\textsuperscript{1}, with a growing number of applications in different fields of science. (Noé et al., 2019; Kanwar et al., 2020; Wirnsberger et al., 2020).

As many real word problems in robotics, physics, chemistry, and the earth sciences are naturally defined on spaces with a non-trivial topology, recent work has focused on building probabilistic deep learning frameworks that can work on manifolds different from the Euclidean space (Davidson et al., 2018, 2019; Falorsi et al., 2018, 2019; Pérez Rey et al., 2019; Nagano et al., 2019). For these type of data the possibility of defining complex reparameterizable densities on manifolds through normalizing flows is of central importance. However, as of today there exist few alternatives, mostly limited to the most basic and simple topologies.

The main obstacle for defining normalizing flows on manifolds is that cur-

\textsuperscript{1}See Papamakarios et al. (2019); Kobyzev et al. (2020) for a comprehensive review of NF.
Currently there is no general methodology for parameterizing maps $F : M \to N$ between two manifolds. Neural networks can only accomplish this for the Euclidean space, $\mathbb{R}^n$. In this work we propose to use vector fields on a manifold $M$ as a flexible way to express diffeomorphic maps from the manifold to itself. As vector fields define an infinitesimal displacement on the manifold for every point, they naturally give rise to diffeomorphisms without needing to impose further constraints. Furthermore, vector fields are significantly easier to parameterize using neural architectures, as they form a free module over the ring of functions on the manifold. In doing so we make it possible to define NF on manifolds. Furthermore, there exists decades old research on how to numerically integrate ODEs on manifolds.\footnote{See Hairer et al. (2006) for a review of the main methods.}

Recently, normalizing flows build using differential equations have proven successful in Euclidean space (Chen et al., 2018; Grathwohl et al., 2019) taking advantage of unrestricted neural network architectures. Using ideas from differential geometry and building on the concepts first introduced in Chen et al. (2018), this work continues this line of research by defining a flexible framework for constructing normalizing flows on manifolds that is trivially extendable to any manifold of interest.

\textbf{1.1 Summary}

\textbf{Chapter 2: } We start by reviewing the fundamental Machine Learning concepts on which we build upon in the rest of the thesis: reparameterizable distributions, normalizing flows and neural ODEs. We additionally review the main works that tried to extend these frameworks to manifold setting.

\textbf{Chapter 3 and Appendix A: } Here we take care of carefully describing the mathematical constructs needed for building a coherent theory of reparameterizable distributions on smooth manifolds. We first define volume forms and densities on manifolds, which are the fundamental objects used for measuring volumes and integrating. We then outline how, using the Riesz extension theorem, a smooth nonnegative density uniquely defines a Radon measure on the space, providing a general methodology for instantiating a base measure on a manifold. After this, we illustrate how the concept of reparameterizable distribution can be generalized to abstract measure spaces via measure pushforward. We conclude by describing how a measure pushforward given by a diffeomorphism operates between densities, giving a change
of variables formula to use when building normalizing flows on manifolds.

**Chapter 4:** In this chapter, we first delineate how vector fields and ODEs on a manifold $M$ can be defined in the context of differential geometry, and explain how they give rise to diffeomorphisms on $M$ through their associated flow. We then apply the notions developed in Chapter 3 to the specific flow case, demonstrating how flows defined by vector fields allow defining continuous normalizing flows on manifolds. As in the Euclidean setting, the change of density is given by integrating the divergence on the ODE solutions, where now the divergence is a generalized quantity that depends on the geometry of the space.

**Chapter 5:** We describe a general methodology for parameterizing vector fields on smooth manifolds. Using module theory, we show that all vector fields on a manifold can be obtained by linear combinations of a finite set of generating vector fields, with coefficients given by functions on the manifold. We then outline how generating sets of vector fields can be built on embedded submanifolds of $\mathbb{R}^m$ and homogeneous spaces. We conclude by describing how the divergence can be computed when employing a generating set, and give an unbiased Monte Carlo estimator in this context.

**Chapter 6:** Here we show how the adjoint sensitivity method can be generalized to vector fields on manifolds in the context of geometric control theory (Agrachev and Sachkov, 2013), highlighting important connections with symplectic geometry and the Hamiltonian formalism. Similarly, as in the adjoint method in the Euclidean space, to backpropagate through the flow defined by a vector field we have to solve an ODE in an augmented space. In this case, the ODE is given by a vector field on the cotangent bundle $T^*M$, called cotangent lift, which is a lift of the original vector field on $M$. Additionally, we provide expressions for the cotangent lift on local charts, Lie groups, and embedded submanifolds.

**Chapter 7:** We conclude by showcasing the utility of the defined framework building continuous normalizing flows on a wide array of spaces, including the hypersphere $S^n$, the Lie groups $SO(n), SU(n), U(n)$, Stiefel manifolds $V_m(\mathbb{R}^n)$, and the positive definite symmetric matrices $\text{Sym}^+(n)$. 
Chapter 2

Preliminaries

2.1 Reparameterization trick

In many machine applications, we are interested in finding the parameters of a probability density function that maximizes the expected value of an objective function.

More precisely, suppose we have an objective function \( f \in C^1(\mathbb{R}^n \times \mathbb{R}^k) \) and a family \( \mathcal{P} \) of probability measures:

\[
\mathcal{P} := \{ \rho_\lambda dx = \rho(\cdot, \lambda) \ dx \}_{\lambda \in W} \quad \rho(\cdot, \lambda) \in L^1(\mathbb{R}^n, dx), \forall \lambda \in W \subseteq \mathbb{R}^k \tag{2.1}
\]

which are absolutely continuous with respect to the Lebesgue measure \( dx \), and parameterized by an open set of parameters \( W \subseteq \mathbb{R}^k \). Our objective is to find the parameters \( \lambda \in W \) that maximize the expectation\(^1\):

\[
J(\lambda) := \mathbb{E}_{\rho(x, \lambda)} [f(x, \lambda)] = \int_{\mathbb{R}^n} f(x, \lambda) \rho(x, \lambda) \ dx \tag{2.2}
\]

The reparameterization trick allows to efficiently obtain low variance, unbiased, Monte Carlo estimates of the gradient \( \partial_\lambda J(\lambda) \) and therefore makes possible to tackle the above optimization problem using gradient-based optimization techniques, such as Adam (Kingma and Ba, 2015).

A reparameterization trick for the family \( \mathcal{P} \) consists in a base probability

---

\(^1\)Since the purpose of this section is to illustrate how the need to have the reparameterizable densities arises in machine learning, we will assume (as it is commonly done in the machine learning literature) that all the expectations are well defined and that we can always exchange derivatives and integrals.
density function \( s \in L^1(\mathbb{R}^m, dx) \) independent form the parameters \( \lambda \) and a family of maps:

\[
\mathcal{T} := \{ F(\cdot, \lambda) = F_\lambda(\cdot) : \mathbb{R}^m \to \mathbb{R}^n \}_{\lambda \in W} \quad \text{s.t.} \quad F \in C^1(\mathbb{R}^m \times \mathbb{R}^k, \mathbb{R}^n) \quad (2.3)
\]
such that:

\[
y \sim s(y) dy \quad \Rightarrow \quad \Phi_\lambda(y) \sim \rho_\lambda(y) dy \quad (2.4)
\]
This allows us to rewrite the objective function using the change of variables:

\[
J(\lambda) = \mathbb{E}_{\rho(x, \lambda)} [ f(x, \lambda) ] = \mathbb{E}_{s(\varepsilon)} [ f(F_\lambda(\varepsilon), \lambda) ] \quad (2.5)
\]
Since now we have an expectation with respect to \( s \), which is independent from the parameters \( \lambda \), the gradient of the expectation becomes the expectation of the gradient:

\[
\partial_\lambda J(\lambda) = \partial_\lambda \mathbb{E}_{s(\varepsilon)} [ f(F_\lambda(\varepsilon), \lambda) ] = \mathbb{E}_{s(\varepsilon)} [ \partial_\lambda f(F(\varepsilon, \lambda), \lambda) ] \quad (2.6)
\]
Where \( \partial_\lambda \) indicates the gradient with respect to the parameters \( \lambda \). We can now easily compute:

\[
\partial_\lambda J(\lambda) \approx \frac{1}{h} \sum_{i=1}^{h} \partial_\lambda f \left( F(\varepsilon^{(i)}, \lambda) \right) \quad \text{where} \quad \varepsilon^{(i)} \sim s(\varepsilon) d\varepsilon \quad (2.7)
\]
Where gradients are usually computed using automatic differentiation.

The reparameterization trick was first introduced in the context of variational autoencoders (Kingma and Welling, 2014) for Gaussian random variables:

\[
\mathcal{P}_N = \left\{ \mathcal{N}(x|\lambda_0, \lambda_1) dx = \frac{1}{\lambda_1 \sqrt{2\pi}} e^{\frac{1}{2} \left( \frac{x - \lambda_0}{\lambda_1} \right)^2} dx \right\}_{\lambda \in W} , \quad W = \mathbb{R} \times \mathbb{R}_{>0} \quad (2.8)
\]
In this case the reparameterization trick is particularly simple and uses a standard normal \( s(y) := \mathcal{N}(y|0, 1) \) as base distribution and family of transformations:

\[
T_N = \{ F_\lambda(y) = \lambda_0 + \lambda_1 \cdot y \}_{\lambda \in W} \quad (2.9)
\]
The subsequent research focused on broadening the class of reparameterizable distributions\(^2\) (Rezende and Mohamed, 2015; Naesseth et al., 2017; Figurnov\(^3\)) for practical applications, it is also important that we can easily draw samples \( y \sim s(y) dy \)

\(^2\)In this work the expression reparameterizable distributions indicated a class of probability measures for which a reparameterization trick is defined.

\(^3\)For practical applications, it is also important that we can easily draw samples \( y \sim s(y) dy \)
et al., 2018). The interested reader can consult Mohamed et al. (2020) for a review of the reparameterization trick in the broader context of Monte Carlo gradient estimation.

In recent years there was an increasing interest in defining reparameterizable distributions with support on non-euclidean spaces. Davidson et al. (2018) define a reparameterization trick for the von Mises-Fisher distribution (vMF) on the hypersphere $\mathbb{S}^n$, Cao and Aziz (2020) propose the Power Spherical distribution as a substitute for the vMF, improving on scalability and numerical stability. In the context of hyperbolic geometry Nagano et al. (2019); Mathieu et al. (2019) suggest using a wrapped normal distribution as a reparameterizable density on the hyperbolic space $\mathbb{H}^n$. Falorsi et al. (2018, 2019) define a general reparameterization trick on Lie groups, allowing to transform any reparameterizable density on the euclidean space to a reparameterizable distribution on a Lie group $G$, using the exponential map from the Lie algebra $\mathfrak{g}$ to the group $exp : \mathbb{R}^n \cong \mathfrak{g} \to G$.

2.2 Normalizing flows

Normalizing flows (NF) are a general methodology for defining complex reparameterizable families of probability distributions, allowing for fast and efficient sampling and density estimation.

Instead of starting from a family of probability measures and then finding a reparameterization trick for it, Normalizing Flows model an expressive class of probability distributions by specifying a base probability measure $4\ s\ dx$ and a family $\mathcal{T}$ of transformations:

$$\mathcal{T} := \{ \Phi(\cdot, \lambda) = \Phi_\lambda(\cdot) : \mathbb{R}^n \to \mathbb{R}^n \}_{\lambda \in W} \quad \text{s.t.} \quad \Phi \in C^1(\mathbb{R}^n \times \mathbb{R}^k, \mathbb{R}^n)$$

(2.10)

such that $\Phi_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism $\forall \lambda \in W$. This defines a family $\mathcal{P} := \{ \rho_\lambda dx = \rho(\cdot, \lambda) \ dx \}_{\lambda \in W}$ of reparameterizable distributions with probability density:

$$\rho(\Phi_\lambda(x), \lambda) = s(x) \left| \det D\Phi_\lambda^{-1}(x) \right| \quad \forall \lambda \in W$$

(2.11)

( Where the flexibility of the distributions in the family is determined by the transformations in $\mathcal{T}$. To be able to evaluate the probability density at the samples, we need to be capable of efficiently computing the determinant of the inverse Jacobian in Equation (2.10).)

4Where $s \in L^1(\mathbb{R}^n, dx)$ and $\int s(x)dx = 1$
Initially introduced in Tabak et al. (2010) and Tabak and Turner (2013), Normalizing Flows were popularized by Rezende and Mohamed (2015) in the context of variational inference and by Dinh et al. (2015) for density estimation. Since their introduction, flow-based models have been subject to an increasing interest in the machine learning community, and have been successfully used in the context of generative modelling, variational inference and density estimation, with an increasing number of applications in different fields of science (Noé et al., 2019; Kanwar et al., 2020). See Papamakarios et al. (2019); Kobyzev et al. (2020) for a general review of NF.

As many real-world problems are naturally defined on spaces with non-trivial topology, recently there has been a great interest in building normalizing flows that can work on manifolds different from the Euclidean space. However as of today there exist few alternatives, mostly limited to the most basic and simple topologies.

As already mentioned, the main obstacle for defining normalizing flows on manifolds is that there is no general methodology for parameterizing maps $F : M \to N$ between two manifolds. Neural networks, are by construction designed to operate on the Euclidean space $\mathbb{R}^n$.

Gemici et al. (2016) try to sidestep this by first mapping points from the manifold $M$ to $\mathbb{R}^n$, applying a normalizing flow in this space and then mapping back to $M$. However, when the manifold $M$ has a non-trivial topology there exist no continuous and continuously invertible mapping, i.e. a homeomorphism between $M$ and $\mathbb{R}^n$, such that this method is bound to introduce numerical instabilities in the computation and singularities in the density. Similarly, Falorsi et al. (2019) create a flexible class of distributions on Lie groups by first applying a normalizing flow on the Lie algebra (which is isomorphic to $\mathbb{R}^n$), and then pushing it to the group using the exponential map. While the exponential map is not discontinuous, for some particular groups the resulting density can still present singularities when the density in the Lie algebra is not properly constrained.

Boyda et al. (2020) develop gauge equivariant and conjugation equivariant flows on the Lie group $SU(n)$ of special unitary complex matrices, in the context of flow-based sampling for lattice gauge theories.

Rezende et al. (2020) define normalizing flows for distributions on hyperspheres and tori. This is done by first showing how to define diffeomorphisms from the circle to itself by imposing special constraints. The method is then generalized to products of circles, and extended to the hypersphere $S^n$, by mapping it to $S^1 \times [-1, 1]^n$ and imposing additional constraints to ensure that
the overall map is a well-defined diffeomorphism. Bose et al. (2020) define
normalizing flows on hyperbolic spaces by successfully taking into account
the different geometry, however, the definition of diffeomorphisms in hyper-
bolic space is made easier due to the fact that topologically the hyperbolic
space is homeomorphic to the Euclidean one.

All of the above methods rely on the target manifold containing additional
special structure to formulate correct mappings. On the contrary, the frame-
work outlined in this thesis works for any manifold.

2.3 Neural ODEs and continuous normalizing
flows

Ubiquitous in all fields of science, differential equations are the main mod-
elling tool for many physical processes. Recently Chen et al. (2018) showed
how to effectively integrate Ordinary Differential Equations (ODEs) with
Deep Learning frameworks. In the context of deep learning, the introduc-
tion of ODE based models was initially motivated from the observation that
the popular residual network (ResNet) architecture can be interpreted as
an Euler discretization step of a differential equation (Haber and Ruthotto,
2017).

A vector field $Y$ in $\mathbb{R}^n$ is a function $Y : \mathbb{R}^n \rightarrow \mathbb{R}^n$. When $Y$ fulfils suitable
regularity conditions\(^5\), the solution of the associated ODE ($\dot{x}(t) = Y(x(t))$),
at a fixed time $t_1 \in \mathbb{R}$, defines a diffeomorphic map from $\mathbb{R}^n$ to itself. In
practice a numerical method needs to be employed to obtain the solution of
the differential equation.

The key observation of Chen et al. (2018) is that we can threat the ODE
solution as a black-box. This means that in the backward pass we do not have
to differentiate through the operations performed by the numerical solver.
Instead Chen et al. (2018) propose to use the adjoint sensitivity method
(Pontryagin et al., 1962).

Closely related with the Pontryagin Maximum Principle, one of the most
prominent results in control theory, the adjoint sensitivity method allows to
compute Vector-Jacobian Product (VJP) of the ODE solutions with respect
to its inputs. This is done by simulating the dynamics given by the initial
ODE backwards, augmenting it with a linear differential equation on $\mathbb{R}^n$,

\(^5\)Here we assume that $Y \in C^1(\mathbb{R}^n,\mathbb{R}^n)$ and that the vector field is complete.
called the *adjoint equation*:

\[
\dot{a}(t) = a(t) \left( \frac{\partial Y(x)}{\partial x} \right)_{x(t)}, \quad \text{where} \quad \dot{x}(t) = -Y(x(t)) \tag{2.12}
\]

which intuitively can be thought of as a continuous version of the usual chain rule.

Since the ODE, through its solution, defines a diffeomorphism on \( \mathbb{R}^n \), we can use it to define normalizing flows. Suppose we start from a base probability distribution \( \rho_0 dx \), then the ODE solution at time \( t \in \mathbb{R}^n \) maps it to a distribution \( \rho_t dx \). The resulting normalizing flow is denoted as Continuous Normalizing Flow (CNF). Differentiating the change of variables formula we obtain that also the volume change can be found by solving a linear ODE on \( \mathbb{R} \):

\[
\frac{d}{dt} \rho_t(x(t)) = -\text{div}(Y)(x(t)) \rho_t(x(t)) \tag{2.13}
\]

Where \( \text{div}(Y) \) is the **divergence** of the vector field \( Y \), which corresponds to the trace of the Jacobian:

\[
\text{div}(Y) = \text{tr} \left( \frac{\partial Y(z)}{\partial z} \right) \tag{2.14}
\]

Since the Jacobian evaluation has a computational cost proportional to \( n \) evaluations of \( Y \), Grathwohl et al. (2019) propose to use Hutchinson’s trace estimator to compute an unbiased Monte Carlo estimation of the divergence:

\[
\text{div}(Y) = \mathbb{E}_{\rho(\varepsilon)} \left[ \varepsilon^T \frac{\partial Y(z)}{\partial z} \varepsilon \right], \quad \text{where:} \quad \mathbb{E}_{\rho(\varepsilon)} [\varepsilon] = 0, \quad \mathbb{E}_{\rho(\varepsilon)} [\varepsilon^T \varepsilon] = I \tag{2.15}
\]

Where \( \rho(\varepsilon) \) is a probability distribution on \( \mathbb{R}^n \). Now the quantity \( \varepsilon^T \frac{\partial Y(z)}{\partial z} \) can be efficiently computed as a Vector Jacobian product (VJP).

Recently, competing work (Lou et al., 2020; Mathieu and Nickel, 2020) tried to generalize continuous normalizing flows and neural ODEs to manifold settings. While based on the same intuition, the present work significantly differs in how the basic idea is developed.

First of all the above approaches model the manifold \( M \) as being embedded in some higher dimensional Euclidean space \( \mathbb{R}^m \), and then use the adjoint equation on \( \mathbb{R}^m \) to perform backpropagation. Instead, we argue that the adjoint equation can be generalized to manifold setting, resulting on a differential
equation on the cotangent bundle $T^*M$, which is an object intrinsically defined on the manifold, independently from the parameterization chosen. The embedded approach is then derived as a special case. Moreover, we draw insightful connections with Hamiltonian formalism and symplectic geometry. Having an intrinsically defined quantity allows us to choose more freely the practical parameterization method and the numerical integration scheme. Moreover, its manipulation allows us to derive more efficient formulations of the equation for manifolds with additional structure. We show how this can be done in the case of Lie groups. Additionally, it makes easier the definition of regularization methods.

Another substantial difference lies in how vector fields are parameterized. Both Lou et al. (2020) and Mathieu and Nickel (2020) start with a vector field on $\mathbb{R}^m$, which is later projected on $TM$. While the orthogonal projection on $TM$ is always well defined, and has a simple expression for the hypersphere $S^n$, for more complicated spaces it can be quite complex to compute. For example, if the manifold is described as the 0 level set of a smooth function $f : \mathbb{R}^m \to \mathbb{R}^k$, the orthogonal projection on $TM$ is given by $I - D_f^+ D_f$, where $D_f^+$ is the pseudo-inverse of the Jacobian, which might be expensive to compute and to further differentiate. Instead, we propose to parameterize vector fields using a generating set for the module of vector fields, which reduces the problem of parameterizing vector fields on $M$ to the easier problem of parameterizing functions on $M$. We argue that using a generating set is a more flexible methodology, which allows an easier parameterization on a wider class of manifolds.

For what concerns the volume change computation, both the above methods compute the divergence using local coordinates in a local chart. While by definition every smooth manifold admits a parameterization by smooth charts, many manifolds are defined implicitly, and finding charts can be quite complicated and lead to difficult expressions. Moreover, the divergence computation on a local chart is in general significantly more complex than the divergence computation on $\mathbb{R}^n$, since it involves the determinant of the $n \times n$ symmetric matrix that represents the metric tensor in local coordinates. While for the hypersphere and the hyperbolic space this has a simple expression, in general, this computation has $O(n^3)$ cost. We argue that, when vector fields are parameterized using a generating set formed by vector fields of known divergence, the divergence computation admits an expression with a complexity similar to the euclidean case and does not explicitly involve the Riemannian metric. We then show that this is always possible for all Homogeneous spaces, which always admit a generating set formed by vector fields with zero divergence.
Lou et al. (2020) Proposes to use the normal coordinates, with Riemann or Lie exp map as charts. While normal coordinates are defined in a neighbourhood of every point, and the exp map admits simple formulas in some particular cases, in general, its computational complexity is cubic in the Lie group case, and for the Riemannian case, it involves the solution of an ODE on the tangent bundle $TM$. 
Chapter 3

Densities and measure pushforwards

The overall objective of this chapter is to describe how to properly define normalizing flows on smooth manifolds. Before doing it we need to understand what is the correct mathematical formalism to deal with this problem.

In mathematics probability distributions can properly defined using measure theory. A measure $\mu$ on a set $\mathcal{X}$ gives a way to measure the "volume" of certain subsets of $\mathcal{X}$. If we call the family of subsets that can be measured $\Sigma \subseteq \mathcal{P}(\mathcal{X})$, a measure $\mu$ on $\mathcal{X}$ is then a function $\mu : \Sigma \to [0, +\infty]$. In order for the measure to be well defined, both the collection $\Sigma$ and the function $\mu$ need to satisfy additional properties. In particular $\Sigma$ needs to be a $\sigma$-algebra: a collection of subsets that contains $\emptyset$ and is closed under complement and countable union. We call the triple $(\mathcal{X}, \Sigma, \mu)$ a measure space, a probability distribution on the set $\mathcal{X}$ is then a measure space for which $\mu(\mathcal{X}) = 1$.

The notion of measure is deeply tied with the definition of integral. As a matter of fact, given a measure space $(\mathcal{X}, \Sigma, \mu)$ we can always define the Lebesgue integral $\int f \, d\mu$. The set of functions for which the integral is defined are called measurable functions. They are all the functions $f$ such that $\{x \in \mathcal{X} : f(x) < t\} \in \Sigma$, $\forall t \in \mathbb{R}$. For probability distributions the integral of a measurable function can be also called expectation: $\mathbb{E}_{\mu}[f] := \int f \, d\mu$.

Since we will work on a manifold $M$, we want to have probability distributions that are compatible with the topological structure of the manifold. This

---

1 in general $\Sigma$ it will be smaller that the collection of all subsets of $\mathcal{X}$
2 See Definition 1.3 and 1.18 in Rudin (1987)
means that we want to be able to measure all the open sets of \( M \). In general, if \( \mathcal{X} \) is a topological space, the smallest sigma algebra that contains all open sets of \( \mathcal{X} \) is called the \( \sigma \)-algebra of Borel sets \( \mathcal{B}(\mathcal{X}) \) of \( \mathcal{X} \), and is denoted as \( \mathcal{B}(\mathcal{X}) \).

Using this definition we want \( \Sigma \subseteq \mathcal{B}(M) \). When this happens we call the measure a Borel Measure.

On a topological space \( \mathcal{X} \), we can define Borel measures by giving a linear functional that integrates continuous functions with compact support:
\[
\Lambda f = \int f \, d\mu, \quad \forall f \in C_c(\mathcal{X}).
\]

This important result is known as Riesz representation theorem. In a nutshell, the theorem shows that there always exists unique a Radon measure \( \mu \) on Borel sets such that the corresponding Lebesgue integral coincides with the original integral defined on \( C_c(\mathcal{X}) \):
\[
\Lambda f = \int f \, d\mu, \quad \forall f \in C_c(\mathcal{X}).
\]

Once we have defined a "base" measure \( \mu \) on a space \( \mathcal{X} \), we can automatically define new measures on \( \mathcal{X} \) using measurable functions:
\[
[f \mu](E) := \int_E f \, d\mu, \quad \forall E \in \Sigma,
\]
where \( f \) is measurable positive. The Radon-Nikodym theorem tells us that this set of measures coincides exactly with the measures that are absolutely continuous with respect\( \mu \).

Restricting to absolutely continuous measures is very convenient when designing practical algorithms. In fact, fixed a base measure on the space, the problem of specifying a measure is then reduced to giving a function that integrates to 1. We call this function probability density. Moreover, the KL divergence is finite only for a.c. measures:
\[
\text{KL}(f \mu || g \mu) := \int f \ln(f/g) \, d\mu
\]
where \( f, g \) are measurable positive.

We can also redefine the reparameterization trick in measure theoretic terms. This corresponds to the concept of measure pushforward. Given two measurable spaces \( (\mathcal{X}_1, \Sigma_1, \mu_1) \) and \( (\mathcal{X}_2, \Sigma_2, \mu_2) \) and a measurable map \( F : \mathcal{X} \to \mathcal{Y} \), the push-forward measure is defined as \([F \# \mu_1](E) := \mu_1(F^{-1}(E)), \forall E \in \Sigma_2\).

The reparameterizability property is given by the change of variable formula:
\[
\int f \, d[F \# \mu_1] = \int f \circ F \, d\mu_1.
\]
In general the measure pushforward is completely independent from \( \mu_2 \). Therefore when designing practical algorithms, in or-
order to work with densities, one needs to ensure\footnote{In this case, in order to be able to compute KL divergences one needs to compute the density of $F_\#\mu_1$ w.r.t to $\mu_2$.} that $F_\#\mu_1 \ll \mu_2$.

The considerations of the previous paragraphs give us a guideline on what is the correct formalism that we have to use to extend Bayesian machine learning models to a broader class of spaces. However, even if necessary, they provide little guidance on how to proceed in practice when working with smooth manifolds. That is, in how to choose the base measure for the space, and on how to compute the density of the pushforward measures. To do that we need to use the additional smooth manifold structure of our space. Since every manifold locally behaves like $\mathbb{R}^n$, we will briefly review how the general approach that was described above is instantiated in the Euclidean case, hoping to generalize it to smooth manifolds.

In $\mathbb{R}^n$ the standard choice as a basis density is the Lebesgue measure $dx$. The Lebesgue measure is obtained by applying the Riesz Representation theorem to the Riemann integral: $\int f(x)dx$, where $f \in C_c(\mathbb{R}^n)$. Defined the Lebesgue measure, one can simply work with absolutely continuous measures just by specifying densities. Starting from a probability measure $f dx$, given a diffeomorphism $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ the resulting pushforward measure is still absolutely continuous, and its density is given by the standard change of variables formula in $\mathbb{R}^n$: $F_\#[f dx] = f \circ F^{-1} \left| \det \left( \frac{\partial F(z)}{\partial z} \right) \right| dx$, where $\frac{\partial F(z)}{\partial z}$ is the Jacobian of the inverse transformation. This gives us exactly the change of variable formula used for normalizing flows in $\mathbb{R}^n$.

Trying to generalize what is done in $\mathbb{R}^n$ is not immediately possible. The main reason is that the integral of continuous functions of compact support is in general not well defined on a smooth manifold $M$. One idea could be to define the integral evaluating the function on charts, integrating the local chart expressions, and then combining the results (using the partition of unity). For example, fix a chart $(U, \varphi)$ and a function $f \in C_c(U)$, we could try to define the integral of $f$ as $\int f(\varphi^{-1}(x))dx$, however this expression is not well defined, as it depends on $\varphi$.

In order to be able to integrate functions on smooth manifolds, we need to add additional structure. In the mathematical literature, this structure is given by volume forms. For precise definitions and explanations of these and other objects defined in this section, consult Chapters 14,15,16 of Lee (2013) and Appendix A.

Intuitively, volume forms give us the oriented volume of infinitesimally small patches in the manifold. Consider $M$ a smooth manifold of dimension $m$ with
a volume form $\omega$. Fix a point $q \in M$, and $m$ small independent displacements $(\Delta_i)_{i \in [m]}$. If the displacements are infinitesimally small we can consider them as tangent vectors at $q$: $\Delta_i \in T_q M$. The volume form at $q$, $\omega_q$, is then a multilinear function that gives the oriented volume of the oriented parallelepiped defined by the $\Delta_i$s: $\omega_q(\Delta_1, \cdots, \Delta_m) \in \mathbb{R}$. Summing infinitesimally small volume patches gives us the integral of the volume form. Accordingly to this, volume forms are objects that can be naturally integrated on a smooth manifold. However, in order to have a volume form that is nowhere vanishing, we need the manifold to be orientable. A manifold is orientable if we can assign an orientation\(^{12}\) to each tangent space $T_q M$ in a continuous way on $M$.

Despite these considerations, we are interested in nonoriented volumes in $M$. In order to do this, we need to consider the absolute value of volume forms on $M$. The resulting object is called a density. Contrary to volume forms, on every smooth manifold, there exists a nowhere vanishing smooth density $\mu$. Moreover, all continuous densities can be expressed as $f\mu$, where $f \in C(M)$ is a continuous function. Since, like volume forms, densities can be integrated on manifolds, assigning $f\mu$ to its integral for every $f \in C_c(M)$ defines a positive linear functional on $C_c(M)$. For the Riesz representation theorem, this defines a measure Borel measure on $M$.\(^{13}\) We have then accomplished our objective of defining a Radon measure on a smooth manifold.\(^{14}\) On semi Riemannian manifolds, there is a standard way of defining a nonvanishing density, it is called the semi Riemannian density, and it is defined as the density such that $\mu_q(E_1, \cdots, E_m) = 1$ for every point $q \in M$ and every orthonormal frame $(E_1, \cdots, E_m)$ of $T_q M$.

---

\(^{12}\) An orientation on a vector space is an equivalence class of ordered bases. Two bases have the same orientation if the change of bases matrix has positive determinant.

\(^{13}\) This justifies using the notation $\mu$ for the densities.

\(^{14}\) The careful reader may have noticed that the resulting measure will depend on the initial density $\mu$, however since all the other densities can be obtained by pointwise rescaling $\mu$ by a function, all the possible other measures defined by a density will all be absolutely continuous with respect to the first one. Therefore the initial choice does not influence the class of measures that we can express.
3.1 Reparameterization trick as measure pushforward

In the previous section, we have outlined how to define a Radon\(^{15}\) regular measure on any smooth manifold using a smooth positive density. We have established that we will work with probability measures that are absolutely continuous with respect to the base measure. Moreover, we have seen that this class of measures is independent from the initial choice of the initial density. For precise definitions, proofs, and details on this construction we refer the reader to Appendix A. In the rest of the Chapter we will assume that the reader is familiar with all the concepts and the notation defined there.

Established the theoretical framework in which our objects are defined, we are interested to develop a general reparameterization trick for this class of measures. In order to achieve this, we need to understand what happens under transformations.

In general terms, we could describe the reparameterization trick as a procedure that takes a probability measure on a space \(X\) and defines a new measure on another space \(Y\) by "pushing" the initial measure to \(Y\) using a transformation \(F : X \rightarrow Y\). A correct formalization of this idea is that the transformation defines a pushforward measure:

**Definition 3.0.1.** \(^{16}\) Let \(F : X \rightarrow Y\) be a measurable function. Where \((X, \Sigma_2)\) and \((Y, \Sigma_2)\) are two measurable spaces. Given a measure \(\nu\) on \(X\) we define the pushforward measure \(F_#\nu\) on \(Y\) as:

\[
(F_#\nu)(E) = \nu(F^{-1}(E)), \quad \forall E \in \Sigma_2.
\]

It is straightforward to check that \(F_#\nu\) is indeed a measure. Notice also that \(F_#\nu\) is supported on the range of \(F\).

If \(\nu\) is a probability measure, then, since \(F_#\nu(Y) = \nu(X) = 1\), also \(F_#\nu(Y)\) is a probability measure.

The new measure on \(Y\) is then parameterized using a base measure on \(X\) and a transformation \(F : X \rightarrow Y\). We can evaluate expectations with respect to the pushforward measure by taking samples from the initial measure and transforming them through \(F\), as stated by the following change of variables formula.

\(^{15}\)A radon measure is a measure defined on Borel sets, that is finite on all compact sets, outer regular on all Borel sets, and inner regular on open sets.

\(^{16}\)See Section 7.7 in Teschl (1998)
CHAPTER 3. DENSITIES AND MEASURE PUSHFORWARDS

Theorem 3.1. 17 Let $X$, $Y$, $\nu$, $F$ as in Definition 3.0.1 and let $g : Y \to \mathbb{R}$ be a Borel function. Then the Borel function $g \circ F : X \to \mathbb{R}$ is integrable with respect to $\nu$ if and only if $g$ is integrable with respect to the pushforward measure $F_#\nu$ and in this case the integrals coincide

$$\int_Y g dF_#\nu = \int_X g \circ F d\nu \quad (3.1)$$

If $\nu$ is a probability measure, we can rewrite the previous formula using expectations:

$$E_{F_#\nu} [g] = E_\nu [g \circ F] \quad (3.2)$$

3.1.1 Differential geometric prospective: pullback of inverse

Let $M, N$ be smooth $n$ dimensional manifolds and $\mu_M, \mu_N$ smooth densities respectively on $M$ and $N$. With $\mu_N$ positive. We then have the regular Radon measures $\tilde{\mu}_M$ and $\tilde{\mu}_N$ induced by $\mu_M$ and $\mu_N$. If we take a diffeomorphism $F : M \to N$ this induces a measure $F_#\tilde{\mu}_M$ on $N$. The following theorem tells us that the defined pushforward measure is absolutely continuous with respect to $\mu_N$ and it is exactly the measure induced from the pullback density $(F^{-1})^*\mu_M$

Theorem 3.2. Let $M, N, \mu_M, \mu_N$ as defined above. If $F : M \to N$ is a diffeomorphism, we have that $F_#\tilde{\mu}_M = (F^{-1})^*\mu_M$. This means that pushing forward the measure induced by a density on a manifold, is the same as inducing a measure from the pullback density. In particular this means $F_#\tilde{\mu}_M \ll \mu_N$ and that there exist $f \in C^\infty(N)$ such that $F_#\tilde{\mu}_M = f\mu_N$

Proof. We first prove that $F_#\tilde{\mu}_M$ is Radon. To achieve this, similarly as done in the proof of Proposition A.2.1, it is sufficient to show that $F_#\tilde{\mu}_M$ is finite on compact sets. Let $K \subseteq N$ compact:

$$F_#\tilde{\mu}_M(K) = \int_{F^{-1}(K)} d\tilde{\mu}_M = \tilde{\mu}_M(F^{-1}(K)) < +\infty$$

Where the last inequality follows from the fact that $F^{-1}(K)$ is compact and $\tilde{\mu}_M$ is radon. Then from Corollary 7.6 in Folland (1999) $F_#\tilde{\mu}_M$. Moreover $F_#\tilde{\mu}_M$ and $(F^{-1})^*\mu_M$ extend the same positive Linear functional in $C_c(N)$.

17 Theorem 9.15 (Teschl, 1998)
If fact, let $g \in C_c(N)$, we have:

$$
\int_N gd\bar{F}#\bar{\mu}_M = \int_M g \circ F d\bar{\mu}_M = (3.3)
$$

$$
= \int_M (g \circ F) \cdot \mu_M = \int_N g \cdot (F^{-1})^* \mu_M = \int_N gd(F^{-1})^*\mu_M (3.4)
$$

Where we have used the fact that since $F^{-1}$ is continuous $g \circ F$ has compact support. Then the thesis follows from the Riesz representation theorem. \( \square \)

Let now also $\mu_M$ be smooth positive, and consider absolutely continuous densities with respect to $\bar{\mu}_M$. Let $f : M \to \mathbb{R}$ measurable, it is then easy to verify using the change of variable formula that:

$$
F#f \bar{\mu}_M = (f \circ F)F#\bar{\mu}_M (3.5)
$$

The pushforward will then be absolutely continuous with respect to $\bar{\mu}_N$ and to compute its Radon-Nikodym derivative we need to know the Radon-Nikodym derivative of $F#\bar{\mu}_M$.

The last theorem fundamentally tells us that when working with measures induced by densities on smooth manifolds, and diffeomorphic transformation, the pushforward measure is completely determined by how the underlying density transforms. Therefore, when working within this framework, it is always sufficient to work directly with densities, specifying how they are defined and how they change under transformation. The analytic objects (the measures) are completely determined by the underlying geometric objects (densities, which are sections of vector bundles). This allows us from now on to work directly with the geometric objects, using mainly geometric arguments.

### 3.1.2 Computing the local volume change

Keeping the same notation as before, in the previous subsection we saw that given a density $\mu \in \Gamma^\infty(M, \mathcal{D}M)$ we are interested in computing $(F^{-1})^*\mu \in \Gamma^\infty(N, \mathcal{D}N)$, which is a section of the density bundle $\mathcal{D}N$. For some particular $F$ we can try to compute its expression in closed form, however, in many practical Machine Learning applications, a closed-form expression might be impossible or computationally intractable to compute. On the other hand, when working with Monte Carlo estimates, or when performing maximum likelihood, we are only interested in being able to compute the density value at one point. For example:
Given a point $q \in N$, we want to compute the density at the point:

$$\left[(F^{-1})^*\mu\right](q)$$

Given a sample $q \in M$ with density $\mu(q)$ we want to compute the pullback density at the transformed sample $\left[(F^{-1})^*\mu\right](F(q))$

To perform these operation we lift the diffeomorphic map $F : M \to N$ to a vector bundle isomorphism $F^{D} : D M \to D N$:

**Definition 3.2.1.** Let $(D M, \pi_M, M)$, $(D N, \pi_N, N)$, density bundles with base spaces respectively $M, N$, smooth $n$ dimensional manifolds. Let $F : M \to N$ diffeomorphism, we call the lift of $F$ to the density bundle the vector bundle isomorphism $F^{D} : D M \to D N$, defined in the following way: let $\psi \in D M$

$$F^{D}(\psi)(v_1, \cdots, v_n) := \psi(dF^{-1}v_1, \cdots, dF^{-1}v_n) \quad \forall v_1, \cdots, v_n \in T_F(\pi_M(\psi))N$$

Which is equivalent to:

$$F^{D}(\psi) = \left((F^{-1})^* (\psi)\right)_{\pi_N(\psi)}$$

**Prop 3.2.1.** The following diagram commutes:

$$\begin{array}{ccc}
M & \xrightarrow{F} & N \\
\downarrow{\mu} & & \downarrow{(F^{-1})^*\mu} \\
D M & \xrightarrow{F^{D}} & D N
\end{array} \tag{3.6}$$

**Proof.** It immediately follows from the definition. \qed

The lift augments the function $F$ with the information on how the density changes. Being able to compute this map allows to perform the operations described above:

- Given $q \in N$, its density is given by: $\left[(F^{-1})^*\mu\right](q) = F^{D}(\mu(F^{-1}(q)))$
- Given a sample $q \in M$ with density $\mu(q)$, the density of the transformed sample is: $\left[(F^{-1})^*\mu\right](F(q)) = F^{D}(\mu(q))$

As we described earlier, to parameterize probability measures on $M$ and $N$ we take two smooth positive densities $\mu_M \in \Gamma^\infty(M, D M)$, $\mu_N \in \Gamma^\infty(N, D N)$. Since $\{\mu_M\}$ and $\{\mu_N\}$ are global frames for the line bundles $D M$ and $D N$. 

respectively, they induce the vector bundle isomorphisms: $\mathcal{D}M \cong M \times \mathbb{R}$, and $\mathcal{D}N \cong N \times \mathbb{R}$. With this identification, probability measures are given by $L^1$ functions on $M$ and $N$ respectively and the vector bundle isomorphism $F^\mathcal{D}$ is completely determined by the map

$$J_{F^{-1}} : M \to \mathbb{R}$$

$$q \mapsto J_{F^{-1}}(q) = \frac{F^\mathcal{D}(\mu_M(q))}{\mu_N(F(q))} \quad \forall q \in M$$

Where $J_{F^{-1}}(q) \in \mathbb{R}$ is a (positive) real number\(^\text{18}\) that depends on the transformation $F$ and the point $q$\(^\text{19}\). This means that globally on the manifold we have a positive function $J_{F^{-1}} : M \to \mathbb{R}$ that gives the local volume change. The notation for $J_{F^{-1}}$ is given by the fact that for $\mathbb{R}^n$ it corresponds to the absolute value of the determinant of the Jacobian of the inverse transformation.

In fact, since $F^\mathcal{D}$ is a vector bundle isomorphism, fixed a point $q \in M$, $F^\mathcal{D}$ restricted to the fiber is a linear map $F^\mathcal{D}|_{\mathcal{D}M_q} : \mathcal{D}M_q \to \mathcal{D}N_{F(q)}$. Since $\{\mu_M(q)\}$, and $\{\mu_N(F(q))\}$ form a basis respectively of $\mathcal{D}M_q$ and $\mathcal{D}N_{F(q)}$ the linear map is completely described by its effect on the basis:

$$F|_{\mathcal{D}M_q} : \mathcal{D}M_q \to \mathcal{D}N_{F(q)}$$

$$a_0\mu_M(q) \mapsto F^\mathcal{D}(a_0\mu_M(q)) = a_0F^\mathcal{D}(\mu_M(q)) = a_0J_{F^{-1}}(q) \cdot \mu_N(F(q))$$

The previous diagram 3.6, can then be rewritten as to:

\[ 
\begin{array}{c c c c c c}
M & F & N \\
(id, f) & & & (id, f \cdot \frac{(F^{-1})^*\mu_M}{\mu_N}) \\
M \times \mathbb{R} & (F, \cdot J_{F^{-1}}) & N \times \mathbb{R}
\end{array}
\]

3.1.3 Computing the volume change: practical advice

How to compute the function $J_{F^{-1}}$ in practice depends on the precise data structure used to parameterize the manifold and the function $F$. In this sec-

\(^{18}\)The division is well defined since $\mu_N$ is a positive density. Since $\mu_M$ is also positive and $F$ is a diffeomorphism, the resulting fraction is positive.

\(^{19}\) $J_{F^{-1}}(q)$ depends on the two base positive densities $\mu_M, \mu_N$ as well. But, since we assume them fixed beforehand, we drop the dependency for convenience.
tion, we provide a theoretical discussion and useful formulas that could be used to tackle this problem. We are going to assume that we are able to compute either the differential or the pullback using an automatic differentiation package.

By definition $J_{F^{-1}}$ is the fraction of the two densities $(F^{-1})^* \mu_M$ and $\mu_N$, to obtain its value at a point $q \in M$ we can evaluate the two densities on the same basis of $T_{F(q)} N$ and take the fraction of the results:

$$J_{F^{-1}}(q) = \frac{(F^{-1})^* \mu_M(v_1, \ldots, v_n)}{\mu_N(v_1, \ldots, v_n)} = \frac{\mu_M(dF^{-1}(v_1), \ldots, dF^{-1}(v_n))}{\mu_N(v_1, \ldots, v_n)}$$ (3.10)

$$\forall (v_1, \ldots, v_n) \in GL(T_{F(q)} N)$$ (3.11)

If $(M, g_M), (N, g_N)$ are semi riemannian manifolds and $v_1, \ldots, v_n$ is an orthonormal basis of $T_{F(q)} M$ the expression further simplifies to:

$$J_{F^{-1}}(q) = \mu_M(dF^{-1}(v_1), \ldots, dF^{-1}(v_n)) = | \det g_M(dF^{-1}(v_i), u_j) |$$ (3.12)

$$\forall (v_1, \ldots, v_n) \in O(T_{F(q)} N, g_N), \forall (u_1, \ldots, u_n) \in O(T_q M, g_M)$$ (3.13)

The above expressions all depend on the differential of the inverse map, however, the analytical expression of the inverse might not be available. We can circumvent this by observing that, since $F$ is a diffeomorphism, we can define a basis of $T_{F(q)} N$ by pushing forward a basis of $T_q M$:

$$J_{F^{-1}}(q) = \frac{(F^{-1})^* \mu_M(dF(u_1), \ldots, dF(u_n))}{\mu_N(dF(u_1), \ldots, dF(u_n))} = \frac{\mu_M(u_1, \ldots, u_n)}{\mu_N(dF(u_1), \ldots, dF(u_n))}$$ (3.14)

$$\forall (u_1, \ldots, u_n) \in GL(T_p M)$$ (3.15)

Also in this case we have a simplified expression for the Riemannian case:

$$J_{F^{-1}}(q) = \frac{1}{\mu_N(dF(u_1), \ldots, dF(u_n))} = | \det g_N(dF(u_i), v_j) |^{-1}$$ (3.16)

$$\forall (u_1, \ldots, u_n) \in O(T_q M, g_M), \forall (v_1, \ldots, v_n) \in O(T_{F(q)} N, g_N)$$ (3.17)

3.1.4 Summary

To define a reparameterization trick on a smooth manifold $N$ we can proceed as the following
• Choose a starting space $M$, where $M$ is a smooth manifold diffeomorphic to $N$.

• Define smooth positive densities $\mu_M$, $\mu_N$, on $M$, $N$ respectively.

• Choose a starting distribution by specifying a measurable function $f : M \to \mathbb{R}$. The probability measure is then $f \cdot \mu_M$. We need to be able to sample from it.

• Take a diffeomorphism $F : M \to N$, $F$ will generally depend on some parameters.

• Compute the volume change term $J_{F^{-1}} : M \to \mathbb{R}$ (Alternatively or in addition one can compute the volume change term of the inverse: $J_F : N \to \mathbb{R}$)

• The Radon-Nikodym derivative of the transformed distribution is given by $f(F^{-1}(q)) \cdot J_{F^{-1}}(F^{-1}(q)), \forall q \in N$.

• Given a sample $q \in M$ the value of the Radon-Nikodym derivative of the transformed distribution computed at the mapped point $F(q)$ is given by $f(q) \cdot J_{F^{-1}}(q), \forall q \in M$. 
Chapter 4

Continuous normalizing flows on Manifolds

4.1 Integral curves and flows

4.1.1 Vector fields

A vector field\(^1\) on a smooth manifold \(M\), is defined as a section of the tangent bundle \(TM\).

**Definition 4.0.1.** A *vector field* on a smooth manifold \(M\) is an element of \(\Gamma (M, TM)\), i.e. a section of the tangent bundle \(TM\). More precisely, denoted by \(\pi\) the natural projection \(\pi : TM \rightarrow M\) a vector field is a continuous map:

\[
X : M \rightarrow TM \quad \text{such that} \quad \pi \circ X = \text{id}_M
\] (4.1)

A *smooth vector field* is a vector field for which the map \(X\) is smooth.

**Definition 4.0.2.** A *time dependent vector field* is a continuous map\(^2\)

\(X : \mathbb{R} \times M \rightarrow TM\) where for every time \(t \in \mathbb{R}\), \(X_t := X(t, \cdot) : M \rightarrow TM\) is a vector field on \(M\).

Every vector field \(X\) automatically determines a time dependent vector field \(X'\) simply by setting \(X'_t = X \forall t \in \mathbb{R}\). We will often refer to a time dependent

\(^1\)We refer to (Lee, 2013, ch. 8-9)

\(^2\)More generally, a time dependent vector field should be defined as a map \(J \times M \rightarrow TM\), where \(J \subseteq \mathbb{R}\) is an interval. However since we are only interested in complete vector fields, we restrict to the particular case where \(J = \mathbb{R}\)
vector field as a **nonautonomous vector field** and to a vector field as an **autonomous vector field**. Vector fields allow to generalize the concept of Ordinary Differential Equation to manifolds. We begin by defining **integral curves**:

**Definition 4.0.3.** Let \( X : \mathbb{R} \times M \to TM \) be a vector field and \( J \subseteq \mathbb{R} \) an interval. A differentiable curve \( \gamma : J \to M \) is an **integral curve** of \( X \) if:

\[
\dot{\gamma}(t) = X(t, \gamma(t)) \quad \forall t \in J
\]

If fix a "starting time" \( t_0 \in J \), the point \( \gamma(t_0) \in M \) is called starting point of \( \gamma \). We call **maximal integral curve** an integral curve that cannot be extended to a larger interval \( J \subset I \).

Locally, in a smooth coordinate chart \( (U; x_i) \), we have that \( X|_{\mathbb{R} \times U} = \sum_{i=1}^n f_i \partial x_i \), with \( f_i \in C(\mathbb{R} \times U), \forall i \in [n] \). Equation 4.3 can then be written as

\[
\dot{x}_i(\gamma(t))\partial x_i \bigg|_{\gamma(t)} = f_i(t, \gamma(t))\partial x_i \bigg|_{\gamma(t)} \quad \forall i \in [n]
\]

Which is equivalent to locally solving a system of ODE on the Euclidean space. Under suitable regularity conditions \(^3\) we can then (locally) apply the existence and uniqueness theorem, ensuring that, for a fixed starting point \( q_0 \in M \), for a fixed starting time \( t_0 \in \mathbb{R} \), and for a small enough interval there exist unique an integral curve of \( X \).

**Theorem 4.1.** \(^4\) Let \( X \) be a smooth time dependent vector field, then for any point \( (t_0, p_0) \in \mathbb{R} \times M \) there exist a unique maximal integral curve \( \gamma : J \to M \) with starting point \( q_0 \), and starting time \( t_0 \) denoted by \( \gamma(t; t_0, q_0) \). We call \( \gamma \) a solution of of the Cauchy problem:

\[
\begin{cases}
\dot{\gamma}(t) = X(t, \gamma(t)) \\
\gamma(t_0) = q_0
\end{cases}
\]

Moreover the map \( (t, q) \to \gamma(t; t_0, q) \) is smooth on a neighborhood of \( (t_0, q_0) \).  

The solution is unique in the sense that, if two solutions \( \gamma_1 : J_1 \to M \) and \( \gamma_2 : J_2 \to M \) exist on two different intervals containing \( t_0 \), then the two solutions agree on the intersection \( J_1 \cap J_2 \). This theorem is equivalent to stating that there exists unique a maximal integral curve with starting point \( q_0 \) and starting time \( t_0 \). We we call it **maximal solution** of the Cauchy problem. Because of their time invariance, for autonomous vector fields the solutions to the Cauchy problem have the special property that \( \gamma(t; t_0, q) = \gamma(t - t_0; 0, q) \).

\(^3\) \( f_i \in C^1 \forall i \in [n] \) is sufficient.  

\(^4\) Theorem 2.15 in Agrachev et al. (2019).
Reducing to the autonomous case

We have already seen that every autonomous vector field can be considered as time dependent vector field. In this section we will see that the converse, in some sense, is also true. This means that we can reduce the study of nonautonomous vector fields to autonomous ones. Let $X : \mathbb{R} \times M \to TM$ be a time dependent vector field. We can define an autonomous vector field $\tilde{X} \in \Gamma(\mathbb{R} \times M, T\mathbb{R} \times TM)$ on the extended space $\mathbb{R} \times M$:

$$
\tilde{X}(s,q) := \left( \frac{\partial}{\partial t} \bigg|_{t=s}, X(s,q) \right)
$$

It’s then straightforward to verify that $\gamma(t; t_0, q_0)$ is a solution of the Cauchy problem for $X$ with initial point $q_0$ an initial time $t_0$ if and only if $(t, \gamma(t; t_0, q_0))$ is the solution of the Cauchy problem for $\tilde{X}$ with initial point $(t_0, q_0)$ and initial time $0$. Therefore considering a time dependent vector field on a manifold $M$ is equivalent to consider a particular type of vector fields on $\mathbb{R} \times M$.

Complete vector field

A time-dependent vector field $X$ is complete if for every $(t_0, q_0) \in \mathbb{R} \times M$, the maximal solution $\gamma(t; t_0, q_0)$ of the Cauchy problem 4.4 is defined on all $\mathbb{R}$.

In general not all vector fields on a smooth manifolds are complete. A sufficient condition is given by the following theorem:

**Theorem 4.2.** Every compactly supported smooth vector field on a smooth manifold is complete.

This means that in a compact manifold all vector fields are complete. In the Euclidean space is well known that every sublinear vector field is complete:

**Theorem 4.3.** Let $X \in \Gamma^\infty(\mathbb{R}^n, T\mathbb{R}^n)$ a smooth vector field such that:

$$
\exists C_1, C_2 > 0 \text{ s.t. } |X(x)| \leq C_1 x + C_2 \quad \forall x \in \mathbb{R}^n
$$

Then $X$ is complete

For complete Riemannian manifolds we can obtain a similar result using Gronwall estimates (on Riemannian manifolds):

---

5Theorem 9.16 in Lee (2013).
6Remark 2.6 in Agrachev et al. (2019).
Theorem 4.4. Let \((M, g)\) a complete Riemannian manifold and \(X \in \Gamma^\infty(M, TM)\) a smooth vector field on \(M\). If
\[
C := \sup_{p \in M}\{\|\nabla X(p)\|_g\} < +\infty \tag{4.5}
\]
Then \(X\) is complete, where \(\|\nabla X(p)\|_g = \sup_{Y \in T_p M}\{\|\nabla YX(p)\|_g \cdot \|Y\|_g\}\).

Proof. Suppose by contradiction that there exists a maximal integral curve \(\gamma : J \to M\) such that the interval \(J\) has a finite least upper bound \(b := \sup_{t \in J}\{t\} < +\infty\). Then by the global escape lemma (Lemma 9.19 in Lee (2013)) \(\forall t_0 \in J \gamma([t_0, b))\) is not contained in any compact set of \(M\). Since by Hopf Rinov theorem all closed balls \(B(q, R), \forall q \in M, \forall R > 0\) are compact sets, this implies that:
\[
\sup_{t \in [t_0, b]}\{\ell(\gamma([t_0, t]))\} = +\infty \tag{4.6}
\]
Where \(\ell(\cdot)\) denotes the Riemannian length of the curve. Now let \(t_0 \in J\) be such that \(t_0 - \Delta t \in J\) where \(\Delta t := b - t_0\). We then have that
\[
\ell(\gamma([t_0, t_0 + s])) \leq \ell(\gamma([t_0 - \Delta t + s, t_0 + s])) \leq \ell(\gamma([t_0 - \Delta t, t_0])) e^{Cs} \leq \ell(\gamma([t_0 - \Delta t, t_0])) e^{C\Delta t} \forall s \in [0, \Delta t] \tag{4.7}
\]
Where the first inequality holds because \(s \in [0, \Delta t] \Rightarrow [t_0, t_0 + s] \subset [t_0 - \Delta t + s, t_0 + s]\); the second follows from Proposition 1 in Kunzinger et al. (2006). The above equation is contradiction with equation (4.6), therefore \(J\) has no finite least upper bound. Similarly we can prove that there is not maximal solution with no finite greatest lower bound, this can be done just by considering the field \(-X\). We then have proven the field \(X\) is complete. \(\Box\)

4.1.2 Flows

Instead of considering the single trajectory of a point, we are interested in considering the solution of the differential equation globally as a function \(M \to M\) defined at every point \(q\) by the solution of the Cauchy problem (supposing that the starting time and the final time are fixed). Restricting for now to the autonomous case, this means considering the family of maps \(\phi^t : M \to M, \phi^t(q) = \gamma(t; 0, q)\) where \(t \in \mathbb{R}\). In this case we say that the vector field generates a flow.

However in general \(\phi^t(q)\) exists only for a subset \(\mathcal{D} \subseteq \mathbb{R} \times M\). When the vector field \(X\) is complete, it generates a flow defined in all \(\mathcal{D} = \mathbb{R} \times M\),
in this case we call it **global flow**. A global flow is abstractly defined as following:

**Definition 4.4.1.** A **global flow** is a continuous left $\mathbb{R}$-action on a manifold $M$; that is, a continuous map $\phi : M \times \mathbb{R} \to M$ satisfying the following properties for all $s,t \in \mathbb{R}$ and $p \in M$:

$$
\phi(t, \phi(s, p)) = \phi(t + s, p), \quad \phi(0, p) = p \quad (4.9)
$$

Given a global flow $\phi$ on $M$, we define the following functions$^7$:

$$
\phi^t : M \to M \quad \forall t \in \mathbb{R} \quad (4.10)
$$

$$
p \mapsto \phi^t(p) := \phi(t, p) \quad (4.11)
$$

$$
\phi^{(p)} : \mathbb{R} \to M \quad \forall p \in M \quad (4.12)
$$

$$
p \mapsto \phi^{(p)} := \phi(t, p) \quad (4.13)
$$

From the group laws it immediately follows that each $\phi^t$ is invertible:

$$
(\phi^t)^{-1} = \phi^{-t} \quad \forall t \in \mathbb{R} \quad (4.14)
$$

This means that $\{\phi^t\}_{t \in \mathbb{R}}$ is a family of homeomorphisms.

Given a smooth global flow $\phi$, we can define a vector field $X \in \Gamma^\infty(M, TM)$ in the following way:

$$
X_p = \frac{d}{dt} \bigg|_{t=0} \phi^{(p)}(t) = \frac{d}{dt} \bigg|_{t=0} \phi^t(p), \quad \forall p \in M \quad (4.15)
$$

This means that $X_p$ is well defined as the tangent vector of the curve $\phi^{(p)}$ at $0$. We call this vector field the **infinitesimal generator** of $\phi$, the name is clarified by the following Proposition:

**Prop 4.4.1.** $^8$ Let $\phi : \mathbb{R} \times M \to M$ be a smooth global flow on a smooth manifold $M$. The infinitesimal generator $X$ of $\phi$ is a smooth complete vector field on $M$; and each curve $\phi^{(p)}$ is an integral curve of $X$.

We therefore have that every smooth global flow is generated by a complete vector field. The converse is also true, this result is known as the **fundamental theorem theorem of flows**, we give here the simplified version for complete vector fields:

---

$^7$In the rest of the thesis, we will often consider a flow as a family formed by the maps $\{\phi^t\}_{t \in \mathbb{R}}$.

$^8$Proposition 9.7 in Lee (2013)
Theorem 4.5. Let $X$ be a smooth vector field on a smooth manifold $M$. There is a unique smooth maximal flow $\phi : \mathbb{R} \to M$ whose infinitesimal generator is $X$. This flow has the following property: For each $p \in M$, the curve $\phi^{(p)} : \mathbb{R} \to M$ is the unique maximal integral curve of $X$ starting at $p$.

This result tells us that given a complete vector field on a smooth manifold, this automatically determines a family of diffeomorphisms on the manifold, that can be obtained by finding the integral curves. Since in the rest of the work we will only consider complete vector fields and global flows we will refer to them simply, as flows.

Time dependent flows

As we saw, the solutions of a time dependent vector field depend on both the initial and final time $t_0$ and $t_1$, and not merely on the difference $t_1 - t_0$ as in the autonomous case. For this reason, a nonautonomous vector field cannot define a flow. Instead, we have to consider time dependent flows. To see how these objects arise consider $X$ a complete time dependent vector field on a manifold $M$ and $\tilde{X}$ its associated vector field on $\mathbb{R} \times M$. Then $\forall t_0, t_1 \in \mathbb{R}$ we can then define $\phi^{t_1, t_0}_X : M \to M$ as the (unique) map that that satisfies:

$$\phi^{t_1, t_0}_X ((t_0, q)) = (t_1, \phi^{t_1, t_0}_X (q))$$

(4.16)

It’s then straightforward to verify that this corresponds to mapping each point $q \in M$ to the solution at time $t_1$ of the Cauchy problem with initial point $q$ and initial time $t_0$:

$$\phi^{t_1, t_0}_X : M \to M$$

$$q \mapsto \gamma(t_1; t_0, q)$$

(4.17)

(4.18)

A time dependent flow is then defined as the following:

Definition 4.5.1. A global time dependent flow on a smooth manifold is a two parameter family of continuous maps $\{\phi^{t_1, t_0}\}_{t_0, t_1 \in \mathbb{R}}$ that satisfy the following conditions:

1. $\phi^{t,t} = Id \quad \forall t \in \mathbb{R}$
2. $\phi^{t_2, t_1} \circ \phi^{t_1, t_0} = \phi^{t_2, t_0} \quad \forall t_0, t_1, t_2 \in \mathbb{R}$

From the definition it immediately follows that

$$\left(\phi^{t_1, t_0}\right)^{-1} = \phi^{t_0, t_1} \quad \forall t_0, t_1 \in \mathbb{R}$$

(4.19)
And therefore that $\phi^{t_1,t_0}$ is a homeomorphism $\forall t_0, t_1 \in \mathbb{R}$.

Consider now the family $\{\phi^{t_1,t_0}_{X}\}_{t_0,t_1 \in \mathbb{R}}$ defined above, using the properties of flows is then straightforward to verify that it verifies the axioms for a time dependent flow. Moreover, from our construction and the fundamental theorem of flows it immediately follows that every complete time dependent vector field generates a time dependent flow. Conversely, as in the autonomous case, given a smooth time dependent flow $\{\phi^{t_1,t_0}\}_{t_0,t_1}$ we can define a time dependent vector field, called the infinitesimal generator, in the following way

$$X_t(q) := \frac{d}{ds} \bigg|_{s=t} \phi^{s,t}(q) \quad \forall q \in M \forall t \in \mathbb{R} \quad (4.20)$$

### 4.2 Measure pushforward induced by the flow

#### 4.2.1 Divergence of a vector field

**Definition 4.5.2.** Let $M$ a smooth manifold and $X \in \Gamma^{\infty}(M, TM)$ a smooth vector field. Given a positive $\mu \in \Gamma^{\infty}(M, DM)$ we define the divergence of $X$ with respect to $\mu$ as the function $\text{div}_\mu(X) \in C^{\infty}(M)$ such that

$$\mathcal{L}_X(\mu) = \text{div}_\mu(X) \cdot \mu \quad (4.21)$$

Similarly, given a non vanishing smooth volume form $\omega \in \Gamma^{\infty}(M, \Lambda^n T^*M)$ we can define the divergence of $X$ with respect to $\omega$ as the function $\text{div}_\omega(X) \in C^{\infty}(M)$ such that:

$$\mathcal{L}_X(\omega) = \text{div}_\omega(X) \cdot \omega \quad (4.22)$$

The following Lemma shows that the divergence with respect to a form and the divergence with respect to a density are essentially the same object

**Lemma 4.5.1.** Let $M$ a smooth manifold and $X \in \Gamma^{\infty}(M, TM)$ a smooth vector field. Given a non vanishing smooth volume form $\omega \in \Gamma^{\infty}(M, \Lambda^n T^*M)$ we have:

$$\text{div}_\omega(X) = \text{div}_{|\omega|}(X) \quad (4.23)$$

**Proof.** Fix $q \in M$ we have that:

$$\mathcal{L}_X(\omega)_q = \lim_{t \to 0} \frac{((\phi_X^{-t})^* \omega)_q - \omega_q}{t} \quad (4.24)$$
CHAPTER 4. CNF ON MANIFOLDS

Since $\omega$ is nowhere vanishing, for small enough $t \in \mathbb{R}$ we can define $a_t \in \mathbb{R}$ such that:

$$((\phi^{-t}_X)^* \omega)_q = a_t \omega_q$$  \hspace{1cm} (4.25)

Since the field and the volume form are both smooth, the number $a_t$ depends continuously on $t$. Now plugging it back:

$$\mathcal{L}_X (\omega)_q = \lim_{t \to 0} \frac{a_t \omega_q - \omega_q}{t} = \lim_{t \to 0} \frac{a_t - 1}{t} \omega_q$$  \hspace{1cm} (4.26)

Now we can see that:

$$\mathcal{L}_X (|\omega|)_q = \lim_{t \to 0} \frac{((\phi^{-t}_X)^* |\omega|)_q - |\omega|_q}{t} = \lim_{t \to 0} \frac{|((\phi^{-t}_X)^* \omega)_q| - |\omega|_q}{t}$$  \hspace{1cm} (4.27)

$$= \lim_{t \to 0} \frac{|a_t \omega_q| - |\omega|_q}{t} = \lim_{t \to 0} \frac{a_t |\omega|_q - |\omega|_q}{t} = \lim_{t \to 0} \frac{a_t - 1}{t} |\omega|_q$$  \hspace{1cm} (4.28)

Where we have used that since $\lim_{t \to 0} a_t = 1$ we can assume for small enough $t$ that $a_t$ is positive and therefore $|a_t \omega_q| = a_t |\omega_q| = a_t |\omega|_q$. Therefore comparing Equation (4.28) and Equation (4.26) we see that we have proven:

$$\text{div}_{\omega}(X) (q) = \text{div}_{|\omega|}(X) (q) \quad \forall q \in M$$  \hspace{1cm} (4.29)

From the properties of the Lie derivative it follows that:

**Prop 4.5.1.** Let $M$ be a smooth manifold, $X, Y \in \Gamma^\infty (M, TM)$ smooth vector fields and $f \in C^\infty (M)$ a smooth function, Given a non vanishing density $\mu \in \Gamma^\infty (M, D^M)$:

1. $\text{div}_{\mu}(X + Y) = \text{div}_{\mu}(X) + \text{div}_{\mu}(Y)$
2. $\text{div}_{\mu}(fX) = X(f) + f \text{div}_{\mu}(X) = df(X) + f \text{div}_{\mu}(X)$

**Divergence on local coordinates**

Let $(U; x_i), U \subseteq M$ be a local chart, $X \in \Gamma^\infty (M, TM)$ be a smooth vector field and $\omega \in \Gamma^\infty (M, \Lambda^n T^*M)$ a smooth nonvanishing volume form \footnote{For a local expression of the divergence with respect to a smooth nonvanishing density $\mu \in \Gamma^\infty (M, D^M)$ we can consider, eventually restricting $U$, $\mu|_U = |\omega||_U$ and use Lemma 4.5.1.} with
local expressions:
\[
\omega \big|_U = h \, dx_1 \wedge \cdots \wedge dx_n \\
X \big|_U = \sum_{i=1}^{n} f_i \partial x_i
\]

Where \( f_i \in C^\infty(U), \forall i \in [n] \) and \( h \in C^\infty(U), \ h(x) \neq 0 \ \forall x \in U \).

\[
\mathcal{L}_X(\omega) \big|_U = \mathcal{L}_X(h \, dx_1 \wedge \cdots \wedge dx_n) = \\
= \mathcal{L}_X(h) \, dx_1 \wedge \cdots \wedge dx_n + h \cdot \mathcal{L}_X(dx_1 \wedge \cdots \wedge dx_n) = \\
= \left( \sum_{i=1}^{n} f_i \frac{\partial h}{\partial x_i} \right) \cdot dx_1 \wedge \cdots \wedge dx_n + \\
+ h \cdot \sum_{i=1}^{n} dx_1 \wedge \cdots \mathcal{L}_X(dx_i) \wedge \cdots \wedge dx_n = \\
= \left( \frac{1}{h} \sum_{i=1}^{n} \frac{\partial (hf_i)}{\partial x_i} \right) \cdot h \, dx_1 \wedge \cdots \wedge dx_n
\]

Where in the last passage we used that
\[
\mathcal{L}_X(dx_i) = d(dx_i(X)) = d(f_i) = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \, dx_j
\]

We therefore have proven that the divergence has the following expression in local coordinates:
\[
\text{div}_\omega(X) \big|_U = \frac{1}{h} \sum_{i=1}^{n} \frac{\partial (hf_i)}{\partial x_i}
\]

For the special case of a semi-Riemannian manifold \((M,g)\), using Equation (A.5) we have \( h = \sqrt{|\det g_{ij}|} \), this gives us the expression for the semi-Riemannian divergence in local coordinates:
\[
\text{div}_\omega(X) \big|_U = \frac{1}{\sqrt{|\det g_{ij}|}} \sum_{i=1}^{n} \frac{\partial (f_i \sqrt{|\det g_{ij}|})}{\partial x_i}
\]

Divergence on semi-Riemannian manifolds

Consider \((M,g)\) a semi-Riemannian manifold (with connection \(\nabla\)), and a smooth vector field \(X \in \Gamma^\infty(M,TM)\). Given a point \(q \in M\), and \(E_1, \cdots, E_n\)
a orthonormal frame for $T_q M$, the divergence can be expressed using the covariant derivative\textsuperscript{10}:
\[
\text{div}_{\mu}(X) = \sum_{i=1} g(E_i, E_i) \cdot g(\nabla E_i X, E_i) 
\] (4.40)

### 4.2.2 Continuity equation on Manifolds

Let $X$ be a time dependent complete vector field on a smooth manifold $M$, and $\{\phi^{t,s}_X\}$ its time dependent flow. Let $\mu \in \Gamma^\infty(M, \mathcal{D}M)$ be a smooth positive density, we are interested on the change of variables induced by the diffeomorphisms $\phi^{t,s}_X$.

Consider an initial density $\rho_0 \mu$ where $\rho_0 \in C^\infty(M)$, we define $\rho_t$ as the smooth function such that
\[
\rho_t \mu = (\phi^{0,t}_X)^* (\rho_0 \mu) 
\] (4.41)

**Theorem 4.6.** Let $M$, $\mu$, $X$, $\rho_t$ as defined above. Then the function $\rho \in C^\infty(M \times \mathbb{R})$, $\rho(\cdot, t) = \rho_t$ satisfies the following linear PDE:
\[
X_t(\rho_t) + \rho_t \text{div}_\mu(X_t) = -\partial_t \rho 
\] (4.42)

**Proof.** We begin by fixing $q \in M$ and $t \in \mathbb{R}$, we then know that there exists an open neighborhood $U \ni q$ and a volume form $\omega$ such that:
\[
\mu \mid_U = |\omega| 
\] (4.43)

Since the flow, as a map $\mathbb{R} \times M \times \mathbb{R} \to \mathbb{R} \times M \times \mathbb{R}$ is continuous, there exists an open interval $I \ni t$ and a open neighborhood $U \supseteq V \ni q$ such that:
\[
\phi^{s,t}_X(V) \subseteq U \quad \forall s \in I 
\] (4.44)

In this set we can work directly with the volume form $\omega$:
\[
(\rho_t \omega)_q = \left( (\phi^{t,s}_X)^* \left( (\phi^{s,t}_X)^* (\rho_t \omega) \right) \right)_q = \left( (\phi^{t,s}_X)^* (\rho_s \omega) \right)_q \quad \forall s \in I 
\] (4.45)

Applying $\frac{d}{ds}\mid_{s=t}$ to each side of the equation and using Proposition 22.15 of Lee (2013) we obtain
\[
0 \cdot \omega_q = \left( \mathcal{L}_{X_t}(\rho_t \omega) + \frac{d}{ds}\mid_{s=t} \rho_t \cdot \omega \right)_q = 
\] (4.46)
\[
= \left( \text{div}_\omega(X_t) + X_t(\rho_t) + \frac{d}{ds}\mid_{s=t} \rho_t \right) \cdot \omega_q 
\] (4.47)

From Lemma 4.5.1 and the generality of $q$ the thesis follows. \hfill \Box

\textsuperscript{10}Definition 47 (O’Neill, 1983)
4.2.3 Continuity equation on characteristic curves

Albeit insightful, directly using Theorem 4.6 for computing the change of density has limited practical usefulness, as using numerical methods for solving PDEs scale poorly with the dimensionality of the manifold. As we already observed in Chapter 3, in the context of normalizing flows, we are specifically interested in computing the value of the density $\rho_t \mu$ on the trajectory $\phi_t^{s,0}$.

Following the notation defined in chapter 3 we are interested in the lifted map $(\phi_s^t)^D : DM \rightarrow DM$: such that

$$(\phi_s^t)^D ((\rho_s \mu)_q) = (\rho_t \mu)_{\phi_s^t(q)} \quad \forall q \in M, \forall s, t \in \mathbb{R}$$

(4.48)

The family of maps $\{(\phi_s^t)^D\}_{s,t \in \mathbb{R}}$ gives us the value of the transformed density, computed at the integral curves of the vector field $X$, which is the quantity that we are interested in. A key observation is that the family $\{(\phi_s^t)^D\}_{s,t \in \mathbb{R}}$ defines a time dependent flow on the density bundle $DM$.

**Prop 4.6.1.** Let $M$ be a smooth manifold and $X$ a complete smooth time-dependent vector field and $\{(\phi_s^t)^D\}_{s,t \in \mathbb{R}}$ its time dependent flow. Then family $\{(\phi_s^t)^D\}_{s,t \in \mathbb{R}}$ defines time dependent flow on the density bundle $DM$ and therefore generates a time dependent vector field $X^D$ on $DM$.

**Proof.** Since $\phi_s^t$ is a vector bundle isomorphism, it is in particular a diffeomorphism. Let $v_q \in DM$ with base point $q := \pi_M(v) \in M$. Using the definition of lift to the density bundle:

$$(\phi_s^t)^D \circ (\phi_s^t)^D (v_q) = (\phi_s^t)^D \left( (\phi_s^t)^* v_q \right)_{\phi_s^t(q)} =$$

$$= \left( (\phi_s^t)^* \circ (\phi_s^t)^* \right) v_q \phi_s^t \circ \phi_s^t (q) = \left( (\phi_s^t \circ \phi_s^t)^* \right) v_q \phi_s^t \circ \phi_s^t (q) =$$

$$= \left( (\phi_s^t)^* \right) v_q \phi_s^t (q) = (\phi_s^t)^D (v_q) \quad \forall r, s, t \in \mathbb{R}$$

To find an explicit expression for $X^D$, we fix a smooth positive density $\mu \in \Gamma^\infty (M, DM)$ and consider the induced vector bundle isomorphism $\mathbb{R} \times M \rightarrow DM$. Via this global trivialization, the vector field\footnote{With an abuse of notation we identify the vector field $X^D$ on $DM$ and the vector field on $\mathbb{R} \times M$ induced by the vector bundle isomorphism. Since the isomorphism depends explicitly on $\mu$ the expression that we will find will explicitly depend on $\mu$ even if the original vector fields was defined independently.} $X^D$ is a time dependent
vector field on \( \mathbb{R} \times M \), linear on the first component (since \((\phi^r_s)^D\) is linear on the fiber) and such that the projection to \( M \) corresponds with \( X \). Using the linearity, fixed a time \( t \in \mathbb{R} \), it is sufficient to evaluate the vector field \( X^D_t \) at \((1, q) \in \mathbb{R} \times M \). In order to compute this value, we first fix an initial density \( \rho_t \mu \) where \( \rho_t \equiv 1 \) and then define \( \rho_s \mu := \left( \phi^{s,t}_X \right)^D (\rho_t \mu) \) be the solution of the Cauchy problem associated with the field \( X^D \), with initial point \((1, q)\) and initial time \( t \):

\[
a(s)\mu_{\phi^{s,t}_X(q)} = \left( \phi^{s,t}_X \right)^D (\rho_t(q)\mu_q) = \rho_s \left( \phi^{s,t}_X(q) \right) \mu_{\phi^{s,t}_X(q)}
\]

(4.49)

We therefore have that \( a(s) = \rho_s \left( \phi^{s,t}_X(q) \right) \) gives the value of transformed density along the flow trajectory. We then notice that this value is also given by the solution of the continuity equation computed on the trajectory given by the flow of the vector field \( X \). To find the value of \( X^D_t (1, q) \) with respect to its first component it is then sufficient to compute \( \dot{a}(t) \) using the continuity equation:

\[
\dot{a}(t) = \frac{d}{ds} \bigg|_{s=t} \left( \rho_s \left( \phi^{s,t}_X(q) \right) \right) = \left( \frac{d}{ds} \bigg|_{s=t} \rho_s \right)(q) + \rho_t \left( \frac{d}{ds} \bigg|_{s=t} \phi^{s,t}_X(q) \right) =
\]

\[
= -\text{div}_\mu(X_t)(q) - X_t(\rho_t) + X_t(\rho_t) = -\text{div}_\mu(X_t)(q)
\]

(4.50)

(4.51)

The method of solving first order PDEs computing the value of the solution on the flow of an associated vector field is known as method of characteristics.

We then have found the following expression for \( X^D_t \):

\[
X^D_t : \mathbb{R} \times M \rightarrow T\mathbb{R} \times TM
(a, q) \mapsto (-a \text{div}_\mu(X_t)(q) \cdot \partial_{x_0}, X_t(q))
\]

(4.52)

(4.53)

Where \( \partial_{x_0} \) is the coordinate field for \( T\mathbb{R} \).

**Integrating the density on characteristic curves**

Suppose we are given an initial density \( \rho_0 \mu \). We are interested in computing the value of \( \rho_t \left( \phi^{t,0}_X(q) \right) \) for a given \( t \in \mathbb{R} \) and \( q \in M \). This corresponds to finding the solution to the Cauchy problem for \( X^D \) with initial point \((\rho_0(q), q)\)

\[12\]see Chapter 9 of Lee (2013).
and initial time 0. By the linearity of $X^D$ in its first component the solution is given by:

$$
\rho_t (\phi_{X}^{t,0} (q)) = -\exp \left( \int_0^t \text{div}_\mu (X_t) \left( \phi_{X}^{t,0} (q) \right) dt \right) \cdot \rho_0 (q)
$$

(4.54)

In many of the applications we are interested in $\log \rho_t$:

$$
\log \rho_t (\phi_{X}^{t,0} (q)) = -\int_0^t \text{div}_\mu (X_t) \left( \phi_{X}^{t,0} (q) \right) dt + \log \rho_0 (q)
$$

(4.55)
Chapter 5

Parameterizing vector fields

Given a manifold $M$, we saw in last chapter that vector fields naturally give rise to diffeomorphisms on $M$, which can then be used to define continuous normalizing flows on $M$.

We are then left with the problem of parameterizing an expressive enough set of vector fields on the manifold. When we try to parameterize a large set of functions in a modular way, we look at neural networks as a natural solution, however, they can only parameterize functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$, and thus there is no straightforward way to use them for this task.

Finding the best way of parameterizing vector fields on manifolds is an interesting problem with no unique solution, how to tackle it will largely depend on how the manifold is defined and what data structure is used to parameterize it in practice. Nevertheless, all the objects and methods discussed in the rest of the paper are defined independently from the specific parameterization method chosen. Therefore, if in the future a better way of parameterizing vector fields will emerge, they will still be applicable.

Notwithstanding the above, in this section, we will try to give some guidance on how to approach this problem. In the first part, we will show how, using a generating set, it can be reduced to the much easier task of parameterizing functions on manifolds. We will then give some practical advice in the case where the manifold is a Homogenous space, or it is described using an embedding in $\mathbb{R}^m$. For the rest of the thesis, given a function $f : M \rightarrow \mathbb{R}^m$ we will indicate with $f_i : M \rightarrow \mathbb{R}$ its $i$-th component, such that $f = (f_1, \cdots, f_m)$. 
5.0.1 Local frames and global constraints

We begin by analyzing how we parameterize vector fields in $\mathbb{R}^n$, to investigate to what extent we can generalize this procedure. As we saw in Secton 2.3, in the Euclidean space vector fields are simply functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. In a more geometrical language, the function $f$ defines the vector field $X$ in the following way:

$$X = f_1 \partial x_1 + \cdots + f_n \partial x_n.$$  \hfill (5.1)

The converse is also true: for every vector field $X$ there exist a unique continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that Equation (5.1) holds. On a generic $n$-dimensional smooth manifold this is only true locally. This means that there exists a open cover $\{U_i\}_{i \in I}$ of $M^1$, called the trivialization cover, such that $TM$ restricted to each $U_i$ is isomorphic to the trivial bundle. This is equivalent to saying that for every set $U_i$ there exist $n$ smooth vector fields $E_1^{(i)}, \cdots, E_n^{(i)} \in \Gamma^\infty(U_i, TU_i))$ such that for every smooth vector field $X \in \Gamma^\infty(M, TM)$ there exists a unique smooth function $f : U_i \rightarrow \mathbb{R}^n$ such that:

$$X|_{U_i} = f_1 E_1^{(i)} + \cdots + f_n E_n^{(i)}. \hfill (5.2)$$

We then can call $E_1^{(i)}, \cdots, E_n^{(i)}$ a local frame. A local frame that is defined on an open domain $U = M$ (this means on the entire manifold) is called a global frame. On a manifold there exists plenty of local frames, in fact given a smooth local chart $(U; x_i)$ the fields $\partial_{x_1}, \cdots, \partial_{x_n} \in \Gamma^\infty(U, TU)$ form a local frame called coordinate frame. In the special case of $\mathbb{R}^n$, its coordinate frame is a global frame. Unfortunately, in general, not every manifold has a global frame, the simplest example is the sphere $S^2$. In the sphere case it is well known that there exists no vector field that is everywhere nonzero, this result goes by the hairy ball theorem. It is then clear that no pair of vector fields $E_1, E_2 \in \Gamma^\infty(M, TM)$ can form a global frame, in fact, there will always be a point $q \in S^2$ such that:

$$\text{span} \left( (E_1)_q, (E_2)_q \right) \leq 1 < 2 = \dim(T_qM). \hfill (5.3)$$

The manifolds for which a global frame $E_1, \cdots, E_n \in \Gamma^\infty(M, TM)$ exists are called parallelizable manifolds, for this class we can parameterize all

---

1Assuming that the manifold is second countable, there exists $I$ that is finite and has cardinality $n + 1$, see Lemma 7.1 in Walschap (2004).
smooth vector fields on \( M \) in the same way as we did on \( \mathbb{R}^n \). This means choosing a smooth function \( f : M \to \mathbb{R} \) and defining a vector field \( X \):

\[
X = f_1 E_1 + \cdots + f_n E_n. \tag{5.4}
\]

A manifold is parallelizable iff its tangent bundle is isomorphic to the trivial bundle: \( \mathbb{R}^n \times M \cong TM \). A global frame gives an explicit isomorphism:

\[
\mathbb{R}^n \times M \to TM, \quad (\beta, q) \mapsto \beta_1 (E_1)_q + \cdots + \beta_n (E_n)_q.
\]

An important and large class class of parallelizable manifolds is given by Lie Groups, which are smooth manifold which additionally posses a group structure compatible with the manifold structure.

### 5.0.2 Lie groups

A **Lie group** \( G \) is a smooth manifold with the additional structure of a group, where the group multiplication and inversion are smooth maps. Lie groups are an important instrument in physics where they are used to model continuous symmetries. Many relevant Lie groups arise as subgroups of the matrix groups \( \text{GL}_n(\mathbb{R}) \) and \( \text{GL}_n(\mathbb{C}) \) of real and complex invertible matrices with matrix multiplications as a group product.

The **Lie algebra** \( \mathfrak{g} \) of a Lie group \( G \) is the tangent space of the group at the identity element \( \mathfrak{g} := T_e G \). The Lie algebra \( \mathfrak{g} \) can be identified with the space of (right) left invariant vector fields on \( G \). In fact any vector \( v \in \mathfrak{g} \) defines a left invariant vector field \( v^L \) and a right invariant vector field \( v^R \) in the following way:

\[
v^L_a := dL_a(v), \quad v^R_a := dR_a(v), \quad \forall a \in G. \tag{5.5}
\]

Where \( L_a, R_a : G \to G \) are respectively the left and the right group multiplication. Conversely any left (right) invariant vector field \( V \in \Gamma^\infty (G, TG) \) gives a Lie algebra element \( V_e \in \mathfrak{g} \). With this identification we can define the Lie bracket in \( \mathfrak{g} \) using the Lie bracket between the associated left invariant vector fields:

\[
[v, w] := [v^L, w^L], \quad \forall v, w \in \mathfrak{g}. \tag{5.6}
\]

A fundamental property of Lie groups is that they are parallelizable manifolds. A basis \( \{v_1, \cdots, v_n\} \subset T_e G = \mathfrak{g} \) defines a **global frame** \( \{E_i\}_{i=1}^n \) for \( TG \), where \( E_i := v_i^L \), or \( E_i := v_i^R \).
Any scalar product \( \langle \cdot, \cdot \rangle_g \) on \( g \) defines a left invariant Riemannian metric \( g \) on \( G \):

\[
g(v_a, w_a) := \langle dL_{a^{-1}}(v_a), dL_{a^{-1}}(w_a) \rangle_g, \quad \forall a \in G, \forall v_a, w_a \in T_a G.
\] (5.7)

This, in turn, induces a left-invariant Riemannian density \( \mu_g \), which is unique up to a normalizing constant (which depends on the initial scalar product choice). The associated left-invariant Borel measure is known as (left) Haar measure. A similar construction can be done to define a right invariant volume form on \( G \). A Lie group is unimodular if its left and right invariant volume forms coincide. Examples of unimodular groups are compact Lie groups and semisimple Lie groups. For proofs, additional details and background we refer to Lee (2013) Chapters 7, 16 and Falorsi et al. (2019) Appendix D.

Given \( v \in g \), the exponential map is defined as \( \exp(v) := \gamma(1) \) where \( \gamma : \mathbb{R} \to G \) is the only 1-parameter subgroup such that \( \dot{\gamma}(0) = v \). The exponential map \( \exp : g \to G \) describes the flow of left and right invariant vector fields:

\[
\phi^L_t v = a \exp(tv), \quad \phi^R_t v(a) = \exp(tv)a, \quad \forall a \in G, \forall v \in g. \] (5.8)

Notice that left invariant vector fields act by right translation and vice-versa. Therefore, given any right invariant vector field, \( v^R \in \Gamma^\infty_G(TM) \) its flow \( \Phi^R_t : G \to G \) is an isometry with respect to the left invariant Riemannian metric \( g \). This means that \( v^R \) has 0 divergence, in fact:

\[
\text{div}_{\mu_g}(v^R) = L(\mu_g) = \frac{d}{dt} \bigg|_{t=0} \left( (\phi^R_t)^\ast \mu_g \right) = \frac{d}{dt} \bigg|_{t=0} \left( L_{\exp(tv)} \right)^\ast \mu_g = \frac{d}{dt} \bigg|_{t=0} (\mu_g) = 0 \cdot \mu_g, \] (5.9)

which implies \( \text{div}_{\mu_g}(v^R) = 0 \). We therefore can obtain a global frame \( \{v_i^R\}_{i=1}^n \) formed by zero divergence vector fields. When the Lie group is unimodular we can use left and right invariant vector fields interchangeably.

### 5.1 Generators of vector fields

We have seen that for parallelizable manifolds, once we have defined a global frame, we have a bijective correspondence between functions \( C^\infty(M, \mathbb{R}^n) \) and smooth vector fields:

\[
C^\infty(M, \mathbb{R}^n) \to \Gamma^\infty(M, TM), \quad f \mapsto f_1 E_1 + \cdots + f_n E_n. \] (5.11)
For non parallelizable manifolds, we fail to find a global frame because, given any \( n \) vector fields \( \{E_i\}_{i=1}^n \), there always exist points \( q \) where \( \{(E_i)_q\}_{i=1}^n \) fail to span all \( T_qM \):

\[
\text{span} \left( \{(E_i)_q\}_{i=1}^n \right) \subsetneq T_qM.
\]

The idea is then to add vector fields to the set \( \{(E_i)_q\}_{i=1}^n \), giving up on the injectivity constraint, until they "generate" all \( \Gamma^\infty(M, TM) \). To make this statement more precise we have to use the language of modules. In fact in general the space of smooth sections of a vector bundle \( (E, \pi, M) \) forms a module over the ring \( C^\infty(M) \) of the smooth functions on \( M \).

**Definition 5.0.1.** A finite set of vector fields \( \{X_i\}_{i=1}^m \subset \Gamma^\infty(M, TM), m \in \mathbb{N}_{>0} \) is a generating set for the \( C^\infty(M) \)-module of the smooth vector fields on \( M \) if for every vector field \( X \in \Gamma^\infty(M, TM) \) there exist \( \{f_i\}_{i=1}^m \subset C^\infty(M) \) such that:

\[
X = f_1X_1 + \cdots + f_mX_m. \tag{5.13}
\]

If there exist a generating set for \( \Gamma^\infty(M, TM) \) we then say that \( \Gamma^\infty(M, TM) \) is finitely generated.

**Lemma 5.0.1.** Let \( M \) be a smooth manifold and let \( \{X_i\}_{i=1}^m \subset \Gamma^\infty(M, TM), m \in \mathbb{N}_{>0} \) a set of smooth vector fields such that \( \text{span} \left( \{(X_i)_q\}_{i=1}^n \right) = T_qM, \forall q \in M \). Then \( \{X_i\}_{i=1}^m \) is a generating set for \( \Gamma^\infty(M, TM) \).

**Proof.** Consider the open sets

\[
U_I := \{ q \in M | \{X_i(q)\}_{i \in I} \text{ are linearly independent} \}
\]

where \( I \subset \{1, \cdots, m\} \) is any subset of indices of cardinality \( n \). Let \( \mathcal{I} := \{ |I| \subset [m], \#(I) = n, U_I \text{ is not empty} \} \). To see that these sets are open, observe that in a local coordinate chart \( (U; x_i) \) we can write \( X_i|_U = \sum_{j=1}^n a_{ij}^U \partial x_j, \forall i \in [m], a_{ij}^U \in C^\infty(U) \). In local coordinates the linear independence of \( \{X_i(q)\}_{i \in I} \), is equivalent to \( \det(A(x)) \neq 0 \). Where if \( I = i_1, \cdots, i_n \) \( A(x) \) is defined as \( A(x)_{jk} = a_{ij}^U \). From the definition of \( U_I \) it descends that the family \( \{U_I| I \subset \{1, \cdots, m\}, \#(I) = n, U_I \text{ is not empty} \} \) forms a open trivialization cover. Fixed \( I \in \mathcal{I} \) there exist smooth functions \( \{f_i^I\}_{i \in I} \subset C^\infty(U_I) \) such that

\[
X|_{U_i} = \sum_{i \in I} f_i^I E_i|_{U_i} \tag{5.14}
\]
CHAPTER 5. PARAMETERIZING VECTOR FIELDS

Now let \( \{\psi_I\}_{i \in I} \) be a smooth partition of unity subordinate to \( \{U_I\}_{i \in I} \).

Defining

\[ j^I_i := \begin{cases} \\
\psi_I \cdot f^I_i, & \text{on } U_I, \\
0, & \text{on } M \setminus \text{supp}(\psi_I). \end{cases} \quad \forall I \in \mathcal{I}, \forall i \in I. \quad (5.15) \]

We have that:

\[ \psi_I \cdot X = \sum_{i \in I} j^I_i X_i \quad (5.16) \]

And therefore

\[ X = \sum_{j=1}^m \left( \sum_{I \in \mathcal{I} \text{ s.t. } j \in I} j^I_i \right) X_j \quad (5.17) \]

**Theorem 5.1.** Let \( M \) be a (second countable) smooth manifold \( M \). Then the module of smooth vector fields \( \Gamma^\infty (M, TM) \) is finitely generated.

**Proof.** Since \( M \) is second countable we can apply Lemma 7.1 in Walschap (2004) and say that there exist an open trivialization cover \( \{U_i\}_{i=0}^n \), where \( n \) is the dimension of \( M \). We denote with \( E_1^{(i)}, \ldots, E_n^{(i)} \) the local frame relative to the domain \( U_i \subseteq M \). Now let \( \{\psi_i\}_{i=0}^n \) be a smooth partition of unity subordinate to \( \{U_i\}_{i=0}^n \). We define the global vector fields on \( M \):

\[ \tilde{E}_j^{(i)} := \begin{cases} \\
\psi_i \cdot E_j^{(i)}, & \text{on } U_i, \\
0, & \text{on } M \setminus \text{supp}(\psi_i). \end{cases} \quad \forall i \in [n], \forall j \in [n]. \quad (5.18) \]

Using Lemma 5.0.1 we have that \( \{\tilde{E}_j^{(i)}\}_{j \in [n]} \) is a generating set of \( \Gamma^\infty (M, TM) \).

From this Theorem and the definition of generating set we can extract a general methodology to parameterize all vector fields on smooth manifolds:

1. choose a suitable generating set \( \{X_i\}_{i=1}^m \),
2. find a way of parameterizing functions \( f_i : M \to \mathbb{R} \),
3. model a generic vector field \( X \) as a linear combination:

\[ X = f_1 X_1 + \cdots + f_m X_m. \quad (5.19) \]
The above proof also tells us that a simple and general recipe to obtain a generating set is to take a collection of local frames and multiply them by a smooth partition of unity. The efficiency of this framework is determined by the cardinality of the generating set: lower cardinality requires parameterization of fewer functions. The proof gives us an initial upper bound on the lowest cardinality of the generating set we can achieve for a generic manifold: $n^2 + n$ where $n$ is the dimension of the manifold. We will see that using the Whitney embedding theorem, and the fact that any smooth manifold admits a Riemannian metric, this number can be further reduced to $2n + 1$.

### 5.1.1 Time dependent vector fields

When parameterizing time-dependent vector fields, we need to model a vector field $X_t$ for every time $t \in \mathbb{R}$. Using a generating set we can easily accomplish this by parametrizing a function $f : \mathbb{R} \times M \to \mathbb{R}^m$ and defining:

$$X_t := f_1(t, \cdot)X_1 + \cdots + f_m(t, \cdot)X_m.$$  \hfill (5.20)

### 5.2 Divergence using a generating set

Suppose now we are using a generating set $\{X_i\}_{i=1}^m$ to parameterize a vector field $X = \sum_{i=1}^m f_i X_i$, $f \in C^\infty(M, \mathbb{R}^m)$. From the properties of the Lie derivative it is straightforward to see that its divergence decomposes in:

$$\text{div}_\mu(X) = \sum_{i=1}^m X_i(f_i) + \sum_{i=1}^m f_i \text{div}_\mu(X_i) =$$  \hfill (5.21)

$$= \sum_{i=1}^m df_i(X_i) + \sum_{i=1}^m f_i \text{div}_\mu(X_i).$$  \hfill (5.22)

Let’s analyze the above expression: the first term involves the derivative of the parameterizing functions $f_i$ on the vector field directions $X_i$, and it is the same as the divergence in the Euclidean space (with $X_i = \partial_{x_i}$). The second term involves the divergence of the vector fields in the generating set. Therefore if we are using a generating set formed by vector fields with known, or easy divergence expression we can use Equation (5.21) for divergence computation without relying on local charts. We will show later that this is always the case for homogeneous spaces (see Theorem 5.4)

As already noticed by Grathwohl et al. (2019), the first term in Equation 5.21 has complexity proportional to $m$ evaluations of the function $f$. Therefore,
when integrating the divergence along on flow curves in Equation (4.55), it can be expensive to compute at every step of the numerical solver. To mitigate this issue we show that the Monte Carlo unbiased estimator of the divergence presented in Grathwohl et al. (2019) can be generalized to manifold setting:

**Theorem 5.2.** Consider a finite set of vector fields \( \{X_i\}_{i=1}^m \subset \Gamma^\infty(M, TM) \) and a function \( f \in C^\infty(M, \mathbb{R}^m) \) on a smooth manifold \( M \). Then for every probability distribution \( \rho(\varepsilon) \) on \( \mathbb{R}^m \) with zero mean \( \mathbb{E}_{\rho(\varepsilon)}[\varepsilon] = 0 \) and identity covariance \( \text{Cov}_{\rho(\varepsilon)}[\varepsilon] = I \) it holds:

\[
\sum_{i=1}^m df_i(X_i) = \mathbb{E}_{\rho(\varepsilon)} \left[ f^* \varepsilon \left( \sum_{i=1}^m \varepsilon_i X_i \right) \right] \tag{5.23}
\]

Moreover if the function \( f \) can can be expressed as the composition of two smooth functions \( f = f^{(2)} \circ f^{(1)} \), the above term can be rewritten as:

\[
= \mathbb{E}_{\rho(\varepsilon)} \left[ f^{(2)*} \varepsilon \left( df^{(1)} \left( \sum_{i=1}^m \varepsilon_i X_i \right) \right) \right] . \tag{5.24}
\]

**Proof.** First of all we notice that the LHS of equation (5.23) can be rewritten as:

\[
\sum_{i=1}^m df_i(X_i) = \text{tr} (A) \tag{5.25}
\]

Where \( A \) is the matrix function that contains all the partial derivatives:

\[
A_{ij} = X_i(f_j) = df_j(X_i) \tag{5.26}
\]

We can then use Hutchinson’s trace estimator (Hutchinson, 1989) to obtain

\[
\text{tr} (A) = \mathbb{E}_{\rho(\varepsilon)} \left[ \varepsilon^T A \varepsilon \right] = \mathbb{E}_{\rho(\varepsilon)} \left[ \sum_{i,j=1}^m \varepsilon_i df_i (\varepsilon_j X_j) \right] = \mathbb{E}_{\rho(\varepsilon)} \left[ \sum_{i=1}^m f^*_i \varepsilon_i \left( \sum_{j=1}^m \varepsilon_j X_j \right) \right] = \mathbb{E}_{\rho(\varepsilon)} \left[ f^* \varepsilon \left( \sum_{i=1}^m \varepsilon_i X_i \right) \right] . \tag{5.27}
\]

Where we used the linearity of the differential and the pullback. Equation (5.24) can be obtained observing that the pullback is the transpose of the differential. \( \square \)
5.3 Homogeneous spaces

Definition 5.2.1. Let $N, M$ smooth manifolds and $F : M \to N$ a smooth function between them. Given $X \in \Gamma^\infty(M, TM)$ and $Y \in \Gamma^\infty(N, TN)$ smooth vector fields respectively on $M$ and $N$, we say that they are $F$-related if

$$dF_p(X_p) = Y_{F(p)}, \quad \forall p \in M. \quad (5.29)$$

A homogeneous space is a manifold equipped with a transitive Lie group action:

Definition 5.2.2. A smooth manifold $M$ is homogeneous if a Lie group $G$ acts transitively on $M$, i.e.:

There exists a smooth map $G \times M \to M, (a, x) \mapsto a.x$ such that:

- $(ab).x = a.(b.x)$,
- $e.x = x, \quad \forall x \in M,$
- for any $x, y \in M$ there exists an element $a \in G$ such that $a.x = y$.

We can construct homogeneous spaces taking a Lie group $G$ and $H < G$ a closed subgroup. Then the quotient manifold $G/H$ is a homogeneous space with action $a.(bH) := abH$ and the projection map $\pi : G \to G/H$ is a smooth submersion. In particular, any Lie group $G = G/\{e\}$ can be considered a homogeneous space with the action given by left multiplication. The following theorem tells us that the above construction completely characterizes homogeneous spaces:

Theorem 5.3. $^3$ Let $G$ be a Lie group, let $M$ be a homogeneous $G$-space, and let $q$ be any point of $M$. The isotropy group $G_q := \{a \in G : a.q = q\}$ is a closed subgroup of $G$, and the map:

$$F : G/G_q \to M \quad (5.30)$$

$$aG_q \mapsto a.q \quad (5.31)$$

is an equivariant diffeomorphism.

$^2$Theorem 21.17 Lee (2013).
$^3$Theorem 21.18 Lee (2013).
CHAPTER 5. PARAMETERIZING VECTOR FIELDS

Given $v \in \mathfrak{g}$ an element of the Lie algebra, we can define a vector field $	ilde{v} \in \Gamma^{\infty}(G/H, TG/H)$ on $G/H$ associated to it:

$$\tilde{v}_{aH} := \frac{d}{dt} \bigg|_{t=0} \exp(tv).aH, \quad \forall a \in G$$  \hspace{1cm} (5.32)

Since $\exp(tv).aH = (\exp(tv)a).H$ we have that $\tilde{v}$ and $v^R$ are $\pi$-related$^4$. Since $\pi : G \to G/H$ is a smooth submersion and we know that there exist a generating set of $\Gamma^{\infty}(G, TG)$ made by right-invariant vector fields $\{v^R_i\}_{i=1}^m$, then there exist a set $\{v_i\}_{i=1}^m \subset \mathfrak{g}$ of elements of the Lie algebra such that $\{\tilde{v}_i\}_{i=1}^m \subset \Gamma^{\infty}(G/H, TG/H)$ is a generating set for $\Gamma^{\infty}(G/H, TG/H)$. We therefore have proven the following result:

**Prop 5.3.1.** Let $M$ a be homogeneous $G$-space. Every basis$^5$ $\{v_i\}_{i=1}^m \subset \mathfrak{g}$ of the Lie algebra $\mathfrak{g}$ induces a generating set $\{\tilde{v}_i\}_{i=1}^m \subset \Gamma^{\infty}(M, TM)$ for the smooth vector fields on $M$.

Let now $\mu$ be a group action invariant$^6$ density on $G/H$. Then, given $v^R$ any right-invariant vector field on $G$, we have that the flow of $\tilde{v}$ is given by the group action of $\exp(tv)$, and therefore $\text{div}_\mu(\tilde{v}) = 0$:

$$\text{div}_\mu(\tilde{v})\mu_g = \mathcal{L}(\mu_g) = \frac{d}{dt} \bigg|_{t=0} \left((\phi_{\tilde{v}}^t)^* \mu_g\right) = \frac{d}{dt} \bigg|_{t=0} (\mu_g) = 0 \cdot \mu_g.$$  \hspace{1cm} (5.33)

We thus have proven the following theorem:

**Theorem 5.4.** Let $M$ a homogeneous $G$-space and let $\mu$ be a $G$ invariant smooth positive density on $M$. For every element $v \in \mathfrak{g}$ of the Lie algebra we have that the associated vector field $\tilde{v}$ on $M$ has zero divergence:

$$\text{div}_\mu(\tilde{v}) = 0.$$  \hspace{1cm} (5.34)

We conclude this section by giving a sufficient condition for defining a group action invariant Riemannian metric on a homogeneous space:

$^4$Proposition 9.6 Lee (2013).

$^5$Depending on the space and the group, taking only a subset of the basis might be sufficient for inducing a generating set. An example is the hyperbolic space $\mathbb{H}^n$ with action by the special indefinite orthogonal group. $SO^+(n,1)$. In fact in this case, a basis for the vector subspace of Lorentz boosts in $\mathfrak{so}(n,1)$ gives rise to a generating set for the vector fields on $\mathbb{H}^n$.

$^6$See sections 2.3.2 and 2.3.3 in Howard (1994) for construction of invariant volume forms and Riemannian metrics on homogeneous spaces.
Prop 5.4.1. \(^7\) Let \( G \) be a Lie group and \( K \) a closed subgroup of \( G \). Let \( g \) be a Riemannian metric on \( G \) which is left invariant under elements of \( G \) and also right invariant under elements of \( K \). Then there is a unique Riemannian metric \( h \) on \( G/K \) so that the natural map \( \pi : G \rightarrow G/K \) is a Riemannian submersion. This metric is invariant under the action of \( G \) on \( G/K \).

Conversely if \( K \) is compact and \( h \) is an invariant Riemannian metric on \( G/K \) then there is a Riemannian metric \( g \) on \( G \) which is left invariant under all elements of \( G \) and right invariant under elements of \( K \). We will say that the metric \( g \) is adapted to the metric \( h \).

5.4 Embedded submanifolds of \( \mathbb{R}^m \)

A general way to work in practice with manifolds is using embedded submanifolds of \( \mathbb{R}^m \). An embedding for a manifold \( M \) is a continuous injective function \( \iota : M \hookrightarrow \mathbb{R}^m \) such that \( \iota : M \rightarrow \iota(M) \) is a homeomorphism. The embedding is smooth if \( \iota \) is smooth and \( M \) is diffeomorphic to its image. In this case \( \iota(M) \) is a smooth submanifold of \( \mathbb{R}^m \). For all practical purposes we can directly identify \( M \) as a submanifold of \( \mathbb{R}^m \), the function \( \iota : M \hookrightarrow \mathbb{R}^m \) then simply denotes the inclusion. Through this identification, we can then consider the tangent space \( T_qM, q \in M \) as a vector subspace of \( T_q\mathbb{R}^m \). An embedding is said proper if \( \iota(M) \) is a closed set in \( \mathbb{R}^m \).\(^8\)

Theorem 5.5 (Whitney Embedding Theorem, 6.15 in Lee (2013)). Every smooth \( n \)-dimensional manifold admits a proper smooth embedding in \( \mathbb{R}^{2n+1} \)

The Whitney embedding theorem tells us that parameterizing manifolds as submanifolds of the Euclidean space gives us a general methodology to work with manifolds. Developing algorithms that assume that the manifold is given as an embedded submanifold of \( \mathbb{R}^m \) is therefore of outstanding importance.

For embedded submanifolds parameterizing functions is extremely easy, and can be simply done via restriction: given a smooth function \( f : \mathbb{R}^m \rightarrow \mathbb{R} \), \( f \circ \iota \) then defines a smooth function from \( M \) to \( \mathbb{R} \).

Unfortunately, for vector fields, it is not as easy. In fact, in general, given a vector field \( X \in \Gamma(\mathbb{R}^m, T\mathbb{R}^m) \) this does not restrict to a vector field on a submanifold \( M \subseteq \mathbb{R}^m \), as, in general, given \( q \in M \) we have \( X_q \not\in T_qM \subseteq T\mathbb{R}^m_q \). In order for \( X \) to restrict to a vector field on a submanifold \( M \), we

\(^7\)Proposition 2.3.14 in Howard (1994).

\(^8\)Requiring that the embedding is proper excludes embeddings of the form \( U \hookrightarrow M \) where \( U \) is an open subset of \( M \).
need for $X$ to be tangent to the submanifold:

$$X_q \in T_q M \subseteq T_q \mathbb{R}^m, \quad \forall q \in M. \quad (5.35)$$

A tangent vector field then defines a vector field on the submanifold:

**Lemma 5.5.1.** Let $M$ be a smoothly embedded submanifold of $\mathbb{R}^m$, and let $\iota : M \hookrightarrow \mathbb{R}^m$ denote the inclusion map. If a smooth vector field $Y \in \Gamma^\infty (\mathbb{R}^m, T\mathbb{R}^m)$ is tangent to $M$ there is a unique smooth vector field on $M$, denoted by $Y|_M$, that is $\iota$-related to $Y$. Conversely a vector field $\overline{Y} \in \Gamma^\infty (\mathbb{R}^m, T\mathbb{R}^m)$ that is $\iota$-related to $Y$ is tangent to $M$.

**Proof.** See proof of Proposition 8.23 in Lee (2013) \hfill $\square$

More importantly, we can parameterize all vector fields on an embedded submanifold using tangent vector fields:

**Prop 5.5.1.** Let $M$ be a properly embedded submanifold of $\mathbb{R}^m$, and let $\iota : M \hookrightarrow \mathbb{R}^m$ denote the inclusion map. For any smooth vector field $X \in \Gamma^\infty (M, TM)$ there exist a smooth vector field $\overline{X} \in \Gamma^\infty (\mathbb{R}^m, T\mathbb{R}^m)$ tangent to $M$ such that:

$$\overline{X}|_M = X. \quad (5.36)$$

We call $\overline{X}$ an extension of $X$.

**Proof.** Let $U \subseteq \mathbb{R}^m$ be a tubular neighborhood of $M$, then by Proposition 6.25 of Lee (2013), there exist a smooth map $r : U \rightarrow M$ that is both a retraction and a smooth submersion. Then, since $r$ is a submersion there exist a vector field $\overline{X} \in \Gamma^\infty (U, TU)$ that is $r$-related to $X$ \footnote{See for example Exercise 8-18 of Lee (2013)}. This means $dr_z \overline{X}_z = X_{r(z)} \forall z \in \mathbb{R}^m$. Since $r$ is a retraction

$$dt_q \circ dr_q = Id_{T_q \mathbb{R}^m} \Rightarrow (dt_q \circ dr_q) \overline{X}_q = \overline{X}_q \Rightarrow dt_q X_q = \overline{X}_q \quad \forall q \in M, \quad (5.37)$$

$\overline{X}$ is $\iota$-related to $X$ and therefore tangent to $M$ and such that $\overline{X}|_M = X$. Then $\overline{X}$ can be used to define a tangent vector field on all $\mathbb{R}^m$ using a smooth partition of unity subordinate to the open cover $\{\mathbb{R}^m \setminus M, U\}$. \hfill $\square$

From the proof of the theorem, it’s clear that the extension of $X$ is not unique. Our objective is then finding a way to parameterize all vector fields tangent to a submanifold. We first observe that given smooth vector fields
$X, Y \in \Gamma^\infty(\mathbb{R}^m, T\mathbb{R}^m)$ tangent to $M$ and smooth functions $f, g \in C^\infty(\mathbb{R}^m)$ then $fX + gY$ is tangent to $M$. This means that the set of all smooth vector fields tangent to $M$ is a submodule of the module of smooth vector fields on $\mathbb{R}^m$. Following the framework outlined in Section 5.0.1 we then need to find $l$ tangent vector fields $X_1, \ldots, X_l$ such that $X_1|_M, \ldots, X_l|_M$ generates all $\Gamma^\infty(M, TM)$.

### 5.4.1 Embedded Riemannian Submanifolds

If our embedded submanifold manifold is equipped with a Riemannian metric\(^{11}\), the gradient of the embedding gives us a generating set for the tangent bundle. We first prove the following Lemma

**Lemma 5.5.2.** Let $M$ be a embedded submanifold of $N$, and let $\iota: M \hookrightarrow N$ denote the inclusion map. Then

\[
\iota^*: T^*N \to T^*M
\]

\[
\beta_{\iota(q)} \mapsto \iota^*\beta_{\iota(q)}: \quad v \mapsto \beta_{\iota(q)}(dv_q) \quad \forall q \in M, \ \forall \beta \in T^*_q N, \ \forall v_q \in T_q M
\]

the pullback of the inclusion is a surjective vector bundle homomorphism, where by $\iota^*T^*N$ we denote the pullback bundle $\iota^*T^*N = \{(q, \beta) \in M \times T^*N | \pi(\beta) = \iota(q)\}$

**Proof.** Fix $q \in M$. Let $n$ be the dimensionality of $M$ and $m$ the dimensionality of $N$. We need to prove that $\iota^*: T^*_{\iota(q)}N \to T^*_q M$ is surjective. Let $e_1, \ldots, e_n$ a basis for $T_q M$ and $\eta_1, \ldots, \eta_n$ its dual basis. By the linearity of $\iota^*$ it’s then sufficient to prove that there exists $\beta_1, \ldots, \beta_n \in T^*_{\iota(q)}N$ such that for all $i \in \{1, \ldots, n\}$:

\[
\iota^*\beta_i = \eta_i.
\]  

(5.40)

To see this, consider the set $\{d(\eta_1)_{\iota(q)}, \ldots, d(\eta_n)_{\iota(q)}\} \subset T_{\iota(q)} N$. Since $\iota$ is an embedding, $dt$ is injective. Therefore the vectors are linearly independent. We can then complete them to a basis $v_1 := d(\eta_1)_{\iota(q)}, \ldots, w_n := d(\eta_n)_{\iota(q)}, w_{n+1}, \ldots, w_m$ of $T_{\iota(q)} N$. Let $\beta_1, \ldots, \beta_m \in T^*_{\iota(q)} N$ the dual basis. We then have:

\[
\iota^*\beta_i(e_j) = \beta_i \left( d(\eta_j)_{\iota(q)} \right) = \beta_i (w_j) = \delta_{ij} = \eta_i (e_j) \quad \forall i, j \in \{1, \ldots, n\}.
\]  

(5.41)

Thus $\beta_i$ satisfies Equation (5.40) $\forall i \in \{1, \ldots, n\}$

\[\Box\]

\[^{10}\text{Since the extension of a vector field is not unique this is different from finding a generating set for the submodule of vector fields tangent to } M\]

\[^{11}\text{We do not assume the Riemannian metric to be inherited from the ambient space.}\]
Theorem 5.6. Let $(M, g)$ be a embedded submanifold of $\mathbb{R}^m$, and let $z : M \hookrightarrow \mathbb{R}^m$ denote the inclusion map. Then $\{\text{grad}(z_i)\}_{i=1}^m$ is a generating set for the smooth vector fields $\Gamma^\infty(M, TM)$. Where $\text{grad}$ denotes the Riemannian gradient with respect to the metric $g$.

Proof. Consider the differential forms $\{dz_i\}_{i=1}^m \subset \Gamma^\infty(M, T^*M)$. Using Lemma 5.5.2 we have that span $\{(dz_i)(q)\}_{i=1}^m = T_q^*M$, which means that at every point they span the cotangent space at the point. Using the musical isomorphism, this implies that the riemannian gradients $\text{grad}(z_i)$ span the tangent space at every point: span $\{\{(\text{grad } z_i)(q)\}_{i=1}^m = T_qM$. Using Lemma 5.0.1 can conclude that $\{\text{grad}(z_i)\}$ is a generating set for $\Gamma^\infty(M, TM)$.

5.4.2 Generators defined by gradients of laplacian eigenfunctions

In general, given a function $f \in C^\infty(M)$ on a Riemannian manifold $(M, g)$, its laplacian is defined as the divergence of its Riemannian gradient:

$$\Delta f := \text{div}_g(\text{grad } f).$$ (5.42)

Then the divergence of the fields defined in Theorem 5.1 is given by the laplacian of the functions $z_i : M \rightarrow \mathbb{R}$.

A laplacian eigenfunction $f$ is a smooth function such that

$$\Delta f = \lambda f$$ (5.43)

For some $\lambda \in \mathbb{R}$.

Since any closed Riemannian manifolds has an embedding formed by laplacian eigenfunctions, in this case, using Theorem 5.6 we can build a generating set formed by the Riemannian gradient of laplacian eigenfunctions.

An example of particular importance is given by the hypersphere $S^n = \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\}$. The coordinate functions $z_i(x) = x_i$ form an embedding of laplacian eigenfunctions. This gives us $n+1$ vector fields $\{\text{grad}(z_i)\}_{i=1}^{n+1} \subset \Gamma^\infty(\mathbb{R}^{n+1}, T\mathbb{R}^{n+1})$ tangent to $S^n$ such that their restriction to $M$ forms a generating set for $\Gamma^\infty(S^n, T\mathbb{S}^n)$.

$$\text{grad}(z_i)(x) = e_i - \langle x, e_i \rangle x \quad \forall i \in \{1, \cdots, n+1\}$$ (5.44)

$$\Delta z_i(x) = -nx_i$$ (5.45)

Where $e_1, \cdots, e_{n+1} \in \mathbb{R}^{n+1}$ are the vectors of the canonical basis of $\mathbb{R}^{n+1}$.

---

12 See Bates (2014).
13 Section 1.2 Bates (2014).
5.4.3 Isometrically embedded Submanifolds

Let \((M, g)\) be a Riemannian submanifold of \(\mathbb{R}^m\). We can consider \(\mathbb{R}^m\) as a Riemannian manifold with metric \(\langle \cdot, \cdot \rangle\) given by the scalar product. We say that an embedding \(z : M \to \mathbb{R}^m\) is isometric\(^\text{14}\) if \(z^* \langle \cdot, \cdot \rangle = g\). If the manifold \(M\) is isometrically embedded in \(\mathbb{R}^m\) we have that the restriction\(^\text{15}\) \(\mathcal{T}\mathbb{R}^m|_M := \{(q, v) \in M \times T\mathbb{R}^m : v \in T_q\mathbb{R}^m \}\) decomposes in the orthogonal direct sum:

\[\mathcal{T}\mathbb{R}^m|_M = TM \oplus \mathcal{V}M\]  

(5.46)

Where \(\mathcal{V}M := \{(q, v_q) \in T\mathbb{R}^m|_M : \langle v_q, w_q \rangle = 0 \ \forall w_q \in T_qM\}\) is called the normal bundle. We can then define the orthogonal vector bundle projections:

\[\langle \cdot \rangle^\top = \pi^\top : \mathcal{T}\mathbb{R}^m|_M \to TM\]  

(5.47)

\[\langle \cdot \rangle^\perp = \pi^\perp : \mathcal{T}\mathbb{R}^m|_M \to \mathcal{V}M\]  

(5.48)

**Prop 5.6.1.** Let \(M \subseteq \mathbb{R}^m\) be a isometrically embedded submanifold. Given a point \(q \in M\) there always exist a neighborhood \(q \in \overline{U} \subseteq \mathbb{R}^m\), \(U := \overline{U} \cap M\) and a local smooth orthonormal frame \(\{E_i\}_{i=1}^m \subseteq \Gamma^\infty (\overline{U}, TU)\) such that \(\{E_i\}_{i=1}^n \subseteq \Gamma^\infty (U, TU)\) is a local smooth orthonormal frame for \(TM\) and \(\{E_i\}_{i=n+1}^m \subseteq \Gamma^\infty (U, TU)\) is a local smooth orthonormal frame for \(\mathcal{V}M\)

For isometrically embedded submanifolds, grad\((z_i)\) is simply given by the orthogonal projection of the constant coordinate field \(\partial z_i|_M\) to \(TM\).

**Lemma 5.6.1.** Let \(M \subseteq \mathbb{R}^m\) be a isometrically embedded submanifold with embedding \(z : M \to \mathbb{R}^m\). Then the fields \(\text{grad}(z_i) \in \Gamma^\infty (M, TM)\) are given by the orthogonal projection of the constant coordinate fields \(\partial z_i|_M\) to \(TM\):

\[\text{grad}(z_i) = \pi^\top (\partial z_i|_M)\]  

(5.49)

**Proof.** Fix \(q \in M\), let \(\{E_i\}_{i=1}^m \subseteq \Gamma^\infty (\overline{U}, TU)\) be a local smooth orthonormal frame as described in Proposition 5.6.1. If we denote with \(\{\eta_i\}_{i=1}^m \subseteq \Gamma^\infty (\overline{U}, T^*\overline{U})\) the dual coframe, then \(dz_i|_U = \sum_{k=1}^n E_k(z_i)\eta_k\) and

\[\text{grad}(z_i)|_U = \sum_{k=1}^n E_k(z_i) E_k = \sum_{k=1}^n E_k(\partial z_i) E_k|_U = \left( \sum_{k=1}^m E_k(\partial z_i) E_k\right)^\top\]  

(5.50)

\(^\text{14}\)For definitions, proofs and additional results on isometrically embedded submanifolds see Chapter 8 in Lee (2006)

\(^\text{15}\)The restriction of the tangent bundle \(T\mathbb{R}^m\) to \(M\) is coincides with the pullback bundle \(z^*T\mathbb{R}^m\)
CHAPTER 5. PARAMETERIZING VECTOR FIELDS

Where $\bar{z}$ is the extension of the embedding function $z_i : M \subseteq \mathbb{R}^m \to \mathbb{R}$, given by the $i$-th coordinate projection $z_i(x) = x_i \quad \forall x \in \mathbb{R}^m$. We then have that

$$\sum_{k=1}^{m} E_k(z_i)E_k|_U = \sum_{k=1}^{m} \frac{\partial z_i}{\partial x_k} \frac{\partial z_i}{\partial x_k}|_U = \frac{\partial z_i}{\partial x}|_U \quad (5.51)$$

**Definition 5.6.1.** Let $M \subseteq \mathbb{R}^m$ be a isometrically embedded submanifold and let $\nabla, \nabla'$ denote respectively the Levi-civita connection on $M$ and $\mathbb{R}^m$. The second fundamental form $\mathbb{I}$ is the function

$$\mathbb{I} : \Gamma^\infty(M, TM) \times \Gamma^\infty(M, TM) \to \Gamma^\infty(M, \mathcal{V}M)$$

$$(X, Y) \mapsto \mathbb{I}(X, Y) = (\nabla_X Y) \perp = (\nabla'_X Y) \perp \quad (5.52)$$

Where $\bar{X}, \bar{Y}$ are arbitrary extensions of $X$ and $Y$. The **mean curvature** is defined as the trace of the second fundamental form:

$$H : M \to \mathcal{V}M$$

$$q \mapsto H(q) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(E_i, E_i)_q \quad (5.54)$$

Where $\{E_i\}_{i=1}^{n}$ is a local frame in a neighborhood of $q$.

The second fundamental form measures the difference between the connection on $M$ and the connection of the ambient space $\mathbb{R}^m$:

**Theorem 5.7** (The Gauss formula). Let $M \subseteq \mathbb{R}^m$ be a isometrically embedded submanifold, $X, Y \in \Gamma^\infty(M, TM)$ and $\bar{X}, \bar{Y}$ arbitrary extensions to $\mathbb{R}^m$. The following holds on $M$:

$$\bar{\nabla}_X \bar{Y} = \nabla_X Y + \mathbb{I}(X, Y) \quad (5.56)$$

The Laplacian of the embedding functions $z_i : M \to \mathbb{R}$ is given by the mean curvature$^{16}$:

**Prop 5.7.1.** Let $M \subseteq \mathbb{R}^m$ be a isometrically embedded submanifold with $z : M \to \mathbb{R}^m$ isometric embedding, then:

$$\Delta z = nH \quad (5.57)$$

Where the equality holds elementwise.

$^{16}$Proposition 2.3 Chen and Verstraelen (2013)
Proof. Fix \( q \in M \), let \( \{ E_i \}_{i=1}^n \subseteq \Gamma^\infty (\overline{U}, T\overline{U}) \) a local smooth orthonormal frame as described in Proposition 5.6.1. From Equation (4.40) we know that:

\[
\Delta z_i|_U = \sum_{i=1}^{n} \langle E_j, \nabla_{E_j} (\text{grad}(z_i)) \rangle = \sum_{j=1}^{n} E_j (\langle E_j, \text{grad}(z_i) \rangle) - \langle \nabla z_i, \nabla_{E_j} E_j \rangle =
\]

\[
= \sum_{j=1}^{n} E_j (\langle E_j, \partial x_i \rangle) - \langle \partial x_i, \nabla_{E_j} E_j \rangle = \sum_{j=1}^{n} \nabla_{E_j} ([E_j]_i) - [\nabla_{E_j} E_j]_i =
\]

\[
= \sum_{j=1}^{n} \Pi (E_j, E_j) = nH|_U
\]
Chapter 6

Backpropagation through flows

Given a smooth complete vector field $X \in \Gamma^\infty (M, TM)$, with flow $\{ \phi^t_X \}$, we want to compute the pullback $(\phi^t_X)^*\psi$ for a covector $\psi \in T^* M$. The pullback generalizes the Vector Jacobian Product (VJP) to smooth manifolds.

In general, given a diffeomorphism $\Phi : M \to M$, we can lift it to a diffeomorphism $\Phi^T$ on the cotangent bundle:

$$\Phi^T : T^* M \to T^* M$$

$$p_q \mapsto ((\Phi^{-1})^* p_q)_{\Phi(q)}$$

(6.2)

Where the notation $p_q$ indicates that the point $p \in T^* M$ has base point $q \in M$ (i.e. $p_q \in T^*_q M$).

We are therefore interested in computing $(\phi^t_X)^T$. One key property that we will use is that the lift $\Phi^T$ is a symplectomorphism, this means that it pulls back the canonical symplectic form $\alpha \in \Gamma^\infty (T^* M, \Lambda^2 T^* T^* M)$ to itself:

$$(\Phi^T)^* \alpha = \alpha$$

(6.3)

In the next section, we provide a minimal primer on symplectic geometry, defining only the objects that are essential for us. For a more comprehensive introduction on symplectic geometry, we refer the reader to Da Silva (2001).

6.1 A short primer on symplectic geometry

**Definition 6.0.1** (Symplectic form). Let $M$ be a smooth manifold. A smooth 2-form $\alpha \in \Gamma^\infty (M, \Lambda^2 T^* M)$ which is closed and nondegenerate is called a symplectic form.
The form $\alpha$ is nondegenerate if for every $p \in M$ the linear map:

$$\tilde{\alpha}_p : T_p M \to T_p^* M$$

is an isomorphism.

We say that $\alpha$ is closed if $d\alpha = 0$ where $d$ represents the exterior derivative.

**Definition 6.0.2 (Symplectic manifold).** A symplectic manifold is a couple $(M, \alpha)$ where $M$ is a smooth manifold and $\alpha$ is a symplectic form on $M$.

A direct consequence of the definition is that every symplectic manifold is even dimensional. Given symplectic $\alpha$ form on a manifold $M$, the map $\tilde{\alpha} : TM \to T^* M$ denotes the induced vector bundle isomorphism, whose action on each fiber is given by Definition 6.0.1.

**Definition 6.0.3 (Symplectomorphism).** Let $(M_2, \alpha_1)$ and $(M_2, \alpha_2)$ symplectic manifolds, and let $\Phi : M_1 \to M_2$ be a diffeomorphism. Then $\Phi$ is a symplectomorphism if $\Phi^* \alpha_2 = \alpha_1$.

The simplest example of symplectic manifold is the real $2n$-dimensional space $\mathbb{R}^{2n}$ with coordinates $x_1, \ldots, x_n, y_1, \ldots, y_n$, endowed with the symplectic form:

$$\alpha_0 = \sum_{i=1}^{n} dx_i \wedge dy_i$$

The Darboux theorem tells us that any $2n$-dimensional symplectic manifold is locally symplectomorphic to $(\mathbb{R}^{2n}, \alpha_0)$.

**Theorem 6.1 (Darboux).** Let $(M, \alpha)$ a $2n$-dimensional symplectic manifold. Then given a point $q \in M$, there exists a chart $(U; x_1, \ldots, x_n, y_1, \ldots, y_n)$ centered at $q$ such that:

$$\alpha|_U = \sum_{i=1}^{n} dx_i \wedge dy_i$$

Such charts are called Darboux charts.

We can give to the cotangent bundle $T^*M$ a natural symplectic structure.

**Definition 6.1.1 (Tautological 1-form).** Let $M$ a smooth manifold and $T^*M$ its cotangent bundle. Consider the natural projection $\pi : T^*M \to M$, $p_q \mapsto \pi(p_q) = q$, the pullback $\theta := \pi^* : T^*M \to T^*T^*M$ defines a 1-form on $T^*M$ called the tautological 1-form.

For $p \in T^*M$, $\theta_p$ can be defined through its action on a vector $v \in T_p T^*M$:

$$\theta_p(v) = [\pi^* p]v = p(d\pi_p(v)) \quad \forall v \in T_p T^*M$$
CHAPTER 6. BACKPROPAGATION THROUGH FLOWS

The tautological 1-form admits a simple expression in coordinates. Consider \( q \in M \) and a smooth chart \((U; x_i)\) centered at \( q \). The differentials \( (dx_i|_q)_{i \in [n]} \) form a basis for \( T^*_q M \). This means that if we have \( p_q \in T^*_q M \) there exist coefficients \( \xi_1, \ldots, \xi_n \) such that \( p_q = \sum_{i=1}^{n} \xi_i(dx_i|_q) \). This gives us a chart \((T^*U; x_i, \xi_i)\) for the cotangent bundle, called cotangent coordinates. With respect to these coordinates the tautological 1-form has the expression:

\[
\theta|_{T^*U} = \sum_{i=1}^{n} \xi_i \, dx_i
\]

**Definition 6.1.2 (Canonical symplectic form).** Let \( M \) be a smooth manifold. Then the cotangent bundle \( T^*M \) is a symplectic manifold with symplectic form \( \alpha := -d\theta \), where \( \theta \) is the tautological 1-form. \( \alpha \) is called the canonical symplectic form. The expression of \( \alpha \) in cotangent coordinates is:

\[
\alpha|_{T^*U} = \sum_{i=1}^{n} dx_i \wedge d\xi_i.
\]

This shows that the cotangent coordinates give a Darboux chart for \( T^*M \).

**Prop 6.1.1.** Let \( M \) be a smooth manifold and \( \Phi : M \to M \) a diffeomorphism. Consider the lift \( \Phi^T : T^*M \to T^*M \) to the cotangent bundle, then \( (\Phi^T)^* \theta = \theta \) and \( (\Phi^T)^* \alpha = \alpha \). This means that \( \Phi^T \) preserves the tautological 1-form \( \theta \) and the canonical symplectic from \( \alpha \), and therefore is a symplectomorphism.

**Proof.** Let’s first show that \( \forall p \in T^*M \) it holds \( ( (\Phi^T)^* \theta )_p = \theta_p \). Given \( v \in T_p T^*M \), using the definition of pullback and of the tautological 1-form

\[
\left( (\Phi^T)^* \theta \right)_p (v) = \theta_{\Phi^T(p)} \left( d(\Phi^T)_p v \right) = \Phi^T (p) \left( d\pi_{\Phi^T(p)} \circ d\pi_p v \right)
\]

Now, since \( \Phi^T \) is a lift of \( \Phi \) we have the following commutative diagram:

\[
\begin{array}{ccc}
T^*M & \xrightarrow{\Phi^T} & T^*M \\
\pi \downarrow & & \downarrow \pi \\
M & \xrightarrow{\Phi} & M
\end{array}
\]

Therefore \( d\Phi \circ d\pi = d\pi \circ d(\Phi^T) \) and:

\[
\Phi^T (p) \left( d\pi_{\Phi^T(p)} \circ d(\Phi^T)_p v \right) = \Phi^T (p) \left( d\pi_p v \right) = \theta_p (v)
\]

\[
p \left( d(\Phi^{-1})_{\Phi(p)} \circ d\Phi_{\pi(p)} \circ d\pi_p v \right) = p (d\pi_p v) = \theta_p (v)
\]
For the second part of the proof, it is sufficient to use first the fact that the exterior derivative $d$ commutes with pullbacks:

$$(\Phi^T)^* \alpha = (\Phi^T)^* (-d\theta) = -d((\Phi^T)^* \theta) = -d\theta = \alpha$$

\[\square\]

**Definition 6.1.3** (Hamiltonian vector field). Let $(M,\alpha)$ a symplectic manifold and $H : M \to \mathbb{R}$ a smooth function. Since $\alpha$ is degenerate, there exists unique a vector field $X_H$ such that $X_H \lrcorner \alpha = dH$. The vector field $X_H$ is called the **Hamiltonian vector field** with Hamiltonian function $H$.

**Prop 6.1.2.** Let $X_H$ a complete Hamiltonian vector field and $\phi^t_{X_H}$ the global flow that it defines. We have that the flow preserves the symplectic form, i.e. $(\phi^t_{X_H})^* \alpha = \alpha$, $\forall t \in \mathbb{R}$

A vector field whose flow preserves the symplectic form is called **symplectic vector field**. The previous Proposition then tells us that every Hamiltonian vector field is symplectic. The converse is not always true:

**Prop 6.1.3.** A vector field $X$ on a symplectic manifold $(M,\alpha)$ is Hamiltonian if and only if $X \lrcorner \alpha$ is exact.

### 6.2 Cotangent lift

Let us now consider again a complete vector field $X$ and its flow $\{\phi^t_X\}_t$. We can then lift each diffeomorphism $\phi^t_X$ to symplectomorphism $(\phi^t_X)^* : T^*M \to T^*M$ on the cotangent bundle, considered a symplectic manifold with the canonical symplectic form. Using the properties of the pullback and the fact that $\phi$ is a global flow, it can be easily shown that the maps $\{(\phi^t_X)^*\}_{t \in \mathbb{R}}$ define a smooth global flow on the cotangent bundle $T^*M$. We then have that the global flow is induced by a complete vector field $X^{T^*}$. This field is called the **cotangent lift** of the vector field $X$. We therefore have:

$$\phi^t_{X^{T^*}} = (\phi^t_X)^{T^*} \forall t \in \mathbb{R}$$

(6.13)

This means that given the cotangent vector $p_q \in T^*_qM$, to compute the pullback of $p_q$ by $\phi^t_X$ we can solve the Cauchy problem defined by $-X^{T^*}$ with starting point $p_q$.

Since by construction the flow of the cotangent lift is given by a family of symplectomorphisms, then $X^{T^*}$ is a symplectic vector field.
From Chapter 4.2.3 we know that to compute the cotangent lift at every point, we need to differentiate its flow:

\[ X^*_p = X^*_q = \left. \frac{d}{dt} \right|_{t=0} \left( \phi^t_X (p_q) \right) = \left. \frac{d}{dt} \right|_{t=0} \left( \phi^t_X (p_q) \right), \quad \forall p_q \in M \]

(6.14)

However this expression is of little practical use. A global expression for the cotangent lift can be obtained using the symplectic structure of the cotangent bundle. The next theorem proves that \( X^*_p \) is a Hamiltonian vector field:

**Theorem 6.2.** Let \( X \) a vector field on a smooth manifold \( M \) and \( X^*_p \) its cotangent lift on \( T^*M \). The vector field \( X^*_p \) is then a Hamiltonian vector field with respect to the canonical symplectic form on \( T^*M \). The corresponding Hamiltonian function \( H_X \) is:

\[ H_X : T^*M \to \mathbb{R} \]

\[ p_q \mapsto p(X_q) \]

(6.15) \hspace{1cm} (6.16)

**Proof.** We need to prove that \( X^* \lrcorner \alpha \) is exact. By the definition of \( \alpha \) and by Cartan’s magic formula we have

\[ X^* \lrcorner \alpha = -X^* \lrcorner d\theta = d \left( X^* \lrcorner \theta \right) + \mathcal{L}_{X^*} \theta = d \left( X^* \lrcorner \theta \right) \]

Where in the last passage we have used that since \( \phi_{X^*} = (\phi_X)^T \), then by Proposition 6.1.1 \( \mathcal{L}_{X^*} \theta = 0 \).

The Hamiltonian \( H_X \) is therefore \( X^* \lrcorner \theta = \theta (X^*) \). Evaluating the expression at a point \( p_q \in T^*M \) we have

\[ H_X(p_q) = \theta_p \left( \left( X^* \right)_p \right) = p \left( d\pi_p \left( \left( X^* \right)_p \right) \right) = p(X_q) \]

\[ . \]

\[ \square \]

**6.2.1 Using local coordinates**

Now that we know that the cotangent lift is a Hamiltonian vector field, we can compute an explicit expression for it. Let’s begin by calculating an expression for \( X^*_p \) on a local chart \((U; x_i)\). In this chart the vector field will have local expression \( X|_U = \sum_{i=1}^n f_i \partial_{x_i} \), where \( f_i \in C^\infty(U) \). Since \( X^*_p \) is a vector field on \( T^*M \) we can find its components with respect to the frame \( \{\partial_{x_i}, \partial_{\xi_i}\}_{i=1}^n \) adapted to its cotangent coordinates \((T^*U; x_i, \xi_i)\). Since the cotangent lift is
Hamiltonian, we can leverage the fact that in Darboux coordinates a Hamiltonian vector field $Y$ with Hamiltonian $H$ has local expression (Hamilton equations):

$$Y_H|_{T^*U} = \sum_{i=1}^n \left( \frac{\partial H}{\partial \xi_i} \partial_{x_i} - \frac{\partial H}{\partial x_i} \partial_{\xi_i} \right)$$

(6.17)

In our specific case we have that $H_X$ can be written as:

$$H_X|_{T^*U} = \sum_{i=1}^n \xi_i dx_i \left( \sum_{j=1}^n f_j \partial_{x_j} \right) = \sum_{i=1}^n \xi_i f_i$$

(6.18)

Therefore:

$$X^{T^*}|_{T^*U} = \sum_{i=1}^n f_i \partial_{x_i} - \sum_{i=1}^n \left( \sum_{j=1}^n \frac{\partial f_i}{\partial x_j} \xi_j \right) \partial_{\xi_i}$$

(6.19)

Notice that as we expected the expression is linear on the components $\partial_{\xi_i}$ and coincides with $X$ if projected on the components $\partial_{x_i}$. Moreover, this expression is the same as the adjoint equation in Chen et al. (2018). Therefore, for $M = \mathbb{R}^n$, and in local charts, the cotangent lift coincides with the adjoint equation.

### 6.2.2 Using a generating set

Suppose now we are parameterizing vector fields on $M$ using a generating set $\{X_i\}_{i=1}^m \subset \Gamma^\infty(M, TM)$:

$$X = \sum_{i=1}^m f_i X_i \quad f_i \in C^\infty(M)$$

(6.20)

In this case, the cotangent lift can be constructed from the lifts of the elements of the generating set $\{X_i^{T^*}\}_{i=1}^m$. First, we rewrite the Hamiltonian of the vector field as a linear combination of the Hamiltonians of the fields in the generating set:

$$H_X(p_q) = p(X_q) = p \left( \sum_{i=1}^m f_i X_i \right) \sum_{i=1}^m f_i (q) p \left( (X_i)_q \right) =$$

$$= \sum_{i=1}^m f_i (q) H_{X_i}(p_q) \quad \forall p \in T^*_q M, \forall q \in M$$
CHAPTEER 6. BACKPROPAGATION THROUGH FLOWS

therefore:

\[ H_X = \sum_{i=1}^{m} \pi^*(f_i) H_{X_i} \quad \text{and} \quad dH_X = \sum_{i=1}^{m} \pi^*(df_i) H_{X_i} + \sum_{i=1}^{m} \pi^*(f_i) dH_{X_i} \]

The cotangent lift will then be:

\[ X^{T^*} = \sum_{i=1}^{n} \tilde{\alpha}^{-1}(\pi^* df_i) H_{X_i} + \sum_{i=1}^{m} f_i X_i^{T^*} \quad (6.21) \]

Where the vector fields \( \tilde{\alpha}^{-1}(\pi^* df_i) \) are obtained first pulling back the 1-form \( df_i \) on \( M \) to a 1-form on \( T^*M \) via the pullback of the standard projection, and then mapped to a vector field on \( T^*M \) via the bundle isomorphism \( \tilde{\alpha} \). A simple calculation in cotangent coordinates shows that the vector fields \( \tilde{\alpha}^{-1}(\pi^* df_i) \) belong to the vertical bundle \( VT^*M = \{ v \in TT^*M | v \in \ker(d\pi), \pi : T^*M \to M \} \) of the cotangent space:

\[
\pi^* df_i \bigg|_{T^*U} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} dx_j
\]

\[
\Rightarrow \tilde{\alpha}^{-1}(\pi^* df_i) \bigg|_{T^*U} = - \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \partial \xi_j
\]

6.2.3 Using a local frame

Consider a local frame of smooth vector fields \( \{E_i\}_{i=1}^{n} \subset \Gamma^\infty (M, TM) \) over an open domain \( U \subseteq M \), and its dual coframe \( \{\eta_i\}_{i=1}^{n} \subset \Gamma^\infty (M, T^*M) \). From these frames we can define a local frame for the cotangent bundle over the domain \( T^*U \):

**Prop 6.2.1.** Let \( M \) be a smooth \( n \)-dimensional manifold, and \( \{E_i\}_{i=1}^{n} \) a local frame over the open domain \( U \subseteq M \) and \( \{\eta_i\}_{i=1}^{n} \) its dual coframe. Then \( Z_1, \ldots, Z_n, E_1^{T^*}, \ldots, E_n^{T^*} \) is a local frame for \( T^*M \) over the open domain \( T^*U \). Where \( E_i^{T^*} = \tilde{\alpha}^{-1}(dH_{X_i}) \forall i \) are the complete lifts of \( E_i \) and \( Z_i := \tilde{\alpha}^{-1}(\pi^* \eta_i) \forall i \) are vertical vector fields.

**Proof.** Fix \( p \in T^*U \) with base point \( q := \pi(p) \in U \), then for the first part is sufficient to prove that the fields \( \{Z_i, E_i^{T^*}\}_i \) span all \( T_p T^*U \). This can be

\footnote{See Alekseevsky et al. (1994)}
done in local coordinates. Let \((V, x_i)\) a local chart where \(q \in V \subseteq U\). In local cotangent coordinates we can write:

\[ E_i|_{T^*V} = \sum_{i=1}^{n} a_{ij} \partial_{x_j}, \quad \forall i = 1 \cdots n \]

\[ \eta_i|_{T^*V} = \sum_{i=1}^{n} b_{ij} dx_i, \quad \forall i = 1 \cdots n \]

Since the \(\{E_i\}_{i=1}^{n}\) give local frame the for every point \((x_1, \cdots, x_n)\) the matrix \(A(x_1, \cdots, x_n)\) and \(B(x_1, \cdots, x_n)\) defined as \(A(x_1, \cdots, x_n)_{ij} := a_{ij}(x_1, \cdots, x_n)\), \(B(x_1, \cdots, x_n)_{ij} := b_{ij}(x_1, \cdots, x_n)\) are invertible one and one is the inverse of the other. Using Equation (6.19) we have:

\[ E_i^{T}|_{T^*V} = \sum_{j=1}^{n} a_{ij} \partial_{x_j} - \sum_{j=1}^{n} \left( \sum_{k=1}^{n} \frac{\partial a_{ij}}{\partial x_k} \xi_k \right) \partial_{\xi_j} \]  

(6.22)

And:

\[ Z_i = -\sum_{j=1}^{n} b_{ij} \partial_{\xi_j} \]  

(6.23)

Therefore if we define \(A := A(x_1(q), \cdots, x_n(q))\) and \(B := B(x_1(q), \cdots, x_n(q))\) we have:

\[ \begin{pmatrix} A & * \\ 0 & B \end{pmatrix} \begin{pmatrix} \partial_{x_i} |_p \\ \partial_{\xi_i} |_p \end{pmatrix} = \begin{pmatrix} (E_i^{T})_p \\ (Z_i)_p \end{pmatrix} \]  

(6.24)

Since both \(A\) and \(B\) are invertible, the above block matrix is invertible, therefore \(\{(Z_i)_p, (E_i^{T})_p\}\) span all \(T_p T^* U\). The fact that the fields \(Z_i\) are vertical follows immediately form their expression in local coordinates. \(\square\)

Now suppose we have a vector field \(X \in \Gamma^\infty(M, TM)\), with local expression \(X|_U = \sum_{i=1}^{n} f_i E_i\), we can find the local expression for \(X^{T^*}\) with respect to the local frame \(\{Z_i, E_i^{T^*}\}_i\). First, since the fields \(\{E_i\}_i\) form locally a generating set, we can apply Equation (6.19):

\[ X^{T^*}|_{T^* U} = \sum_{i=1}^{n} \alpha_i^{-1} (\pi^* df_i) + \sum_{i=1}^{n} \pi^* f_i E_i^{T^*} \]  

(6.25)

We can then locally rewrite \(df|_U = \sum_{j=1}^{n} E_j(f_j)\eta_i\), combining this expression with the definition of the fields \(Z_i\) we have:

\[ X^{T^*}|_{T^* U} = \sum_{j=1}^{n} \left( \sum_{i=1}^{n} E_j(f_i) H_{E_i} \right) Z_j + \sum_{i=1}^{n} \pi^* f_i E_i^{T^*} \]  

(6.26)
Where $H_{E_i}(p_q) = p \left( (E_i)_q \right)$ is just $i$-th coordinate of $p_q$ with respect to the basis $\{\eta_i|_q\}_{i=1}^m$

### 6.2.4 Parallelizable manifolds

If a manifold $M$ is parallelizable, its tangent and cotangent bundle can be parameterized using $M \times \mathbb{R}^n$. We can find an explicit parameterization using a global frame $\{E_i\}_{i=1}^n \subset \Gamma^\infty(M, TM)$ and its dual global coframe $\{\eta_i\}_{i=1}^n \subset \Gamma^\infty(M, T^*M)$:

$$M \times \mathbb{R}^n \rightarrow T^*M$$

$$(q, \beta) \mapsto \sum_{i=1}^n \beta_i \eta_i|_q$$

With this parameterization the vertical vector fields $Z_i$ become the coordinate fields:

$$Z_i = -\partial_{\beta_i}$$

We can then rewrite the cotangent lifts $\{E_i^{T^*}\}_{i=1}^n$ with respect to the global frame $\{E_i, \partial_{\beta_i}\}_{i=1}^n$ as:

$$E_i^{T^*} = E_i + \sum_{j=1}^n h_{ij} \partial_{\beta_i}$$

Where $h_{ij} \in C^\infty(M \times \mathbb{R}^n)$ is linear in the second argument. Equation (6.26) can be now rewritten in the following form:

$$X^{T^*} = -\sum_{j=1}^n \left( \sum_{i=1}^n E_j(f_i) \beta_i \right) \partial_{\beta_j} + \sum_{i=1}^n f_i E_i^{T^*} =$$

$$= \sum_{j=1}^n \left( \sum_{i=1}^n f_i h_{ij} - E_j(f_i) \beta_i \right) \partial_{\beta_j} + \sum_{i=1}^n f_i E_i$$

From the last expression we observe that for parallelizable manifolds, finding the cotangent lift corresponds with augmenting the initial vector field with a linear vector field on $\mathbb{R}^n$.
6.2.5 Cotangent lift on Lie Groups

Let $G$ be a Lie group with Lie algebra $\mathfrak{g}$. Since $G$ is parallelizable we can identify $T^*G$ with $G \times \mathfrak{g}^*$ via the following isomorphism:

$$G \times \mathfrak{g}^* \rightarrow T^*G$$

$$(a, \psi) \mapsto L_{a^{-1}}^* \psi$$

(6.33)

(6.34)

**Theorem 6.3.** Let $G$ be a Lie group with Lie algebra $\mathfrak{g}$. If we identify, as described above, $T^*G \cong G \times \mathfrak{g}^*$ then given $v \in \mathfrak{g}$ the cotangent lift of the associated left-invariant vector field is:

$$(v^L)^T = (v^L, \text{ad}^*_v)$$

(6.35)

Where $\text{ad}^*_v$ is the coadjoint action, defined as the transpose of the adjoint action $\text{ad}_v : \mathfrak{g} \rightarrow \mathfrak{g}, w \mapsto [v, w]$, and since $\mathfrak{g}^*$ is a vector space we identified $T^*_\psi \mathfrak{g}^*$ with $\mathfrak{g}^*$ for all $\psi \in \mathfrak{g}^*$.

**Proof.** Let’s first prove that the flow of the cotangent lift is given by:

$$\left(\phi^t_{v^L}\right)^T = \left(R_{\exp(vt)}, \text{Ad}^*_{\exp(vt)}\right)$$

(6.36)

We already know from Equation (5.8) that a left invariant vector field $v^L$ has flow $\phi^t_{v^L} = R_{\exp(vt)}$. Therefore his cotangent lift is a diffeomorphism on $T^*G$ is given by $R_{\exp(-vt)}^*$. Now identifying $T^*G$ via the map in Equation (6.33) we have:

$$\left(\phi^t_{v^L}\right)^T(a, \psi) = \left(R_{\exp(vt)}a, L_{a\exp(vt)}^* \circ R_{\exp(-vt)}^* \circ L_{a^{-1}}^* \psi\right) =$$

(6.37)

$$= \left(R_{\exp(vt)}a, L_{\exp(vt)}^* \circ R_{\exp(-vt)}^* \psi\right) =$$

(6.38)

$$= \left(R_{\exp(vt)}a, \text{Ad}^*_{\exp(vt)} \psi\right)$$

(6.39)

Where we have used that the left and the right multiplications commute. The thesis now follows taking the limit for $t \rightarrow 0$. 

Let now $\{v_i\}_{i=1}^n \subset \mathfrak{g}$ be a basis for the lie algebra, and $\{\eta_i\}_{i=1}^n \subset \mathfrak{g}^*$ its dual basis. The basis defines a global frame of left invariant vector fields $\{v_i^L\}_{i=1}^n$. Let $V = \sum_{i=1}^n f_i v_i^L$ a vector field on $G$ parameterized by the smooth function $f : G \rightarrow \mathbb{R}^n$. In order to find an expression for its cotangent lift, we need to

---

2See Proposition 1.3 in Ayala et al. (2009).
compute the functions \( \{ h_{ij} \}_{i,j=1}^n \) present in Equation (6.30). From Theorem 6.3 we then have that

\[
\begin{align*}
    h_{ij}(a, \beta) &= \text{ad}_{v_i}^* \left( \sum_{k=1}^n \beta_k \eta_k \right) (v_j) = \sum_{k=1}^n c_{ij}^k \beta_k
\end{align*}
\] (6.40)

Where \( c_{ij}^k \) are the structure constants of for \( \{ v_i \}_{i=1}^n \) where \( [v_i, v_j] = \sum_{k=1}^n c_{ij}^k v_k \).

The cotangent lift of \( V \) is then:

\[
\begin{align*}
    V^T = \sum_{j=1}^n \left( \sum_{i=1}^n \left( \sum_{k=1}^n f_i c_{ij}^k \beta_k - \sum_{i=1}^n E_j(f_i) \beta_i \right) \right) \partial_{\beta_j} + \sum_{i=1}^n f_i E_i
\end{align*}
\] (6.41)

Which can be suggestively rewritten as:

\[
\begin{align*}
    V^T = \text{ad}^*_{\sum_{i=1}^n f_i v_i} - \sum_{j=1}^n \left( \sum_{i=1}^n v_j(f_i) \beta_i \right) \partial_{\beta_j} + \sum_{i=1}^n f_i E_i
\end{align*}
\] (6.42)

### 6.2.6 Cotangent lift on embedded submanifolds

Let \( M \) be a properly embedded smooth submanifold of \( \mathbb{R}^m \), and let \( \iota : M \hookrightarrow \mathbb{R}^m \) denote the inclusion map. Consider a smooth vector field \( X \in \Gamma^\infty(M, TM) \) and \( \overline{X} \in \Gamma^\infty(\mathbb{R}^m, T\mathbb{R}^m) \) a tangent vector field that extends \( X \). We first observe that since the \( \overline{X} \) and \( X \) are \( \iota \)-related their flows commute with the inclusion map. We, therefore, have that the following diagram commutes:

\[
\begin{array}{ccc}
\mathbb{R}^m & \xrightarrow{\phi^t_X} & \mathbb{R}^m \\
\uparrow \iota & & \uparrow \iota \\
M & \xrightarrow{\phi^t_X} & M
\end{array}
\] (6.43)

Suppose now that we are interested in computing the differential of the function \( f \circ \iota \circ \phi_X^t = f \circ \phi_X^t : M \to \mathbb{R} \) where \( f : \mathbb{R}^m \to \mathbb{R} \). We can then both write:

\[
\begin{align*}
    \text{d} \left( f \circ \iota \circ \phi_X^t \right) &= (\phi_X^t)^* \circ \iota^* \circ \text{df} = \phi_X^{t*} \circ \iota^* \circ \text{df} \\
    &= \iota^* \circ (\phi_X^t)^* \circ \text{df} = \iota^* \circ \phi_X^{t*} \circ \text{df}
\end{align*}
\] (6.44)

(6.45)
This means that we can use the cotangent lift of $\overline{X}$ to compute the pullback of cotangent vectors by the flow of $X$.

Consider now the pullback bundle $\iota^*T^*\mathbb{R}^m$ := \{(x, \beta) \in M \times T^*\mathbb{R}^m|\pi(\beta) = x\}$. Since $M$ is a properly embedded submanifold of $\mathbb{R}^m$, we can consider the pullback bundle $\iota^*T^*\mathbb{R}^m = \{(x, \beta) \in T^*\mathbb{R}^m|\pi(\beta) = x, x \in M\} =: T^*\mathbb{R}^m|_M$, as a properly embedded submanifold of $T^*\mathbb{R}^m$ with inclusion map $(\iota, \text{id})$.

Moreover, since the fiber of $T^*\mathbb{R}^m$ is canonically isomorphic to $\mathbb{R}^m$, we have that $T^*\mathbb{R}^m|_M$ is canonically isomorphic (as a vector bundle) to $M \times \mathbb{R}^m$.

Let’s analyze now the cotangent lift $\overline{X}^{T^*}$. Since $\overline{X}$ is tangent to $M$, it’s easy to show that $\overline{X}^{T^*}$ is tangent to $T^*\mathbb{R}^m|_M$.

**Prop 6.3.1.** Let $M$ be a properly embedded smooth submanifold of $\mathbb{R}^m$, and let $\iota : M \hookrightarrow \mathbb{R}^m$ denote the inclusion map. Consider a smooth vector field $X \in \Gamma^\infty(\mathbb{R}^m, TR^m)$ a tangent vector field that extends $X$. Then the cotangent lift is tangent to the restriction bundle $T^*\mathbb{R}^m|_M = \iota^*T^*\mathbb{R}^m = \{(x, \beta) \in T^*\mathbb{R}^m|\pi(\beta) = x, x \in M\}$.

**Proof.** Consider $(x, \beta) \in T^*\mathbb{R}^m|_M \subset T^*\mathbb{R}^m$. Since $(T^*\mathbb{R}^m|_M)_x = (T^*\mathbb{R}^m)_x$ which means that the fiber of $T^*\mathbb{R}^m|_M$ and $T^*\mathbb{R}^m$ are the same, the the only condition that $\overline{X}^{T^*}$ has to satisfy for being tangent to $T^*\mathbb{R}^m|_M$ $(x, \beta)$ is:

$$\pi_{T^*\mathbb{R}^m} \left( \left( \overline{X}^{T^*} \right)_{(x, \beta)} \right) = d\pi_{\mathbb{R}^m} \left( \left( \overline{X}^{T^*} \right)_{(x, \beta)} \right) \in T_xM \subset T_x\mathbb{R}^m \quad (6.46)$$

Where $\pi_{\mathbb{R}^m} : T^*\mathbb{R}^m \to \mathbb{R}^m$ is cotangent bundle projection. Using the definition of cotangent lift we see that:

$$\pi_{T^*\mathbb{R}^m} \left( \left( \overline{X}^{T^*} \right)_{(x, \beta)} \right) = X_x \in T_xM \subset T_x\mathbb{R}^m. \quad (6.47)$$

Where the equality holds because $\overline{X}$ is tangent to $M$.

The restriction of $\overline{X}^{T^*}$ therefore defines the smooth vector field $\overline{X}^{T^*}|_{T^*\mathbb{R}^m|_M}$ on $T^*\mathbb{R}^m|_M$. Then, by taking the pullback of every map in the commutative diagram (6.44), we have that the following diagram commutes:

$$
\begin{array}{ccc}
T^*\mathbb{R}^m|_M & \xrightarrow{\phi^t|_{T^*\mathbb{R}^m|_M}} & T^*\mathbb{R}^m|_M \\
\downarrow \iota^* & & \downarrow \iota^* \\
T^*M & \xrightarrow{\phi^t|_{T^*M}} & T^*M 
\end{array}
(6.48)
$$
Then by Proposition 9.6 of Lee (2013) $X^T_{|T^*\mathbb{R}^m|_M}$ and $X^T_\ast$ are $\iota^\ast$ related. Moreover by Lemma 5.5.2 $\iota^*T^*\mathbb{R}^m = T^*\mathbb{R}^m|_M$ parameterizes all $T^*M$. As previously observed, when working with embedded manifolds functions on $M$ are parameterized by $f \circ \iota$, where $f : \mathbb{R}^m \to \mathbb{R}$. In this case, the above diagram tells us that in order to compute the pullback through $\phi_X^t$ we can solve a ODE on $\iota^*T^*\mathbb{R}^m = T^*\mathbb{R}^m|_M \cong M \times \mathbb{R}^m$. 
Chapter 7

Experiments

In this chapter we show that the framework derived in the previous chapters can be effectively used to model complex multimodal probability distributions on non-trivial manifolds. This is achieved by training a manifold continuous normalizing flow (MCNF) to match mixture distributions on a variety of spaces: the hypersphere $S^n$, the matrix Lie groups $SO(n)$, $U(n)$, $SU(n)$, Stiefel manifolds $V_m(\mathbb{R}^n)$, and the positive definite symmetric matrices $\text{Sym}^+(n)$. For each manifold we specified a mixture distribution with 4 components, adapted to the specific geometry of the space.

7.1 General setup

7.1.1 Density matching on manifolds

The experiments consist of the following: fixing a smooth manifold $M$ and base smooth positive density $\mu$, we train MCNF to match target density $\mu^*$:

$$\mu^* := \frac{\rho^*}{Z} \mu$$

where $\rho^* \in C(M, \mathbb{R}_{\geq 0})$, $Z := \int \rho^* \mu$.

Where $\rho^*$ is a positive continuous function and $Z \in \mathbb{R}$ is the normalization constant. Denoted by $\mu_\lambda := \rho_\lambda \mu$ the density parameterized by the MCNF
model\textsuperscript{1}, the training objective is to minimize the KL divergence:

\[
\text{KL}(\mu_\lambda||\mu^*) = \int (\ln \rho_\lambda - \ln \rho^*) \rho_\lambda \mu - \ln Z = \mathbb{E}_{\mu_\lambda} [\ln \rho_\lambda - \ln \rho^*] - \ln Z. \tag{7.1}
\]

Where \(\ln Z\) is a constant and can therefore be ignored. The performance of the model is then assessed using KL divergence\textsuperscript{2} and effective sampling size\textsuperscript{3} (ESS):

\[
\text{ESS} \approx \left(\sum_{i=1}^{S} w_i\right)^2 \text{ where } w_i = \frac{\rho^*(q_i)}{\rho_\lambda(q_i)}, \text{ and } q_i \sim \rho_\lambda \mu. \tag{7.2}
\]

### 7.1.2 Mixture of densities

Let \(\mu\) be a smooth positive density on a smooth manifold. In each of the experiments we fit a \(k\) component mixture distribution:

\[
\mu^* = \rho^*(\cdot|\beta, \{W_i\}_{i=1}^k) \mu = \sum_{i=1}^{k} \frac{\rho(\cdot|\beta, W_i)}{Z(\beta)} \mu \tag{7.4}
\]

Where \(\rho \in C(M)\) is the component of the distribution, which is given by an unnormalized probability density function with normalization constant \(0 < Z \in \mathbb{R}\). If \(Z(\beta)\) does not depend on the parameters \(W\) of the components then \(\int \rho^* d\mu = 1\) is a probability density function. The parameter \(\beta\) controls the concentration of the component distribution. The log target density is then given by:

\[
\log \rho^*(\cdot|\beta, \{W_i\}_{i=1}^k) = \log \text{sumexp} \log \rho(\cdot|\beta, W_i) \tag{7.5}
\]

In each of the experiments, the function \(\rho\) is adapted to the specific geometry of the space.

### 7.1.3 Implementation details

All the spaces are considered to be embedded in \(\mathbb{R}^m\) and backpropagation was performed using what is observed in Section 6.2.6. We parameterize vector

\textsuperscript{1}Where \(\rho_\lambda \in C(M, \mathbb{R}_{\geq 0}), 1 = \int \rho_\lambda \mu, \text{ and } \lambda \in \mathbb{R}^k\) denotes the model parameters.

\textsuperscript{2}The normalization constant \(Z\) in the KL divergence is estimated as \(Z \approx \sum_{i=1}^{S} w_i\).

\textsuperscript{3}Kish (1965)
fields choosing a suitable generating set of vector fields. Fixed the generating set, the parameterizing function $f \in C^\infty(M; \mathbb{R}^{m_{\text{gen}}})$ used in Equation 5.19 is parameterized by a neural network with input in the embedding space $\mathbb{R}^m \supset M$, $\tanh$ nonlinearity, and 2 hidden layers with dimension: $[5 \cdot m_{\text{gen}}, 5 \cdot m_{\text{gen}}]$, where $m_{\text{gen}}$ is the cardinality of the generating set used. The parameterized flow was numerically integrated with starting time $0$ and final time $1$, using the Dormand-Prince ODE integration algorithm, with adaptive stepsize, as implemented in Bradbury et al. (2018). Each model was optimized using Adam (Kingma and Ba, 2015) with a learning rate of $5 \cdot 10^{-4}$. Divergence during training was approximated using the estimator described in Theorem 5.2 and sample size of 512.

Each model was scored on KL divergence and effective sampling size (ESS) using a sample of 200000 points and the exact divergence expression given by Equation (5.21).

In the next sections, we will focus individually on each of the spaces we used in the experiments. Describing the type of generating set employed and giving the exact expression of the mixture centres used, as well as reporting the experiments results.

### 7.2 Manifold specific setup and results

#### 7.2.1 Special Orthonormal matrices

The group $SO(n) := \{ A \in \text{GL}_n(\mathbb{R}) : A^\top A = I, \ \det(A) = 1 \}$ is the matrix Lie group formed by $n \times n$ orthonormal matrices with unit determinant. Its Lie algebra $\mathfrak{so}(n) := \{ A \in \text{M}_n(\mathbb{R}) : A^\top = -A \}$ is given by the $n \times n$ skew-symmetric matrices. A basis for $\mathfrak{so}(n)$ is given by the $\frac{n(n-1)}{2}$ matrices:

$$V_{ij} = E_{ij} - E_{ji} \quad i < j \quad \forall i, j \in [n] \quad (7.6)$$

The matrix unit $E_{ij} \in \text{M}_nn(\mathbb{R})$ is the matrix with 1 in the $(i, j)$-entry and 0 everywhere else. For example, when $n = 3$ a basis $\{v_i\}_{i=1}^3$ for $\mathfrak{so}(3)$ is given by:

$$v_{1,2,3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (7.7)$$

We use the generating set $\{v_i^L\}_{i=1}^3 \subset \Gamma^\infty(SO(3), TSO(3))$, when parameterizing vector fields on $SO(3)$ in the experiments.
7.2.2 Stiefel Manifold

The Stiefel manifold \( \mathcal{V}_m(\mathbb{R}^n) = \{Q \in M_{nm}(\mathbb{R}) \mid Q^TQ = I_m\} \) is the manifold formed by all orthonormal \( m \)-frames in \( \mathbb{R}^n \). As a special case, for \( m = 1 \) we have: \( S^n = \mathcal{V}_1(\mathbb{R}^{n+1}) \). When \( m < n \) the Stiefel manifold can be considered a homogeneous-\( SO(n) \) space with action given by the left matrix multiplication: \( (A, Q) \mapsto AQ \).

By Proposition, 5.3.1 the basis \( \{V_{ij}\}_{i<j} \subset \mathfrak{so}(n) \) induces a generating set \( \{\widetilde{V}_{ij}\}_{i<j} \subset \Gamma^\infty(\mathcal{V}_m(\mathbb{R}^n), T\mathcal{V}_m(\mathbb{R}^n)) \) for the smooth vector fields on \( \mathcal{V}_m(\mathbb{R}^n) \), which we will use in the experiments when parameterizing vector fields on \( \mathcal{V}_m(\mathbb{R}^n) \):

\[
\widetilde{V}_{kj}(Q) = V_{kj}Q = E_{kj}Q - E_{jk}Q \in T_Q\mathcal{V}_m(\mathbb{R}^n) \subset M_{nm}(\mathbb{R}) \quad \forall Q \in \mathcal{V}_m(\mathbb{R}^n) \tag{7.8}
\]

where \( 1 \leq k < j \leq n \).

Experiments:

The target density used in the experiments for \( \mathcal{V}_m(\mathbb{R}^n) \) and \( SO(n) \) is a \( k = 4 \) mixture of Langevin distribution\(^4\) with components given by:

\[
\log \rho(Q \mid \beta, W) = \frac{\beta}{m} \text{tr} \left( W^TQ \right) \quad \forall Q \in \mathcal{V}_m(\mathbb{R}^n), SO(n) \tag{7.9}
\]

where \( W \in \mathcal{V}_m(\mathbb{R}^n), SO(n) \). For \( SO(n) \) we used \( m := n - 1 \).

As a base density \( \mu \), and as an initial probability density, we used the \( SO(n) \) action invariant density on \( \mathcal{V}_m(\mathbb{R}^n) \) (normalized to 1). Algorithm 1 in Section 4.4.2 of Camaño-García (2006) was employed for sampling.

In the experiments we fixed \( \beta \in \{10\} \) and the centers \( W_i \in \mathcal{V}_m(\mathbb{R}^n), SO(n) \) were sampled uniformly at random for each run. Results are reported in Table 7.1.

7.2.3 Hypersphere

The hypersphere \( \mathbb{R}^n = \{x \in \mathbb{R}^{n+1} \mid \|x\|_2 = 1\} \) is the set of points in \( \mathbb{R}^{n+1} \) with unit norm.

\(^4\) See Section 3.2 in Camaño-García (2006)
Table 7.1: Evaluation of Manifold Continuous Normalizing Flow (MCNF) on $\mathcal{V}_{m}(\mathbb{R}^n)$ and $SO(n)$. Error is computed over 5 replicas of each experiment.

| Manifold       | $\beta$ | KL [nats] | ESS [%] |
|----------------|---------|-----------|---------|
| $\mathcal{V}_1(\mathbb{R}^3) \cong S^2$ | 10       | 0.003±0.001 | 99.3±0.8 |
| $\mathcal{V}_1(\mathbb{R}^4) \cong S^3$ | 10       | 0.006±0.002 | 98.6±0.5 |
| $\mathcal{V}_2(\mathbb{R}^4)$ | 10       | 0.02±0.006  | 96.5±1.1 |
| $\mathcal{V}_2(\mathbb{R}^5)$ | 10       | 0.02±0.003  | 97.1±0.6 |
| $SO(3)$        | 10       | 0.03±0.007  | 95.3±0.8 |

As we already saw, we can consider the hypersphere $S^n = \mathcal{V}_1(\mathbb{R}^{n+1})$ as a homogeneous $SO(n + 1)$-space. And therefore build a generating set of vector fields given by infinitesimal $SO(n + 1)$ actions. However, as we saw before, the cardinality of this generating set is $\frac{n(n+1)}{2}$, number that grows quadratically with the dimension of the manifold, making it inefficient for building normalizing flows on high dimensional hyperspheres.

Alternatively, we can use the generating set induced by the isometric embedding $z : S^n \hookrightarrow \mathbb{R}^{n+1}$, given by the coordinate functions $z_i(x) = x_i$ and described by in Section 5.4.2. Fortunately, the embedding components are laplacian eigenfunctions. The divergence of the gradient fields is then given by a multiple of the embedding functions.

This gives us $n+1$ vector fields $\{\text{grad}(z_i)\}_{i=1}^{n+1} \subset \Gamma^\infty(\mathbb{R}^{n+1}, T\mathbb{R}^{n+1})$ tangent to $S^n$ such that their restriction to $S^n$ forms a generating set for $\Gamma^\infty(S^n, T\mathbb{S}^n)$:

$$
\text{grad}(z_i)(x) = e_i - (x, e_i)x \quad \forall i \in [n+1] \tag{7.10}
$$

$$
\text{div}_{\mu_g}(\text{grad}(z_i)) = \Delta z_i(x) = -nx_i \tag{7.11}
$$

Where $\mu_g$ is the Riemannian density given by the rotation invariant metric on $S^n$. We have therefore built a generating set with cardinality $n + 1$, which scales linearly with the dimension of the space, and with a simple divergence expression. We use this generating set for $S^n$ in the experiments when scaling to higher dimensions.

---

5Section 1.2 Bates (2014).
6See Section 5.4.2
Experiments

The target density used in the experiments for $\mathbb{S}^n$ is a $k = 4$ mixture of von Mises Fisher (vMF) distributions, with components given by:

$$\log \rho(Q|\beta, W) = \beta \cdot W^\top Q \quad \forall Q \in \mathbb{S}^n \quad \text{where } W \in \mathbb{S}^n$$  \hspace{1cm} (7.12)

As a base density $\mu$, and as an initial probability density, we used the $SO(n)$ action invariant density (normalized to 1) on $\mathbb{S}^n$.

We employed the generating set given by Equation 7.10 for parameterizing vector fields on $\mathbb{S}^n$.

In the experiments we fixed $n \in \{10, 20\}$ add $\beta \in \{10, 20, 50\}$ and the centers $W_i \in \mathbb{S}^n$ were sampled uniformly at random for each run. Results are reported in Table 7.2.

Additional experiment: comparing with previous normalizing flows on spheres

Additionally, we decided to compare our MCNF against the normalizing flow for hyperspheres proposed in Rezende et al. (2020). We, therefore, trained a MCNF to match the same vMF mixture used in the experiments by Rezende et al. (2020) and compared against the best result provided there. Results are reported in Table 7.3, while Figure 7.1 shows the leaned density on $\mathbb{S}^2$.

We observe that the proposed MCNF model can closely match the target densities, with a considerably lower KL divergence and higher ESS than the model by Rezende et al. (2020).

Table 7.2: Evaluation of Manifold Continuous Normalizing Flow (MCNF) on $\mathbb{S}^n$. The experiments on the Highlighted rows were repeated 5 times and reported in the summary Table 7.7

| Manifold | $\beta$ | KL[nats] | ESS [%] |
|----------|---------|----------|---------|
| $\mathbb{S}^{10}$ | 10 | 0.006 | 98.7 |
| $\mathbb{S}^{10}$ | 20 | 0.01±0.003 | 97.2±0.6 |
| $\mathbb{S}^{10}$ | 50 | 0.02 | 95.0 |
| $\mathbb{S}^{20}$ | 10 | 0.01 | 97.32 |
| $\mathbb{S}^{20}$ | 20 | 0.02±0.003 | 95.4±0.4 |
(a) Target  
(b) Model

Figure 7.1: Learned density on $S^2$

| Manifold | Model | KL[nats] | ESS[%] |
|----------|-------|----------|--------|
| $S^2$    | MS    | 0.05±.01 | 90     |
|          | MCNF  | 0.003±.001 | 99.3±.2 |
| $S^3$    | MS    | 0.14     | 84     |
|          | MCNF  | 0.004±.001 | 99.2±0.2 |

Table 7.3: Comparing of MCNF and MS flow proposed in Rezende et al. (2020) on a vMF density matching task. Error is computed over 3 replicas of each experiment.

7.2.4 Unitary matrices

The group $U(n) := \{ A \in \text{GL}_n(\mathbb{C}) : A^\dagger A = I \}$ is the matrix Lie group formed by all the complex $n \times n$ unitary matrices. Its Lie algebra $\mathfrak{u}(n) = \{ A \in \text{M}_{nn}(\mathbb{C}) | A^\dagger = -A \}$ is given by the $n \times n$ skew-Hermitian matrices. A basis for $\mathfrak{u}(n)$ is given by the $n^2$ matrices:

\[
V_{kj} := E_{kj} - E_{jk} \quad iV^\dagger_{kj} := iE_{kj} + iE_{jk} \quad H_j := iE_{jj} \quad \text{where } 1 \leq k < j \leq n
\]  

(7.13)
where $1 \leq k < j \leq n$ and $i$ denotes the imaginary unit. For example for $n = 2, 3$ a basis $\{v_k\}_{k=1}^{n^2}$ is given by:

$$v_{1,2,3,4} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} i & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & i \end{bmatrix}.$$  \hspace{1cm} (7.14)

$$v_{1,2,3,4,5,6,7,8,9} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & i \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$  \hspace{1cm} (7.15)

We use the generating set $\{v^L_k\}_{k=1}^{n^2} \subset \Gamma^\infty (U(n), TU(n))$, when parameterizing vector fields on $U(2)$ and $U(3)$ in the experiments.

### 7.2.5 Special Unitary matrices

The group $SU(n) := \{ A \in GL_n(C) : A^t A = I, \det(U) = 1 \}$ is the matrix Lie group of complex $n \times n$ unitary matrices with unit determinant. The lie algebra $su(n) := \{ A \in M(n,C) : A^t = -A, \tr(A) = 0 \}$ is given by the traceless skew-Hermitian matrices. A basis for $su(n)$ is given by the $n^2 - 1$ matrices:

$$V_{kj} = E_{kj} - E_{jk} \hspace{0.5cm} iV^\dagger_{kj} = iE_{kj} + iE_{jk} \hspace{0.5cm} H_k = iE_{kk} - iE_{k+1k+1}$$  \hspace{1cm} (7.17)

where $1 \leq k < j \leq n$. When $n = 2$ a basis $\{v_i\}_{i=1}^3$ for $su(2)$ is given by Pauli matrices:

$$v_{1,2,3} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}.$$  \hspace{1cm} (7.18)

When $n = 3$ a basis $\{v_i\}_{i=1}^8$ for $su(3)$ is:

$$v_{1,2,3,4,5,6,7,8} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$  \hspace{1cm} (7.19)

$$\begin{bmatrix} 0 & 0 & i \\ 0 & i & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$  \hspace{1cm} (7.20)
Table 7.4: Evaluation of Manifold Continuous Normalizing Flow (MCNF) on $U(n)$ and $SU(n)$. The experiments on the Highlighted rows were repeated 5 times and reported in the summary Table 7.7

| Manifold | $\beta$ | KL[nats] | ESS[\%] |
|----------|---------|----------|----------|
| $SU(2)$  | 5       | 0.001    | 99.9     |
| $SU(2)$  | 10      | 0.005$\pm$0.003 | 98.6$\pm$0.8 |
| $SU(3)$  | 5       | 0.003    | 99.3     |
| $SU(3)$  | 10      | 0.01$\pm$0.003 | 97.1$\pm$0.6 |
| $U(2)$   | 5       | 0.003    | 99.3     |
| $U(2)$   | 10      | 0.02$\pm$0.008 | 95.4$\pm$1.4 |
| $U(3)$   | 5       | 0.006    | 98.9     |
| $U(3)$   | 10      | 0.05$\pm$0.008 | 91.5$\pm$1.3 |

Experiments:

The target density used in the experiments for $U(n)$ and $SU(n)$ is a $k=4$ mixture of distributions with components given by:

$$\log \rho(Q|\beta, W) = \frac{\beta}{n} \text{tr} \left( \text{Real} \left(W^t Q \right) \right) \quad \forall Q \in U(n), SU(n)$$  \hspace{1cm} (7.21)

where $W \in U(n), SU(n)$  \hspace{1cm} (7.22)

As a base density $\mu$, and as an initial probability density, we used the bi-invariant density associated to the Haar measure on $SU(n), U(n)$. The algorithm described in Mezzadri (2006) was employed for sampling.

In the experiments we fixed $n \in \{2, 3\}, \beta \in \{5, 10\}$ and the centers $W_i \in U(n), SU(n)$ were sampled uniformly at random for each run. Results are reported in table 7.4.

Additional experiment: matching conjugation invariant densities on $SU(3)$

To further investigate the ability of MCNF to model multimodal densities, we trained a model to match the following conjugation invariant probability density on $SU(3)$:

$$\log \tilde{\rho}(A|\beta, c) = \frac{\beta}{3} \text{Real} \left( \text{tr} \left( \sum_{j=1}^{3} c_j U_j^2 \right) \right)$$  \hspace{1cm} (7.23)
Table 7.5: ESS of Manifold Continuous Normalizing Flow (MCNF) on conjugation invariant density matching task on $SU(3)$. Results are compared with conjugation equivariant flow (CEF) defined by Kanwar et al. (2020).

| Setup       | Model  | $\beta = 1$ | $\beta = 5$ | $\beta = 9$ |
|-------------|--------|-------------|-------------|-------------|
| $c = (0.17, -0.65, 1.22)$ | CEF    | 97          | 80          | 82          |
|             | MCNF   | 98.3        | 91.2        | 83.5        |
| $c = (0.98, -0.63, -0.21)$ | CEF    | 99          | 91          | 73          |
|             | MCNF   | 99.5        | 98.5        | 97.4        |

Where $c = (c_1, c_2, c_3) \in \mathbb{R}^3$. Results are reported in Table 7.5. As a reference, we report the scores achieved by the conjugation equivariant flow described in Kanwar et al. (2020). Notice that a throughout comparison of the models might be inappropriate since the permutation equivariant flow already incorporates the symmetries of the target distribution. Thus reducing to modelling a flow on the two-dimensional simplex $\Delta_2$. Our model instead parameterizes an unconstrained density on $SU(3)$. Notwithstanding the above a MCNF, can match the performance of the conjugation equivariant flow in all settings, achieving in some cases significantly better results. We interpret these results as additional evidence of the high expressive power of our model.

### 7.2.6 Positive definite symmetric matrices

The manifold $\text{Sym}^+(n) = \{Q \in M_{nn}(\mathbb{R}) \mid Q^T = Q, \quad x^T Q x > 0 \quad \forall x \in \mathbb{R}^n\}$ is formed by all $n \times n$ symmetric positive definite matrices. $\text{Sym}^+(n)$ is a convex cone on the manifold of symmetric matrices $\text{Sym}(n) \cong \mathbb{R}^{n(n+1)/2}$.

$\text{Sym}^+(n)$ can be considered a homogeneous $GL(n)$ space with action given by matrix congruence:

$$GL(n) \times \text{Sym}^+(n) \to \text{Sym}^+(n)$$

$$(A, Q) \mapsto A^T QA$$

By Proposition 5.3.1 the basis $\{E_{kj}\}_{k,j=1}^n \subset \mathfrak{gl}(n)$ induces the generating set $\{\tilde{E}_{kj}\}_{k,j=1}^n \subset \Gamma^\infty \left(\text{Sym}^+(n), T\text{Sym}^+(n)\right)$ for the smooth vector fields on
CHAPTER 7. EXPERIMENTS

\[ \tilde{E}_{kj}(Q) = \frac{d}{dt} \bigg|_{t=0} (I + E_{jk}t) Q (I + E_{kj}t) + o(t) \]

(7.26)

\[ = E_{jk}Q + QE_{jk} \in T_Q \text{Sym}^+(n) \cong \text{Sym}(n) \subset M_{nn}(\mathbb{R}) \quad \forall Q \in \text{Sym}^+(n) \]

(7.27)

(where \( k, j \in [n] \)), which we will use in the experiments when parameterizing vector fields on \( \text{Sym}^+(n) \). Notice that we can identify

\[ T_Q \text{Sym}^+(n) \cong \text{Sym}(n) := \{ Q \in M_{nn}(\mathbb{R}) \mid Q^\top = Q \}, \quad \forall Q \in \text{Sym}^+(n). \]

We can define on \( \text{Sym}^+(n) \) the following group action invariant Riemannian metric:

\[ g_Q(A, B) = \text{tr} \left( Q^{-1} A Q^{-1} B \right) \]

(7.28)

The corresponding group invariant volume density \( \mu_g \) is given by

\[ \mu_g(Q) = \det(Q)^{-\frac{\beta}{2} - \frac{n+1}{2}} \bigg| \bigwedge_{1 \leq k \leq j \leq n} dQ_{kj} \bigg| \quad \forall Q \in \text{Sym}^+(n) \]

(7.29)

Where \( \bigwedge_{1 \leq k \leq j \leq n} dQ_{kj} \) represents the uniform density associated with the Lebesgue measure on \( \mathbb{R}^{rac{n(n+1)}{2}} \). What discussed in section 5.3 then assures us that the generating set described in Equation (7.26) is formed by vector fields with zero divergence with respect to \( \mu_g \).

Experiments:

The target density used in the experiments for \( \text{Sym}^+(n) \) is a \( k = 4 \) mixture of Wishart distributions with components given by:

\[ \tilde{\rho}(Q\mid \beta, W) \bigg| \bigwedge_{1 \leq k \leq j \leq n} dQ_{kj} \bigg| = \frac{\det(Q)^{\frac{\beta-n-1}{2}}}{\det(W)^{\frac{\beta}{2}}} \exp \left( -\frac{1}{2} \text{tr} \left( W^{-1} Q \right) \right) \bigg| \bigwedge_{1 \leq k \leq j \leq n} dQ_{kj} \bigg| \quad \forall Q \in \text{Sym}^+(n) \]

Where \( n + 1 \leq \beta \in \mathbb{N} \) corresponds to the degrees of freedom, and \( W \in \text{Sym}^+(n) \). The above probability density function is given with respect to the density associated with the Lebesgue measure in \( \mathbb{R}^{rac{n(n+1)}{2}} \). If we use as a

\footnote{See Moakher and Zerai (2011).}
Table 7.6: Evaluation of Manifold Continuous Normalizing Flow (MCNF) on $\text{Sym}^+(n)$. The experiments on the Highlighted rows were repeated 5 times and reported in the summary Table 7.7.

| Manifold  | $\beta$ | KL[nats] | ESS [%] |
|-----------|---------|----------|---------|
| $\text{Sym}^+(3)$ | 5       | 0.006    | 98.9    |
| $\text{Sym}^+(3)$ | 10      | 0.007    | 98.4    |
| $\text{Sym}^+(3)$ | 20      | 0.02     | 94.9    |
| $\text{Sym}^+(2)$ | 5       | 0.001    | 99.7    |
| $\text{Sym}^+(2)$ | 10      | 0.004    | 99.2    |
| $\text{Sym}^+(2)$ | 20      | 0.009$\pm$0.002 | 98.01$\pm$0.1 |

base density the congruence invariant density $\mu_g$ defined by Equation (7.29) we can rewrite:

$$
\rho(Q|\beta,W)\mu_g(Q) = \left(\frac{\det(Q)}{\det(W)}\right)^{\frac{\beta}{2}} \exp\left(-\frac{1}{2}\text{tr}\left(W^{-1}Q\right)\right) \mu_g(Q) \quad (7.30)
$$

$$
\forall Q \in \text{Sym}^+(n) \quad (7.31)
$$

Algorithm 1 in Section 4.4.2 of Camaño-Garcia (2006) was employed for sampling. As initial probability density, we used $\rho(\cdot|\beta, I_\beta^2)\mu_g$. In the experiments, we fixed $\beta \in \{5, 10, 20\}$, $n \in \{2, 3\}$ with centers:

$$
W_{1,2,3,4} := \frac{1}{\beta} \widehat{W}_{1,2,3,4} \quad \text{where} \quad (7.32)
$$

$$
\widehat{W}_{1,2,3,4} := \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}. \quad (7.33)
$$

$$
W_{1,2,3,4} := \frac{1}{\beta} \widehat{W}_{1,2,3,4} \quad \text{where} \quad (7.34)
$$

$$
\widehat{W}_{1,2,3,4} := \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}. \quad (7.35)
$$

Results are reported in Table 7.6.

### 7.3 Summary of results and comment

Table 7.7 shows that the proposed MCNF is able to closely match the target densities on all the spaces, with low KL and high ESS (> 90%). The results
Table 7.7: Summary of manifold continuous normalizing flow (MCNF) on mixture density matching task. Error is computed over 5 replicas of each experiment.

| Manifold    | dimension | KL[nats] | ESS[%]  |
|-------------|-----------|----------|---------|
| $V_1(\mathbb{R}^4) \cong S^2$ | 2         | $0.003 \pm 0.001$ | $99.3 \pm 0.8$ |
| $V_1(\mathbb{R}^3) \cong S^3$ | 3         | $0.006 \pm 0.002$ | $98.6 \pm 0.5$ |
| $V_2(\mathbb{R}^4)$          | 5         | $0.02 \pm 0.006$  | $96.5 \pm 1.1$ |
| $V_2(\mathbb{R}^3)$          | 7         | $0.02 \pm 0.003$  | $97.1 \pm 0.6$ |
| $SO(3)$                     | 3         | $0.03 \pm 0.007$  | $95.3 \pm 0.8$ |
| $S^{10}$                     | 10        | $0.01 \pm 0.003$  | $97.2 \pm 0.6$ |
| $S^{20}$                     | 20        | $0.02 \pm 0.003$  | $95.4 \pm 0.4$ |
| $SU(2)$                     | 3         | $0.005 \pm 0.003$ | $98.6 \pm 0.8$ |
| $SU(3)$                     | 8         | $0.01 \pm 0.003$  | $97.1 \pm 0.6$ |
| $U(2)$                      | 4         | $0.02 \pm 0.008$  | $95.4 \pm 1.4$ |
| $U(3)$                      | 9         | $0.05 \pm 0.008$  | $91.5 \pm 1.3$ |
| $\text{Sym}^+(2)$           | 3         | $0.009 \pm 0.002$ | $98.01 \pm 0.1$ |

on $S^{10}$ and $S^{20}$ show that the proposed method can effectively scale to higher dimensional manifolds. Interestingly, the experiments seem to indicate that the model has more difficulty in matching the target distribution when the space is not simply connected ($SO(3)$, $U(2)$, $U(3)$).
Chapter 8

Conclusion

In this thesis we presented a universal methodology for building free form normalizing flows on manifolds. To achieve this objective we generalized neural ordinary differential equations based architectures and continuous normalizing flows to arbitrary smooth manifolds. As manifolds are nontrivial mathematical objects to work with, we strove to build our theory on strong mathematical foundations, recognizing in differential geometry the correct language to define our framework.

Working with intrinsically defined objects on manifolds has several advantages over employing quantities that depend on a particular parameterization or defined on local coordinates. First of all, it decouples the mathematical derivation from a specific algorithmic choice, making it easier to use the most convenient parameterization for each specific space. For example, our definition of cotangent lift allows to freely choose the numerical integration method to use in practice. Secondly, our formulation paves the way for future research, such as the design of theoretically sound regularization methods, or the investigation of stability properties and approximation capabilities of neural ODEs on manifolds.

Notwithstanding the above, we strove to present a framework that can be easily implemented and involves efficient computations. We achieve this by employing a generating set of vector fields in combination with free form neural architectures to parameterize vector fields. We then showed that this formulation leads to a simplified divergence calculation for homogeneous spaces and Lie groups, spaces of outstanding importance in many fields of science and engineering. These claims are supported by our experiments, which prove how the defined framework can be used to successfully train normalizing flows on a wide array of spaces. An aspect that we didn’t consider
in this thesis, and that we leave for future research, is using specific numerical integrators adapted to the manifold structure.

We hope that the methods and algorithms developed in this thesis will allow probabilistic deep learning models to tackle problems in many scientific disciplines outside machine learning.
Appendix A

Densities and induced measure

A.1 volume forms

In the rest of the chapter $M$ will be a smooth manifold of dimension $n$

**Definition A.0.1.** A volume form $\omega$ is a section of the line bundle $\Lambda^n T^* M = \Lambda^{top} T^* M$, that is $\omega \in \Gamma (M, \Lambda^n T^* M)$.

In any smooth chart $(U; x_i)$ a volume form $\omega$ can be written as:

$$\omega \big|_U = f dx_1 \wedge \cdots \wedge dx_n \quad f \in C(U)$$  \hspace{1cm} (A.1)

If the manifold is orientable, then there exists a smooth non vanishing volume form that is positively oriented at each point $^1$. Since $\Lambda^n T^* M$ is a vector bundle of rank 1, orientability implies that the line bundle is isomorphic to the trivial bundle $M \times \mathbb{R}$, such that any non vanishing (smooth) volume form is a (smooth) global frame.

This means that fixed a smooth nonvanishing volume form $\omega \in \Gamma^\infty (M, \Lambda^n T^* M)$ we have a 1 to 1 correspondence between continuous (resp. smooth) functions on $M$ and (resp. smooth) volume forms on $M$:

$$C(M) \rightarrow \Gamma (M, \Lambda^n T^* M) \quad \text{and} \quad C^\infty (M) \rightarrow \Gamma^\infty (M, \Lambda^n T^* M)$$  \hspace{1cm} (A.2)

$$f \mapsto f \omega \quad \quad f \mapsto f \omega$$  \hspace{1cm} (A.3)

The map depends explicitly on the initial choice for $\omega$. On semi-Riemannian manifolds there is a standard choice for this form:

$^1$Proposition 15.5 Lee (2013)
Prop A.0.1.  

Suppose \((M,g)\) is an oriented semi Riemannian \(n\)-manifold, and \(n \geq 1\). There is a unique smooth non vanishing form \(\omega_g \in \Gamma^\infty(N,\Lambda^nT^*M)\), called the semi Riemannian volume form, that satisfies:

\[
\omega_g(E_1, \cdots, E_n) = 1 \quad (A.4)
\]

for every local oriented orthonormal frame\(^3\) \(\{E_i\}\) on \(M\).

In any local coordinates \((U; x_i)\) the semi Riemannian volume form \(\omega_g\) can be written as:

\[
\omega_g|_U = \sqrt{|\det g_{ij}|} \, dx_1 \wedge \cdots \wedge dx_n \quad (A.5)
\]

Where for each point \(q \in U\), \(g_{ij}(q)\) is the symmetric matrix that gives the local coordinate expression for the metric \(g\).

Given a volume form on a smooth manifold we can define its integral. For the rest of the section, we will outline the principal steps of the construction. The interested reader can consult Chapter 16 of Lee (2013) for a complete exposition.

We first define the integral of a volume form \(\omega\) compactly supported on the domain of a smooth chart \((U, \varphi)\) with coordinates \((x_i)\):

\[
\int_M \omega = \int_U f \, dx_1 \wedge \cdots \wedge dx_n := \pm \int_{\varphi(U)} f \circ \varphi^{-1} \, dx_1 \wedge \cdots \wedge dx_n = 4 \quad (A.6)
\]

\[
\int_{\varphi(U)} f(\varphi^{-1}(x)) \, dx \quad (A.7)
\]

Where \(\omega|_U = f \, dx_1 \wedge \cdots \wedge dx_n\) and the sign depends on the orientation. Notice that \(f \circ \varphi^{-1} \, dx_1 \wedge \cdots \wedge dx_n = (\varphi^{-1})^* \omega \in \Lambda^n T^*\varphi(U)\) is a volume form on a compactly supported open set of \(\mathbb{R}^n\), and we define its integral as simply "erasing" the wedges and computing the corresponding Riemann integral. \(^5\) Proposition 16.4 of Lee (2013) ensures that the integral is well defined, i.e.

\(^2\)See for example Proposition 15.29 Lee (2013). Many of the theorems here expressed for semi-Riemannian manifolds, are in the original source expressed and proved only Riemannian manifolds, however, they can effortlessly generalized to semi-Riemannian manifolds. A treatment that includes also semi Riemannian metrics can be found in O’Neill (1983).

\(^3\)The existence of a smooth orthonormal frame around every point in a semi Riemannian manifold is ensured by Corollary 13.8 in Lee (2013).

\(^5\)With a slight abuse of notation we use the \(dx_i\)s both as a basis for \(T^*U \subset T^*M\) and \(T^*\varphi(U) \subset T^*\mathbb{R}^n\). In the first case \(dx_i\) can be interpreted as the differential of the coordinate function \(x_i : U \rightarrow \mathbb{R}\)
that it does not depend on the choice of the smooth chart whose domain contains the support of $\omega$.

Once we have defined the integral for a volume form compactly supported in one chart domain, we can define the general integral of a volume form $\omega$. This is done using the partition of unity to express the initial volume form as a sum of volume forms which are compactly supported in one chart domain. To achieve this, we first cover the support of $\omega$ using a finite collection $\{U_i\}$ of open smooth chart domains. We then take a partition of unity $\alpha_i$ subordinate to the open collection. We then define the integral of $\omega$ over $M$ as:

$$\int_M \omega := \sum_i \int_M \alpha_i \omega$$ (A.8)

It can be shown that this definition does not depend on the particular choice of the open cover and partition of unity.

### A.2 Densities

**Definition A.0.2.** Let $V$ be a $n$-dimensional vector space. A **density on a vector space** $V$ is a function:

$$\mu : V \times \cdots \times V \rightarrow \mathbb{R}$$  \hspace{1cm} (A.9)

Such that for every linear map $T : V \rightarrow V$:

$$\mu(Tv_1, \cdots, Tv_n) = |\det T| \mu(v_1, \cdots, v_n) \quad \forall v_1, \cdots, v_n \in V$$  \hspace{1cm} (A.10)

A density is said positive if $\mu(v_1, \cdots, v_n) > 0$, for every linear independent tuple $(v_i)_{i=[n]}$. A negative density is similarly defined.

The set of all densities on $V$, $\mathcal{D}(V)$ is a 1 dimensional vector space. Given a $n$-form $\omega \in \Lambda^n(V^*)$, we can define a density $|\omega| : V \times \cdots \times V \rightarrow \mathbb{R}$ as

$$|\omega|(v_1, \cdots, v_n) = |\omega(v_1, \cdots, v_n)|$$ (A.11)

On a smooth manifold $M$ we can take the collection of all densities defined on each tangent space $T_q M$, $q \in M$, and give it a vector bundle structure.

---

6This is possible because $\omega$ has compact support

7Prop. 16.5 Lee (2013).

8Prop. 16.35 Lee (2013).
More formally, we define the **density bundle** of $M$ as the set:

$$\mathcal{D}(M) := \bigsqcup_{q \in M} \mathcal{D}(T_q M)$$  \hspace{1cm} (A.12)

Moreover if we define $\pi : \mathcal{D}(M) \to M$ to be the natural projection that sends every element of $\mathcal{D}(T_q M)$ to $q$ we have:

**Prop A.0.2.** \(^9\) $(\mathcal{D}(M), \pi, M)$ with $\mathcal{D}(M)$ and $\pi$ defined as above, is a smooth line bundle over $M$.

A section of $\mathcal{D}M$ is called a density on $M$. Densities are closely connected with volume forms. In fact, a non-vanishing volume form $\omega \in \Lambda^n T^*M$ determines a positive density$^{10}$ $|\omega|$, defined as $|\omega|_q := |\omega_q|$, $\forall q \in M$. Moreover in local coordinates $(U; x_i)$ a density $\mu$ can be expressed as:

$$\mu|_U = f|dx_1 \wedge \cdots \wedge dx_n| \quad f \in C(U)$$  \hspace{1cm} (A.13)

Like differential forms, densities can be pulled back:

**Definition A.0.3.** Let $F : M \to N$ be a smooth map between manifolds of dimension $n$, and $\mu$ a density over $N$. We define a density $F^*\mu$ on $M$, called the pullback density, as:

$$(F^*\mu)_p(v_1, \cdots, v_n) = \mu_{F(p)}(dF_pv_1, \cdots, dF_pv_n)$$  \hspace{1cm} (A.14)

**Prop A.0.3.** \(^11\) Let $G : P \to M$ and $F : M \to N$ be smooth maps between $n$ dimensional manifolds, and let $\mu$ be a density on $N$

- $a \ \forall f \in C^\infty(N), F^*(f\mu) = (f \circ F)F^*\mu$
- $b \ \text{If } \omega \text{ is an } n\text{-form on } N, \text{ then } F^*|\omega| = |F^*\omega|$
- $c \ \text{If } \mu \text{ is smooth, then } F^*\mu \text{ is a smooth density on } M$
- $d \ (F \circ G)^*\mu = G^*(F^*\mu)$

Differently from volume forms, a smooth positive density exists on every smooth manifold:

**Prop A.0.4.** \(^12\) If $M$ is a smooth manifold, there exists a smooth positive density on $M$

---

\(^9\)Prop. 16.36 Lee (2013)

\(^10\)A positive density is a section that maps every point $p$ in the manifold to a positive density on $T_pM$. A positive density is in particular nonvanishing

\(^11\)Prop. 16.38 Lee (2013)

\(^12\)Prop 16.37 Lee (2013)
Since the density bundle is a line bundle, the previous proposition implies that the density bundle is always parallelizable and that any positive density forms a global frame. This means that fixed a smooth positive density $\mu \in \Gamma^\infty (M, D^M)$ we have a 1 to 1 correspondence between continuous (resp. smooth) functions on $M$ and (resp. smooth) densities on $M$:

$$C(M) \to \Gamma (M, D^M) \quad \text{and} \quad C^\infty (M) \to \Gamma^\infty (M, D^M)$$  \hspace{1cm} (A.15)

$$f \mapsto f\mu \quad \text{and} \quad f \mapsto f\mu$$  \hspace{1cm} (A.16)

As with volume forms, these maps explicitly depends on the initial choice for $\mu$. Furthermore, in semi-Riemannian manifolds there is a standard choice for this density that generalizes the semi Riemannian volume form:

**Prop A.0.5.** \(^{13}\) Let $(M,g)$ be a semi Riemannian manifold There is a unique smooth positive density $\mu_g$ on $M$, called the semi Riemannian density, with the property that

$$\mu_g(E_1, \cdots, E_n) = 1$$

for any local orthonormal frame $(E_i)$.

On an oriented semi Riemannian manifold we have $\mu_g = |\omega_g|$.

We can define the integral of compactly supported densities on a manifold. The construction closely mirrors the definition of the integral of volume forms, first defining the integral for densities compactly supported on a chart domain and then generalizing to arbitrary compact support using a partition of unity. For a detailed description of the process, we refer to Lee (2013).

The following proposition illustrates the properties of the integral:

**Prop A.0.6.** \(^{14}\) Suppose $M$ is a $n$-manifolds and $\mu, \eta$ are compactly supported densities on $M$:

- **LINEARITY** If $a, b \in \mathbb{R}$, then:

  $$\int_M a\mu + b\eta = a\int_M \mu + b\int_M \eta.$$

  \hspace{1cm} (A.17)

- **POSITIVITY** If $\mu$ is a positive density, then:

  $$\int_M \mu > 0$$

  \hspace{1cm} (A.18)

This properties fundamentally tell us that this integral is a positive linear functional on the vector space of densities with compact support on a smooth manifold.

\(^{13}\)Proposition 16.45 Lee (2013)

\(^{14}\)Prop. 16.42 Lee (2013)
### A.2.1 Induced measure

Using the correspondence given by Equation (A.15) every nonnegative density defines a positive linear functional on $C_c(M)$. In fact, fixed $\mu \in \Gamma (M, DM)$ nonnegative density we define:

$$L_\mu : C_c(M) \to \mathbb{R}$$

$$f \mapsto \int_M f \mu$$

We can then use the Riesz representation theorem to (uniquely) define a Radon measure on $M$:

**Theorem A.1.**\(^{15}\) Let $\mathcal{X}$ be a locally compact Hausdorff space, and let $L$ be a positive linear functional on $C_c(\mathcal{X})$. Then there exists a unique Radon measure $\mu$ such that:

$$Lf = \int_\mathcal{X} f d\mu, \quad \forall f \in C_c(\mathcal{M})$$

The Radon measure is said to represent the functional $L$.

One direct consequence of the Riesz representation theorem is that if we have two Radon measures $\mu_1$ and $\mu_2$ on a locally compact Hausdorff space such that their Lebesgue integral coincides on continuous and compactly supported functions:

$$\int_{\mathcal{X}} g d\mu_1 = \int_{\mathcal{X}} g d\mu_2 \quad \forall g \in C_c(\mathcal{X})$$

then the two measures coincide ($\mu_1 = \mu_2$).

We can use a positive smooth density and the Riesz representation theorem to define a Radon measure on a smooth manifold.

**Theorem A.2.** Let $M$ be a smooth manifold and $\mu$ a non negative density on $M$, then there exists a unique Radon regular measure $\tilde{\mu}$ such that:

$$\int_M f \mu = \int_M f d\tilde{\mu} \quad \forall f \in C_c(M)$$

We will say that the measure $\tilde{\mu}$ is derived from $\mu$.

---

\(^{15}\)Theorem 7.2 Folland (1999))

\(^{16}\)A radon measure is a measure defined on Borel sets, that is finite on all compact sets, outer regular on all Borel sets, and inner regular on open sets.
**APPENDIX A. DENSITIES AND INDUCED MEASURE**

*Proof.* We first show that we can apply the Riesz representation theorem for the functional $L_\mu$. Since any manifold is a locally compact Hausdorff space, this reduces to proving that $L_\mu$ is a positive linear functional. Using Proposition A.0.6 we see that:

$$L_\mu(\alpha f + \beta g) = \int_M \alpha f \mu + \beta g \mu = \alpha \int_M f \mu + \beta \int_M g \mu = \alpha L_\mu f + \beta L_\mu g$$

for all $f, g \in C_c(M)$, $\forall \alpha, \beta \in \mathbb{R}$

and

$$L_\mu f = \int_M f \mu \geq 0 \quad \forall f \in C_c(M), \ f \geq 0$$

This shows that $L_\mu$ is a linear positive functional on $C_c(M)$, therefore a Radon measure $\tilde{\mu}$ that represents $L_\mu$ exists. For Corollary 7.6 in Folland (1999) any Radon measure in a $\sigma$-compact space is regular. Therefore, since any Manifold is $\sigma$-compact, $\tilde{\mu}$ is also regular.

This choice of measure on $M$ depends on the initial choice for $\mu$. The next proposition shows how two measures derived from two different densities relate to each other:

**Prop A.2.1.** Let $\tilde{\eta}$ and $\tilde{\mu}$ two regular Radon measures derived from two densities $\eta$, $\mu$, with $\mu$ positive. Then $\tilde{\nu} \ll \tilde{\mu}$, that is there exists a continuous function $f$ such that $f \tilde{\nu} = \tilde{\eta}$. If both $\eta$ and $\mu$ are smooth then $f$ is a smooth function.

*Proof.* Since every positive density forms a frame for the line bundle $DM$, we can use Equation (A.15) to state that there exists $f \in C(M)$ such that $\eta = f \mu$. If $\mu$ and $\eta$ are both smooth then also $f$ is smooth.

Consider now the measure $f \tilde{\mu}$, we see that $\tilde{\eta}$ and $f \tilde{\mu}$ represent the same positive linear functional on $C_c(M)$:

$$\int_M g d\tilde{\eta} = \int_M g \eta = \int_M g f \mu = \int_M g d(f \tilde{\mu}) \quad (A.24)$$

If $f \tilde{\mu}$ is a Radon measure then by Riesz representation theorem we can conclude that the two measures coincide.

To show that $f \tilde{\mu}$ is Radon we can use Theorem 7.8 from Folland (1999), that tells us that a in a locally compact space in which every open set is $\sigma$ compact, every Borel measure on $X$ which is finite on compact sets is Radon.

Since by definition every manifold satisfies the topological requirements that
the space needs to satisfy, we are left to prove that $f\tilde{\mu}$ is finite on compact sets. Then, fixed $K \subseteq M$ compact, we have:

$$[f\tilde{\mu}](K) = \int_K f d\tilde{\mu} \leq \sup_K f \int_K d\tilde{\mu} = \sup_K f \tilde{\mu}(K) < +\infty$$ (A.25)

Where in the last equality we have used that, since $\tilde{\mu}$ is Radon, $\tilde{\mu}(K) < +\infty$, and that $\sup_K f < +\infty$ since $f$ is continuous and $K$ compact. Therefore $f\tilde{\mu}$ is finite on compact sets and this concludes our proof. □

In practice, fixed a smooth positive density $\mu$, we will work with absolutely continuous probability measures $g\tilde{\mu}$, where $g \in L^1_\tilde{\mu}(M), \ g \geq 0, \ \int_M g d\tilde{\mu} = 1$. Then Proposition (A.2.1) tells us that the set of measures that we can express in this way does not depend on the initial choice for $\mu$. 
Bibliography

Agrachev, A., Barilari, D., and Boscain, U. (2019). A comprehensive introduction to sub-riemannian geometry.

Agrachev, A. A. and Sachkov, Y. (2013). Control theory from the geometric viewpoint, volume 87. Springer Science & Business Media.

Alekseevsky, D., Grabowski, J., Marmo, G., and Michor, P. W. (1994). Poisson structures on the cotangent bundle of a lie group or a principle bundle and their reductions. Journal of Mathematical Physics, 35(9):4909–4927.

Ayala, V., Rodríguez, J., and Sammartin, L. (2009). Optimality on homogeneous spaces, and the angle system associated with a bilinear control system. SIAM J. Control and Optimization, 48:2636–2650.

Bates, J. (2014). The embedding dimension of laplacian eigenfunction maps. Applied and Computational Harmonic Analysis, 37(3):516–530.

Bose, A. J., Smofsky, A., Liao, R., Panangaden, P., and Hamilton, W. L. (2020). Latent variable modelling with hyperbolic normalizing flows. arXiv preprint arXiv:2002.06336.

Boyda, D., Kanwar, G., Racanière, S., Rezende, D. J., Albergo, M. S., Cranmer, K., Hackett, D. C., and Shanahan, P. E. (2020). Sampling using su(n) gauge equivariant flows.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. (2018). JAX: composable transformations of Python+NumPy programs.

Camaño-García, G. (2006). Statistics on Stiefel manifolds. PhD thesis, Iowa State University.

Cao, N. D. and Aziz, W. (2020). The power spherical distribution. ICML workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models.
Chen, B.-Y. and Verstraelen, L. (2013). Laplace transformations of submanifolds.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, Advances in Neural Information Processing Systems 31, pages 6571–6583. Curran Associates, Inc.

Da Silva, A. C. (2001). Lectures on symplectic geometry, volume 3575. Springer.

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. (2018). Hyperspherical variational auto-encoders. 34th Conference on Uncertainty in Artificial Intelligence (UAI-18).

Davidson, T. R., Tomczak, J. M., and Gavves, E. (2019). Increasing expressivity of a hyperspherical vae. NeurIPS, in Workshop on Bayesian Deep Learning.

Dinh, L., Krueger, D., and Bengio, Y. (2015). Nice: Non-linear independent components estimation. ICLR Workshop.

Falorsi, L., de Haan, P., Davidson, T. R., De Cao, N., Weiler, M., Forré, P., and Cohen, T. S. (2018). Explorations in homeomorphic variational auto-encoding. ICML workshop on Theoretical Foundations and Applications of Deep Generative Models.

Falorsi, L., de Haan, P., Davidson, T. R., and Forré, P. (2019). Reparameterizing distributions on lie groups. AISTATS.

Figurnov, M., Mohamed, S., and Mnih, A. (2018). Implicit reparameterization gradients. In Advances in Neural Information Processing Systems, pages 441–452.

Folland, G. (1999). Real analysis: modern techniques and their applications. Pure and applied mathematics. Wiley.

Gemici, M. C., Rezende, D., and Mohamed, S. (2016). Normalizing flows on riemannian manifolds. arXiv preprint arXiv:1611.02304.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., and Duvenaud, D. (2019). Scalable reversible generative models with free-form continuous dynamics. In International Conference on Learning Representations.
Haber, E. and Ruthotto, L. (2017). Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations. 2nd ed*, volume 31. Springer Science & Business Media.

Howard, R. (1994). Analysis on homogeneous spaces.

Hutchinson, M. (1989). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communication in Statistics- Simulation and Computation*, 18:1059–1076.

Kanwar, G., Albergo, M. S., Boyda, D., Cranmer, K., Hackett, D. C., Racanière, S., Rezende, D. J., and Shanahan, P. E. (2020). Equivariant flow-based sampling for lattice gauge theory. *arXiv preprint arXiv:2003.06413*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kish, L. (1965). Survey sampling. *American Political Science Review*, 59(4):1025–1025.

Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kunzinger, M., Schichl, H., Steinbauer, R., and Vickers, J. A. (2006). Global gronwall estimates for integral curves on riemannian manifolds. *Revista Matemática Complutense*, 19(1):133–137.

Lee, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media.

Lee, J. M. (2013). Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer.

Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S.-N., and Sa, C. D. (2020). Neural manifold ordinary differential equations.
Mathieu, E., Le Lan, C., Maddison, C. J., Tomioka, R., and Teh, Y. W. (2019). Continuous hierarchical representations with poincaré variational auto-encoders. In *Advances in neural information processing systems*, pages 12544–12555.

Mathieu, E. and Nickel, M. (2020). Riemannian continuous normalizing flows.

Mezzadri, F. (2006). How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54.

Moakher, M. and Zerai, M. (2011). The riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision*, 40:171–187.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2020). Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62.

Naesseth, C., Ruiz, F., Linderman, S., and Blei, D. (2017). Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. (2019). A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pages 4693–4702.

Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147.

O’Neill, B. (1983). *Semi-Riemannian Geometry With Applications to Relativity*. Pure and Applied Mathematics. Elsevier Science.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *stat*, 1050:5.

Pontryagin, L. S., Mishchenko, E., Boltyanskii, V., and Gamkrelidze, R. (1962). *The mathematical theory of optimal processes*. Wiley.

Pérez Rey, L. A., Menkovski, V., and Portegies, J. W. (2019). Diffusion variational autoencoders. *arXiv*. 
Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538.

Rezende, D. J., Papamakarios, G., Racanière, S., Albergo, M. S., Kanwar, G., Shanahan, P. E., and Cranmer, K. (2020). Normalizing flows on tori and spheres. *arXiv preprint arXiv:2002.02428*.

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., New York, NY, USA.

Tabak, E. G. and Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164.

Tabak, E. G., Vanden-Eijnden, E., et al. (2010). Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233.

Teschl, G. (1998). Topics in real and functional analysis.

Walschap, G. (2004). *Metric Structures in Differential Geometry*. Graduate Texts in Mathematics. Springer, softcover reprint of hardcover 1st ed. 2004 edition.

Wirnsberger, P., Ballard, A. J., Papamakarios, G., Abercrombie, S., Racanière, S., Pritzel, A., Rezende, D. J., and Blundell, C. (2020). Targeted free energy estimation via learned mappings. *arXiv preprint arXiv:2002.04913*.