

# Projection Geometric Methods for Linear and Non-linear Filtering Problems

by

Ashraf Ahmed

A thesis  
presented to the University of Waterloo  
in fulfillment of the  
thesis requirement for the degree of  
Master of Mathematics  
in  
Statistics

Waterloo, Ontario, Canada, 2024

© Ashraf Ahmed 2024

## **Author's Declaration**

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## **Statement of Contributions**

Ashraf Ahmed was the sole author for Chapters 1, 2, 3 and 5 which were written under the supervision of Dr. Paul Marriott and were not written for publication.

Exceptions to sole authorship of material are as follows:

Research presented in Chapters 4: Dr. Paul Marriott was the primary co-investigator and are co-authors on any publications relating to this work.

## Abstract

In this thesis, we review the infinite-dimensional space containing the solution of a broad range of stochastic filtering problems, and outline the substantive differences between the foundations of finite dimensional information geometry and Pistone's extension to infinite dimensions characterizing the substantive differences between the two geometries with respect to the geometric structures needed for projection theorems such as a dually flat affine manifold preserving the affine and convex geometry of the set of all probability measures with the same support, the notion of orthogonal complement between the different tangent representation which are key for the generalized Pythagorean theorem, and the key notion of exponential and mixture parallel transport needed for projecting a point on a submanifold. We also explore the projection method proposed by Brigo and Pistone for reducing the dimensionality of infinite-dimensional measure-valued evolution equations from the infinite-dimensional space in which they are written, that is, the infinite-dimensional statistical manifold of Pistone, onto a finite-dimensional exponential subfamily using a local generalized projection theorem that is a nonparameteric analog of the generalized projection theorem proposed by Amari. Also, we explore using standard arguments the projection idea in the discrete state space with a focus on building intuition and using computational examples to understand properties of the projection method. We establish two novel results regarding the impact of the boundary and choosing a subfamily that does not contain the initial condition of the problem. We demonstrate, when the evolution process approaches the boundary of the space, the projection method fails completely due to the classical boundary relating to the vanishing of the tangent spaces at the boundary. We also show the impact of choosing a subfamily to project onto that does not contain the initial condition of the problem, showing that, in certain directions, the approximation by projection changes from the true value due to solving a different differential equation than if we are to start from within the low-dimensional manifold. We also study the importance of having sufficient statistics of the exponential subfamily to lie in the span of the left eigenfunctions of the infinitesimal generator of the process we wish to project using computational experiments.

## Acknowledgments

This thesis would not have been possible without the constant support and genuine care that I received from my supervisor, Dr. Paul Marriott. I can sincerely say that working with Dr. Paul has been the highlight of my master's studies.

I express my deep gratitude to the committee members, Dr. Martin Lysy and Dr. Tony Wirjanto, for their precious time and effort to revise this thesis.

My master's years at the University of Waterloo have been full of personal and academic growth. I would like to express my appreciation to the university staff and my peers for nurturing me and providing me with a wonderful environment to learn and grow. In particular, I would like to thank Dr. Ruxandra Moraru, Dr. Shoja'eddin Chenouri, and Dr. Chris Eliasmith, and Dr. Christopher Nielsen for playing an important role in my academic years at Waterloo and beyond.

Most importantly, I am forever grateful to my wife, Alexandra Muton, my friends, and my dog Luna. No words can express the magnitude of the love, support, and patience they provided me throughout my life.

## **Dedication**

Meiner Frau Alex, ohne die ich nicht leben könnte, und meiner besten Freundin Luna, deren tägliches Lächeln mir alles bedeutet.

To my wife Alex, without whom I could not live, and to my best friend Luna, whose smile means everything to me.

# Table of Contents

<b>Author's Declaration</b>	<b>ii</b>
<b>Statement of Contributions</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Acknowledgments</b>	<b>v</b>
<b>Dedication</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 The Filtering Problem . . . . .	1
1.2 General Solution . . . . .	4
1.3 Linear Filters . . . . .	8
1.4 Projection Filter . . . . .	10
<b>2 The foundations of finite dimensional Information Geometry</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 Definitions . . . . .	16
2.3 Affine geometry . . . . .	19

2.4	Convex structures . . . . .	24
2.5	Manifold structure . . . . .	31
2.6	Tangent spaces . . . . .	33
2.7	Riemannian manifold structure . . . . .	41
2.8	Projections and the Pythagorean theorem . . . . .	44
<b>3</b>	<b>Infinite Dimensional Information Geometry</b>	<b>46</b>
3.1	Introduction . . . . .	46
3.2	Overview . . . . .	47
3.3	Definitions . . . . .	51
3.4	Generalized parameter space . . . . .	53
3.5	Manifold structure . . . . .	60
3.6	Tangent spaces . . . . .	70
3.7	Duality and Projections . . . . .	73
<b>4</b>	<b>Examples</b>	<b>82</b>
4.1	Projection in the discrete state space . . . . .	83
4.2	Numerical experiments . . . . .	89
4.2.1	Boundary and initial condition effects . . . . .	89
4.2.2	Visualizing the flow . . . . .	93
4.2.3	Optimal projection of a stationary random walk . . . . .	96
4.2.4	Suboptimal projection of a stationary random walk . . . . .	99
<b>5</b>	<b>Conclusion</b>	<b>101</b>
5.1	Future Directions . . . . .	103
	<b>References</b>	<b>104</b>

# List of Figures

3.1	Infinite-dimensional manifold structure of a set of probability measures that agree on a set of measure zero . . . . .	70
4.1	Evolution of the true equation in red, and its projection in the exponential family in blue with the same initial condition. Right hand panel shows the calculated mean value $\mu(t)$ for both. . . . .	90
4.2	A visualization of the solution to a differential equation in red, its projection on a 2-dimensional exponential family in blue with the same initial conditions, and the calculated mean value for both. . . . .	91
4.3	A visualization of the solution to a differential equation in red, its projection on a 2-dimensional exponential family in blue not containing the initial condition, and the calculated mean value for both. . . . .	92
4.4	The flow of a non-stationary random walk with parameters $\theta = 0.5, \phi = 0.3$ on the closed 80-simplex at different time steps. The vertical gray dashed line is marker for the singular initial distribution . . . . .	94
4.5	The flow of a non-stationary random walk with parameters $\theta = 0.5, \phi = 0.3$ on the closed 240-simplex at different time steps. The vertical gray dashed line is marker for the singular initial distribution . . . . .	95
4.6	Two computed eigenfunctions of the operator associated with a stationary random walk model 4.1.1 with parameters $\theta = 0.35$ and $\phi = 0.35$ . . . . .	96

4.7	Flow of the mean of a stationary random walk on a 50-simplex with parameters $\theta = \phi = 0.35$ and its projection on a 2-dimensional exponential family with a sufficient statistic matching the first two eigenfunctions of the infinitesimal generator of the random walk. The gray dashed vertical line highlights the position of the singular initial condition. Note the differences in the shape of the probability mass functions between the random walk, and its approximation. . . . .	97
4.8	Computed mean of the true solution of a stationary random walk on a 50-simplex with parameters $\theta = \phi = 0.35$ and its projection on a 2-dimensional exponential family with sufficient statistics matching the eigenfunctions of the infinitesimal generator of the random walk. . . . .	98
4.9	Two computed eigenfunctions of the operator associated with a stationary random walk model 4.1.1 and the two sufficient statistics of the low-dimensional exponential family . . . . .	99
4.10	Evolution of probability mass functions of a stationary random walk with parameters $\theta = \phi = 0.35$ on 50-simplex and its projection two different low-dimensional manifolds. (a) shows the evolution of the mass function when the sufficient statistics are exactly the eigenfunctions of the infinitesimal generator. (b) shows the evolution of the mass function when the second sufficient statistic is a nonlinear function of the second eigenfunction. (c) Computed mean of the true solution and its approximation using a projection on a non-optimal 2-dimensional exponential family. . . . .	100

# Chapter 1

## Introduction

### 1.1 The Filtering Problem

Stochastic filtering problems arise naturally in many areas, including engineering, physics, biology, chemistry, finance, and economics. The aim of stochastic filtering is to provide an (online) estimate of the state of a dynamic system that can only be partially observed through a noisy measurement system [32]. The classical and highly influential example of such a procedure is the Kalman filter [52]. Due to the regularity assumptions of the Kalman filter – linearity, independence, and Gaussianity – this filtering problem can be described in a geometric way as an orthogonal projection onto a finite-dimensional affine subspace. This thesis explores the work of Pistone et al. [71, 43, 70, 27, 22, 8, 7] which investigates whether a general filtering problem can be solved by projection methods in infinite-dimensional spaces.

This thesis looks at [23], which proposes a dimensional reduction method from infinite-dimensional spaces containing the solution of a broad range of filtering problems to a carefully chosen finite-dimensional manifold – in fact an exponential family. Dual affine projections between finite-dimensional exponential families are one of the key pillars of the theory of information geometry [5], which generalized the Pythagorean properties of a normal linear regression to a much wider class of finite-dimensional models. We carefully review the foundations of this theory in Chapter 2 to understand Pistone’s extension to infinite-dimensional spaces used in [23]. This extension is reviewed in Chapter 3, and the projection method is explored numerically in Chapter 4. The rest of this chapter carefully defines the geometry of a broad class of filtering problems and explains why it is necessary

to consider the extension of classical Information Geometry to the infinite-dimensional case.

The solution of stochastic filtering problems has attracted the attention of generations of mathematicians, statisticians, and engineers. The origin of the field can be traced back to the work of Kolmogorov and Krein for their study of stochastic filtering in discrete time [32]. Wiener was the first to study the estimation of a continuous-time stationary signal process under independent additive noise [83]. That work led to the development of an optimal filter, the Wiener filter, which had a significant impact in the field of defense and space exploration [32].

Mathematically, the system of interest is modeled using a stochastic process,  $X = \{X_t : t \in T\}$ , referred to as a signal, where  $T$  is an indexing set. The observation process, modeled using another stochastic process  $Y = \{Y_t : t \in T\}$ , is a function of both the signal and some measurement noise modeled by a Wiener process,  $V = \{V_t : t \in T\}$  so that  $Y_t = h(t, X_t, V_t)$  for  $t \in T$ . The objective is to use values of  $Y$  to estimate  $X$ , such that, the estimate  $\hat{X}_t$  have the following three important properties [32]

- *Causality*:  $X_t$  is to be estimated using only the history of  $Y_s$  for  $s \leq t$ .
- *Optimality*: The estimate  $\hat{X}_t$  should minimize the mean square error  $\mathbb{E}[(X - \hat{X}_t)^2]$ .
- *Online/concurrent estimation*: at any (arbitrary) time  $t$ , the estimate  $\hat{X}_t$  should be available.

Kalman [52] and Bucy [26] examined an important special case, known as a linear filter. In this setting, the signal is a stochastic process modeled by a differential equation driven by a Brownian motion process and has a Gaussian initial condition; while the observation process depends linearly on the signal. During the same period, and independently of Kalman and Bucy, Stratonovich [79] developed a modern theory of nonlinear stochastic filtering using a novel method in stochastic integration bearing his name. For a detailed historical account, refer to [32].

Following the theory of stochastic filtering, we shall work with the following framework and notation.

- (i)  $(\Omega, \mathcal{F}, \mathbb{P})$  is a probability space.
- (ii)  $(\mathcal{F}_t : 0 \leq t)$  is an increasing family of  $\sigma$ -algebras,  $\mathcal{F}_0 \subset \mathcal{F}_1 \cdots \subset \mathcal{F}$ , in  $\Omega$ , that is, a filtration of  $\mathcal{F}$ .

- (iii)  $\mathcal{F}$  is complete, that is,  $E_1 \subset E_2 \in \mathcal{F}$  and  $\mathbb{P}(E_2) = 0$  implies  $E_1 \in \mathcal{F}$ , and  $\mathbb{P}(E_1) = 0$ .
- (iv) The filtration  $(\mathcal{F}_t, 0 \leq t)$  is right-continuous.
- (v)  $\mathcal{F}$  (and consequently all elements of the filtration  $(\mathcal{F}_t : 0 \leq t)$ ) contain all the  $\mathbb{P}$ -null sets.
- (vi)  $X = \{X_t : 0 \leq t\}$  is a process adapted to  $(\mathcal{F}_t : 0 \leq t)$  and takes values in the state space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ , where  $\mathcal{B}(\mathcal{X})$  is the associated Borel  $\sigma$ -algebra of  $\mathcal{X}$ .
- (vii)  $\mathcal{P}(\mathcal{X})$  is the set of all probability measures on the state space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ .
- (viii)  $\mathcal{B}(\mathcal{X})$  is a Banach space of real-valued measurable functions on  $\mathcal{X}$  with  $\|\varphi\| = \sup_{x \in \mathcal{X}} |\varphi(x)|$  for all  $\varphi \in \mathcal{B}(\mathcal{X})$ .
- (ix)  $X_0$  is  $\mathcal{F}_0$  measurable.
- (x)  $V = \{V_t, 0 \leq t\}$  is an  $m$ -dimensional standard Wiener process adapted to  $(\mathcal{F}_t : 0 \leq t)$
- (xi)  $V$  is independent of  $X$ .
- (xii)  $h : \mathcal{X} \rightarrow \mathbb{R}^m$  is a  $\mathcal{B}(\mathcal{X})$  measurable function, known as the observation function, which satisfies
 
$$\mathbb{P}\left(\int_0^t \|h(X_s)\| ds < \infty\right) = 1 \quad \forall t \geq 0 \quad (1.1)$$
- (xiii)  $Y = \{Y_t, 0 \leq t\}$  is a process that satisfy
 
$$Y_t = Y_0 + \int_0^t h(X_s) ds + V_t, \quad \forall t \geq 0 \quad (1.2)$$
- (xiv)  $Y_0$  is identically zero (there is no information available from observations initially).
- (xv)  $\sigma(Y_s : 0 \leq s \leq t)$  is the  $\sigma$ -algebra generated by the process  $Y$ .
- (xvi)  $\mathcal{Y}_t = \sigma(Y_s, 0 \leq s \leq t) \cup \{E \in \mathcal{F} : \mathbb{P}(E) = 0\}$  is the augmentation of the filtration associated with the process  $Y$  with all the  $\mathbb{P}$ -null sets.

In its most general form, stochastic filtering is concerned with characterizing the conditional distribution  $\pi_t$  of the signal  $X_t$  given the information available from the process  $Y_t$  up to time  $t$ . More formally, the filtering problem is defined as follows.

**Definition 1.1.1.** (The Filtering Problem)[12, Definition 3.2, Page 48] The filtering problem consists of determining the conditional distribution  $\pi_t$  of the signal  $X$  at time  $t$  given the information accumulated from observing  $Y$  in the interval  $[0, t]$ ; that is, for any  $\mathcal{B}(\mathcal{X})$ -measurable (test) function  $\varphi : \mathcal{X} \rightarrow \mathcal{X}$ , computing

$$\pi_t \varphi = \mathbb{E}[\varphi(X_t) | \mathcal{Y}_t] \tag{1.3}$$

provided that  $\pi_t |\varphi| < \infty$ .

## 1.2 General Solution

An important class of problems is the case where the signal is a solution of the martingale problem for  $(A, \pi_0)$  where  $A$  is an operator on  $\mathcal{B}(\mathcal{X})$ . This particular class of problems is of practical and theoretical importance due to its connection to strong Markov processes [12].

**Definition 1.2.1.** [38] Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $(\mathcal{F}_t : 0 \leq t)$  be a filtration of  $\mathcal{F}$ . A process  $M = (M_t : 0 \leq t)$  on  $(\Omega, \mathcal{F}, \mathbb{P})$  with values in  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  is an  $\{\mathcal{F}_t\}$ -martingale if

1.  $M$  is adapted to the the filtration  $\{\mathcal{F}_t\}$ .
2.  $\mathbb{E}(|M_t|) < \infty, \quad \forall t \geq 0$
3.  $\mathbb{E}[M_{t+s} | \mathcal{F}_t] = M_t, \quad \text{for all } s, t \geq 0$

**Definition 1.2.2.** [38] Suppose that  $\mathcal{X}$  is a metric space. Let  $\mathcal{B}(\mathcal{X})$  be a Banach space of real-valued  $\mathcal{B}(\mathcal{X})$ -measurable functions on  $\mathcal{X}$  with  $\|\varphi\| = \sup_{x \in \mathcal{X}} |\varphi(x)|$  for all  $\varphi \in \mathcal{B}(\mathcal{X})$ . In addition, let  $A \subset \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{X})$ .

1. A solution of the martingale problem for  $A$  is a process  $X = (X_t : 0 \leq t)$  with values in  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  defined on some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  such that for each test function  $\varphi$  in the domain of the operator  $A$ , the process

$$M_t = \varphi(X_t) - \int_0^t A\varphi(X_s) ds$$

is a martingale with respect to the filtration

$$\mathcal{F}_t^* = \sigma(X_t : 0 \leq t) \cup \sigma\left(\int_0^s \varphi(X_u) du : s \leq t, \varphi \in \mathcal{B}(\mathcal{X})\right)$$

2. If  $(\mathcal{A}_t : 0 \leq t)$  is a filtration of  $\mathcal{F}$  such that  $\mathcal{F}_t^* \subset \mathcal{A}_t$  for all  $t \geq 0$ , and  $M_t$  is a  $\{\mathcal{A}_t\}$ -martingale for all test functions  $\varphi$  in the domain of  $A$ , we say that  $X$  is a solution of the martingale problem for  $A$  with respect to  $\{\mathcal{A}_t\}$ .
3. When an initial distribution  $\pi_0 \in \mathcal{P}(\mathcal{X})$  is specified, if  $X$  is a solution of the martingale problem for  $A$  and the distribution of  $X_0$  is  $\pi_0$ , we say  $X$  is a solution of the martingale problems for  $(A, \pi_0)$ .

In addition, for the signal process  $X_t$  being a solution of a martingale problem for  $(A, \pi_0)$ , we assume the following regularity conditions on  $X_t$ .

**Regularity conditions 1.2.1.** We assume throughout that

1. The state space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  of the signal process  $X_t$  is a complete separable metric space.
2. The signal process  $X_t$  has sample paths which are càdlàg, i.e. the sample paths are right continuous with left limits.
3. The signal process  $X_t$  is a solution of the martingale problem for  $(A, \pi_0)$ .

It is important to note that even though the regularity condition in (1) above is not the most general topological restriction, it covers many of the most important problems in practice such as diffusion and Markov jump processes. For extensions, see [42].

The following theorems establish that the conditional distribution process  $\pi$  is well defined [12] under the regularities (1.2.1) above.

**Theorem 1.2.1.** [12, theorem 2.1] Let  $X$  be a process on  $(\Omega, \mathcal{F}, \mathbb{P})$  such that  $X$  is adapted to the filtration  $(\mathcal{F}_t : 0 \leq t)$  with values in a complete separable metric space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ . If

- (a)  $Y$  is a process on  $(\Omega, \mathcal{F}, \mathbb{P})$  that satisfies

$$Y_t = Y_0 + \int_0^t h(X_s) ds + V_t, \quad \forall t \geq 0 \quad (1.4)$$

- (b) The observation function  $h$  satisfy

$$\mathbb{P}\left(\int_0^t \|\pi_s(h)\| ds < \infty\right) = 1 \quad \forall t \geq 0 \quad (1.5)$$

(c)  $V = (V_t : 0 \leq t)$  is an  $m$ -dimensional Wiener process on  $(\Omega, \mathcal{F}, \mathbb{P})$  that is independent of  $X$ .

then

- There exists a  $\mathcal{P}(\mathcal{X})$ -valued  $\mathcal{Y}_t$ -adapted progressively measurable process  $\pi = (\pi_t : 0 \leq t)$ .
- For any test function  $\varphi : \mathcal{X} \rightarrow \mathcal{X}$  that is  $\mathcal{B}(\mathcal{X})$ -measurable, the process  $\pi\varphi := (\pi_t\varphi : 0 \leq t)$  satisfies (1.3), i.e.  $\pi_t\varphi = \mathbb{E}[\varphi(X)|\mathcal{Y}_t]$   $\mathbb{P}$ -almost surely, for all  $t \geq 0$ , provided that  $\pi_t|\varphi| < \infty$
- If  $X$  has càdlàg sample paths, then  $\pi_t$  can be chosen to have càdlàg paths.

In 1958, Stratonovich [79] was the first to solve the filtering problem by studying the nonlinear transformation of conditional Markov processes. In 1960, Kushner [57, 58, 59] derived the evolution equation for the conditional distribution process  $\pi$  using Itô calculus. Independently of Kushner and Stratonovich, Shirayaev [77] rigorously derived the filtering equation that solves the filtering problem in the case of a general observation process where the signal and the observational noise are correlated.

The following theorem establishes the general solution of the filtering problem under assumptions (a), (b), and (c) above.

**Theorem 1.2.2.** [12, Theorem 3.30] Suppose that  $X, Y$  are two processes on  $(\Omega, \mathcal{F}, \mathbb{P})$  satisfying the conditions of Theorem (1.2.1) and the observation function  $h$  satisfies the following conditions:

1.  $X$  is a solution of the martingale problem for  $(A, \pi_0)$ .
2. The observation function  $h$  satisfies the following two regularity conditions

$$\mathbb{P}\left(\int_0^t \|\pi_s(h)\|^2 ds < \infty\right) = 1 \quad \forall t \geq 0 \quad (1.6)$$

$$\mathbb{E}\left[\int_0^t \|h(X_s)\| ds\right] < \infty \quad \forall t \geq 0 \quad (1.7)$$

then, for any  $\varphi$  in the domain of  $A$ , the process  $\pi$  satisfies the following Kushner-Stratonovich equation,

$$\begin{aligned} \pi_t(\varphi) = \pi_0(\varphi) &+ \int_0^t \pi_s(A\varphi)ds \\ &+ \int_0^t (\pi_s(\varphi h^T) - \pi_s(h^T)\pi_s(\varphi))(dY_s - \pi_s(h)ds) \end{aligned} \quad (1.8)$$

for all  $t \geq 0$ .

Using the Kushner-Stratonovich equation (1.8) requires checking conditions (1.6) and (1.7) which are difficult to do (particularly (1.6) since it requires a way to compute  $\pi_s$  for each  $s \in [0, t]$ ). As a result, it is typical to require the following stronger condition (which implies both (1.6) and (1.7))[12].

$$\mathbb{E} \left[ \int_0^t \|h(X_s)\|^2 ds \right] < \infty \quad \forall t \geq 0 \quad (1.9)$$

A rather very important class of signals is a case where  $X$  is diffusion processes that satisfy the Fokker-Planck-Kolmogorov equation.

**Definition 1.2.3.** (Fokker-Planck-Kolmogorov equation)[23] A process  $X = \{X_t : 0 \leq t\}$  with values in  $\mathbb{R}^d$  is said to satisfy the Fokker-Planck-Kolmogorov equation if

1.  $X$  is a diffusion process that satisfies the Itô stochastic differential equation

$$dX_t = f_t(X_t)dt + \tau_t(X_t)dW_t \quad (1.10)$$

where  $f_t$  is a  $d$ -vector valued function,  $\tau$  is a  $d \times p$  matrix function, and  $W = (W_t : 0 \leq t)$  is a  $p$ -dimensional Wiener process.

2. the initial state  $X_0$  is independent of  $W$  and has a probability density  $\pi_0$  w.r.t. the Lebesgue measure on  $\mathbb{R}^n$ , with finite moments of any order and with  $\pi_0$  almost surely positive.
3.  $f$  is once continuously differentiable with respect to  $x$  and continuous w.r.t  $t$
4. the function  $a_t(\tau_t) := \tau_t(x)\tau_t(x)^T$  is twice continuously differentiable w.r.t.  $x$  and continuous w.r.t.  $t$ .

5. there exists  $K > 0$  such that

$$2x' f_t(x) + \|a_t(x)\| \leq K(1 + |X|^2),$$

for all  $t \geq 0$ , and for all  $x \in \mathbb{R}^d$

6. the law of  $X_t$  is absolutely continuous and its density  $p_t(x)$  at  $x$  is twice continuously differentiable with respect to  $x$  and once continuously differentiable w.r.t  $t$ , and satisfy the Fokker-Planck-Kolmogorov equation

$$\frac{\partial \pi_t}{\partial t} = A_t^* \pi_t \tag{1.11}$$

where the backward diffusion operator  $A_t$  is defined by

$$A_t = \sum_{i=1}^d f_i \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^d a_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}$$

and its dual (forward) operator is given by

$$A_t^* \pi = - \sum_{i=1}^d \frac{\partial}{\partial x_i} (f_i \pi) + \frac{1}{2} \sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x_j} (a_{i,j} \pi)$$

7.  $\pi_t(x)$  is positive for all  $t \geq 0$  and almost all  $x \in \mathbb{R}^d$

*Remark.* Assumptions (3), (4), and (5) imply local Lipschitz continuity.

*Remark.* Assumption (6) holds under conditions given by the boundedness of  $f_t$  and  $a_t$  plus the uniform ellipticity of  $a_t$ , see [80, Theorem 9.1.9].

One reason for the popularity of the Fokker-Planck-Kolmogorov equation is knowledge of the differential operators  $A^*$ ,  $A$ . As a result, one can write them with respect to their eigenfunctions and develop methods to approximate their behaviors. As we shall see later, knowledge of the operators is going to play a key role in the approximation using the projection filter.

## 1.3 Linear Filters

Even though the Kushner-Stratonovich equation can be described mathematically, it is not clear whether the solution is tractable. In 1959 and independently of Stratonovich,

the linear filter in discrete time was introduced by Kalman [51], and independently by Bucy [26]. In 1961, Kalman and Bucy collaborated [52] to extend the solution to cover the continuous time and showed that the problem can be solved explicitly, that is,  $\pi$  has a closed-form solution.

The Kalman-Bucy filter is a solution to a filtering problem where the signal  $X$  is a solution of a constant coefficient stochastic differential equation with Gaussian initial condition and a linear observation function.

**Definition 1.3.1.** [12, ] A solution to the filtering problem is said to be a Kalman-Bucy filter or simply a linear filter if

1. The signal process  $X$  satisfies the evolution equation.

$$X_t = X_0 + \int_0^t (\alpha_s X_s + \beta_s) ds + \int_0^t \sigma_s dV_s \quad (1.12)$$

where, for any  $s \geq 0$ ,  $\alpha_s \in \mathbb{R}^{d \times d}$ ,  $\beta_s \in \mathbb{R}^d$ ,  $\sigma_s \in \mathbb{R}^{d \times p}$ ,  $V$  is a  $p$ -dimensional Brownian motion.

2.  $X$  takes values in  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ .
3. The distribution of  $X_0$  is  $N(x_0, r_0)$  where,  $x_0 \in \mathbb{R}^d$ , and  $r_0 \in \mathbb{R}^{d \times d}$ .
4. The observation function  $h : \mathbb{R}^d \rightarrow \mathbb{R}^m$  is a linear function.
5.  $X_0$  is independent of  $V$ .

The simplicity and tractability of the Kalman-Bucy filter led to many successful applications in the estimation and control of linear dynamical systems [12] and motivated many mathematicians and scientists to search for other filtering problems where the solution is explicit and finite dimensional.

In 1981, Beněs succeeded in extending the class of problems to include signals that satisfy stochastic differential equations with a non-constant drift term and satisfy a quite restrictive condition, known as the Beněs condition. Beněs showed that for this particular class of problems, the solution is finite-dimensional [16].

**Definition 1.3.2.** A solution to the filtering problem is said to be a Beneš filter if

1.  $X$  takes values in  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ .

2. The distribution of  $X_0$  is  $N(x_0, r_0)$  where,  $x_0 \in \mathbb{R}^d$ , and  $r_0 \in \mathbb{R}^{d \times d}$ .
3. The signal process  $X$  satisfies the evolution equation.

$$X_t = X_0 + \int_0^t f(X_s) ds + \sigma V \quad (1.13)$$

where,  $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is differentiable and globally Lipschitz (i.e.  $\exists K \in \mathbb{R}$  such that  $|f(x) - f(y)| \leq K|x - y|$ ),  $\sigma \in \mathbb{R}^{d \times p}$ , and  $V$  is a  $p$ -dimensional Brownian motion.

4.  $X_0$  is independent of  $V$ .
5. The observation function  $h : \mathbb{R}^d \rightarrow \mathbb{R}^m$  is a linear function.
6. The functions  $f, h$  satisfy the following Beneš condition

$$\frac{d}{dx} f(x) + \frac{1}{\sigma^2} f^2(x) + h^2(x) \quad (1.14)$$

equals a second-order polynomial with a positive leading-order coefficient.

It is important to note that the Kalman-Bucy filter is a special case of the Beneš filter, and satisfy the Beneš condition (1.14) [64].

## 1.4 Projection Filter

After the great success of the linear filter, others have searched for problems that could be solved in a closed form, or that have exact solutions [12]. As it turns out, the linear filter is rather special in the theory of stochastic filtering. Early results have alluded to that solutions to stochastic filters are not exact, in general.

This result was first established for the case where the observation function  $h$  is the cubic of the signal (i.e.  $h(X) = X^3$ ) [48], this is a so-called cubic sensor problem.

**Definition 1.4.1.** [48] The filtering problem is said to be a cubic sensor problem if

1. The signal process  $X$  takes values in  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ .
2.  $X$  satisfy the evolution equation

$$X_t = V_t$$

where  $V$  is a 1-dimensional Wiener process independent of  $X_0$ .

3. The observation function is  $h(X) = X^3$ , that is, the observation process  $Y$  satisfy

$$Y_t = Y_0 + \int_0^t X_s^3 ds + W_t, \quad \forall t \geq 0$$

where  $W$  is 1-dimensional Wiener process that is independent of  $X_0$ , and  $V$ .

4.  $Y_0$  is identically zero.

To extend linear theory to nonlinear settings, there was a realization that tools from differential geometry could provide a nice fit [12]. Brockett and Clark [24], and Brockett and Mitter [64] used tools from Lie theory to study under what conditions the solution to the filtering problem lies on a finite-dimensional surface. As part of this Lie theoretic approach, results have been discovered that prove, in general, that the solution to stochastic filtering problems is infinite-dimensional [24, 64].

**Definition 1.4.2.** [64, page 175] Let  $\pi_t$  be the conditional distribution process in (1.3). By a finite dimensional filter for the  $\pi_t\varphi$ , we mean a stochastic process  $Z = \{Z_t : 0 \leq t\}$  that satisfy

$$dZ_t = \alpha(Z_t)dt + \beta(Z_t)dW_t$$

defined on a finite-dimensional manifold  $\mathcal{M}$ , so that  $Z_t \in \mathcal{M}$ , and  $\alpha(Z_t)$ , and  $\beta(\xi_t)$  are smooth vector fields on  $\mathcal{M}$ , together with a smooth output map

$$\pi_t\varphi = \gamma(Z_t)$$

which computes  $\pi_t\varphi$ .

Intuitively speaking, the solution to a stochastic filtering problem is said to be finite-dimensional if the conditional distribution process evolves on an a priori finite-dimensional manifold of probability distributions.

This infinite dimensionality can easily be illustrated by considering the time evolution of the first moment for the conditional distribution  $\pi_t$  [60]. Observe that, one can write the Kushner-Stratonovich equation (1.8) as follows

$$\begin{aligned} \pi_t(\varphi) = \pi_0(\varphi) &+ \int_0^t \pi_s(A\varphi)ds \\ &+ \int_0^t \text{cov}_{\pi_t}[\varphi(X_t), h(X_t)^T] \Sigma_y^{-1} (dY_s - h(X_s)\pi_s(h)ds) \end{aligned} \tag{1.15}$$

where  $\text{cov}_{\pi_t}$  is a posterior covariance. Let  $f = A\varphi$  and set  $\varphi(X) = X$ , we can derive the time evolution of the first moment as follows:

$$\begin{aligned} \mu(t) = \mu_0 + \int_0^t \mu[f(X_t)]dt \\ + \int_0^t \text{cov}_{\pi_t}(X, h(X_t)^T)\Sigma_y^{-1}(dY_s - \mu[h(X_s)]ds) \end{aligned} \tag{1.16}$$

where  $\mu[f(X_t)], \mu[h(X_t)]$  are the mean values for the random variables  $f(X_t)$ , and  $h(X_t)$  respectively.

Note that, for any non-trivial observation function  $h$ , any moment equation will depend on higher-order moments of  $X$  due to  $\text{cov}_{\pi_t}(X, h(X_t)^T)$ , i.e. the posterior covariance between the observation function  $h$ , and the test function  $\varphi$ . In effect, this amounts to a closure problem when  $f$  is nonlinear [60]. In fact, the Fokker-Planck-Kolmogorov Equation 1.11 presents such a closure problem [60].

For such a topic with a large diversity of applications, many approximation methods have been developed, such as linearization methods, approximation using finite-dimensional filters, spectral methods, particle methods, and projection methods. Projection methods appeal to us, especially in the case where the geometry is that of statistical manifolds, because they allow a deeper understanding of the geometric structures at play and could establish a notion of optimality related to maximum likelihood estimation [23].

The projection filter was first introduced by Hanzon in [47] where he proposed to separately project the vector fields associated with the drift and the diffusion part of the Kushner-Stratonovich equation on a finite-dimensional exponential model using an induced  $L^2$  distance [9]. The projected finite-dimensional evolution process, called a projection filter, was used to approximate the evolution of the full optimal filter. The projection filter was later formulated precisely in [19], during the PhD studies of Damiano Brigo, where it was shown that exponential families played an important role allowing the correction step in the filtering algorithm to be exact [9].

**Definition 1.4.3.** (Projection Filter) A projection filter is an algorithm which provides an approximation of the conditional distribution process solving a general stochastic filtering problem by projecting it to a finite-dimensional family of probability measures.

In that period, 1995-1999, Brigo applied this method to small observation noise in nonlinear filtering [18], provided an approximation for the infinite-dimensional Fokker-Planck-Kolmogorov equation [22], and used it for volatility modeling in finance [21].

It is important to recognize in what follows that even though the projection filter used finite-dimensional exponential families, the projection chosen relies on the  $L^2$  structure on the space of square roots of densities, i.e. the map  $p \mapsto \sqrt{p}$ , which corresponds to minimizing using the Hellinger distance [71].

In 2011, Brigo returned to the problem of filtering after collaborating with John Armstrong [8], where they used an  $L^2$  structure on the space of densities themselves, that is,  $p \mapsto p$ , leading to what they termed the  $L^2$  direct metric approach [8, 7]. Under an  $L^2$  direct metric approach and as anticipated by Brigo in an earlier preprint, the  $L^2$  direct metric approach works best with mixture families rather than exponential models.

It is important to note that even though the projection method works well with mixture families of probability densities, the direct metric  $L^2$  structure is not compatible with the metric induced by the statistic manifold as developed in information geometry [23], raising the question of the statistical interpretability of the results. This highlights the point that projection methods are impacted by the geometry of the space in which the projection is taking place. As a result, the ideal space for projection to take place is the space in which the evolution equation is written [23]. This space is infinite-dimensional in general in the case of general stochastic filtering and in particular for Fokker-Planck-Kolmogorov Equation 1.11. This emphasizes the importance of studying the finite-dimensional geometry of exponential families which is thought to be better linked to their affine and convex structures, and statistically relevant objects and methods such as divergence, Fisher information, and maximum likelihood estimation.

An infinite-dimensional geometric structure on the set of all probability measures that are dominated by an arbitrary measure on an infinite measure space was introduced by Pistone in [71], and further developed in [70, 43, 27]. This infinite dimensional space contains the evolution of stochastic processes whose probability measures agree on a set of measure zero making the space an ideal place for the approximating infinite-dimensional stochastic processes by projecting them on finite-dimensional submodels including exponential and mixture families. Brigo and Pistone [23] extended the projection method by approximating the evolution of an infinite-dimensional Fokker-Planck-Kolmogorov equation by a finite-dimensional exponential and mixture submodels that are embedded in the infinite-dimensional space.

To examine the projection method in [23], in Chapter 2, we examine the key Pythagorean projection theorem needed for projecting between exponential models and the underlying structures required, such as a dually flat manifold structure and orthogonality of tangent space representations with respect to the Fisher Metric. We also highlight boundary regularity challenges when considering the smooth tangent structures.

Even though chapter 2 is a review of a well-established finite-dimensional geometry in the literature, our contribution is through rigorously presenting all the material together connecting the affine and convex geometries with the manifold and tangent geometries, which are typically discussed separately in the literature. In addition, we emphasize the nature of the boundary in the different mathematical constructions which are overlooked in the literature except for the case when the sample spaces are finite.

We follow that in Chapter 3 by reviewing, in detail, the challenges of Pistone's extension of exponential families paying close attention to boundary problems, dual structures, orthogonality of the dual tangent spaces, and a local Pythagorean projection theorem. Our contribution in this chapter is the presentation of the infinite-dimensional information geometry as an extension to the finite-dimensional geometry, highlighting the important similarity and differences between the two geometries and emphasizing the difficulties encountered as a result of the extension from finite-dimensions to infinite-dimensions.

In Chapter 4 we study properties of the projection filter using numerical examples with a focus on boundary problems, the nature of the projected evolution, and the different choices of the finite-dimensional submodel. In addition, we explore the projection filter in an infinite-like environment with more emphasis on understanding the problems in computational terms rather than analytical ones. We close by discussing the learned properties of the projection filter and highlight key questions and future research directions.

# Chapter 2

## The foundations of finite dimensional Information Geometry

### 2.1 Introduction

In this chapter, we review the foundations of finite-dimensional Information Geometry, [5], which are key to understanding dimensional reducing projections between exponential families. We look at the framework of this geometry in more detail than is usual in the literature since we need to explore Pistone's extension, [71], to the infinite-dimensional case needed for filtering problems. Since infinite-dimensional geometry can be highly technical, it was felt necessary to have a strong finite-dimensional foundation to build on.

The theory of exponential families and their geometry is, of course, a key part of Theoretical Statistics, [37]. The concept was independently introduced by Koopman [55], Pitman [72], and Darrois [35] motivated by the search for probability distributions that possess desirable statistical inference qualities, as proposed by Fisher in his groundbreaking paper [39] on maximum likelihood. After defining exponential families in Section 2.3 we look at their affine geometric properties; intuitively, they can be viewed as convex subsets of a finite-dimensional subspace of a very general affine space. Their convexity properties are described in Section 2.4. They also have differential geometric properties being smooth manifolds, see Section 2.5, whose (dual) tangent bundle and Riemannian structure are described in 2.6

One of the distinguishing features of information geometry is its so-called dual structures. This duality is related to the duality of convex geometry, but also to what Amari

calls its dual affine geometry. Both of these related notations are going to be important for the projection tools used in studying the filtering problem. A second important part of the underlying geometry is the existence of boundaries. These will play an important role in the evaluation of the methods in [23].

It may be useful to consider an intuitive, if informal, summary of the key Pythagorean projection theorem in Information Geometry with all terms being defined later in this chapter. It concerns the inferential behavior of a finite-dimensional exponential family  $p_\theta$ , when the data generation process  $p^*$  does not lie in the family. Define the asymptotic limiting distribution of the MLE in the family to be labeled  $\hat{\theta}^*$ . Then it is well-known that: (i)  $\hat{\theta}^*$  is the argmin of the Kullback-Leibler divergence from  $p^*$  to the exponential family, (ii) the exponential family is flat (affine) in the so-called (+1)-affine space, (iii) there is a curve connecting  $p^*$  and  $p_{\hat{\theta}^*}$  that is flat in the dual (-1)-affine space, and (iv) the (+1)- and (-1)-affine spaces are Fisher orthogonal. We can therefore consider the Kullback-Leibler projection as resulting in the best model-based estimate of the DGP. This motivates the projection of the solution of a filtering problem to a finite-dimensional exponential family using the Kullback-Leibler divergence proposed by [23]. The chapter then gives the formal mathematical details of this general idea.

## 2.2 Definitions

Exponential families are parametric statistical models in which the associated set of probability densities (or mass functions in the finite or countable case) can be expressed in a particular exponential form.

**Definition 2.2.1.** [13, page 111] Let  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be a probability space and  $\mathcal{P}(\mathcal{X})$  be the power set of all probability measures on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ . A subset  $P \subset \mathcal{P}(\mathcal{X})$  is said to take the exponential family form if there exists:

1. a  $\sigma$ -finite measure  $\nu_0$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  that dominates  $P$  (that is, for all  $p \in P, p \ll \nu_0$ ).
2. real-valued functions  $a, \alpha = (\alpha_1, \dots, \alpha_d)$  on  $\mathcal{P}(\mathcal{X})$  where  $d > 0$  is a positive integer.
3. real-valued measurable functions  $b, u^1, \dots, u^d$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  such that  $b \geq 0$ .

and, for every  $p \in P$

$$\frac{dp}{d\nu_0}(x) = b(x) \exp \left( \sum_{i=1}^d \alpha^i(p) u^i(x) - \log(a(p)) \right) \quad (2.1)$$

then, (2.1) is called an *exponential representation* of the densities (or mass functions – for simplicity we will just use the term density to cover both cases) of the probability measures in  $P$  with respect to the reference measure  $\nu_0$ . In addition, the representation is said to be *minimal* if the positive integer  $d$ , known as the *order* of  $P$ , denoted by  $\text{order}(P)$ , is the smallest integer such that the densities of the probability measures in  $P$  are in the exponential representation (2.1).

It is important to note that any set of probability measures in the exponential family form has several exponential representations [13]. For example, let  $c \in \mathbb{R}, c \neq 0$  then,  $c \cdot \alpha$ , and  $\frac{1}{c} \cdot u$  be another exponential representation for  $P$ . In addition, one can show that, if there exists another representation of  $P$ , i.e. there exists  $\tilde{d}, \tilde{a}, \tilde{\alpha}, \tilde{u}$  that satisfy Definition (2.2.1) then, there exists an affine transformation from  $u$  to  $\tilde{u}$ , and from  $\alpha$  to  $\tilde{\alpha}$  [13, Lemma 8.1, page 112]. However, it is not directly clear from Definition (2.2.1) why such an affine relationship exists. We will have more to say about that later.

According to the theory of exponential families, one important parameterization is the so-called canonical parameterization.

**Definition 2.2.2.** [13] Let  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be a probability space, and  $\nu_0, P, d, a, \alpha, b, u$  as per Definition (2.2.1) such that, for all  $p \in P$ ,  $P$  has the exponential representation (2.1). Let  $\Theta = \text{range}(\alpha)$ .

1. Define the map  $\mathbf{p} : \mathbb{R}^d \rightarrow P$  by  $\mathbf{p} \triangleq \alpha^{-1}$ . The parameterization  $(\Theta, \mathbf{p})$  is called *canonical*.
2. If the representation of  $p \in P$  is minimal, then  $(\Theta, \mathbf{p})$  is called *minimal canonical*.
3. The set  $\{P \ni p_\theta \triangleq \mathbf{p}(\theta) : \theta \in \Theta\}$  will denote a canonical parameterization of  $P$ .
4. Define  $K(\theta) = \log(a(p_\theta))$ . For every  $P_\theta$  in the canonical parameterization of  $P$  with

$$\frac{dp_\theta}{d\nu_0}(x) = p(x; \theta) = b(x) \exp \left( \sum_{i=1}^d \theta^i u^i(x) - K(\theta) \right) \quad (2.2)$$

being the exponential representation considered,  $S$  will stand for a (common) support of  $p(x; \theta)\nu_0$ ,  $\theta \in \Theta$ , and  $C$  is a convex closure of  $S$ .

5. If one chooses a representation such that  $0 \in \Theta$ , and  $a(0) = 1$  (which implies  $\frac{dp_0}{d\nu_0} = b$ ) then, for every  $p \in P$  represented by

$$\frac{dp_\theta}{dp_0}(x) = \exp \left( \sum_{i=1}^d \theta^i u^i(x) - K(\theta) \right) \quad (2.3)$$

is said to be in *standard representation*.

6. If (2.3) is minimal, then we say that (2.3) is *minimal standard*.
7.  $P$  is said to be *full* if the following regularity condition holds

$$\left\{ \theta : \int_{\mathcal{X}} b(x) \exp \left( \sum_{i=1}^d \theta^i u^i(x) \right) d\nu_0 < \infty \right\} = \Theta \triangleq \text{range}(\alpha) \quad (2.4)$$

8.  $P$  is said to be *regular* if it is full and if for some minimal canonical parameterization, the set  $\Theta$  is open (w.r.t the standard topology on  $\mathbb{R}^d$ ).

For the purpose of statistical inference, one is typically interested in a particular value (e.g. the maximum likelihood value) calculated in a given “preferred” parameterization. This typically requires the application of differential calculus to the parameters. In addition, we are also interested in knowing whether our statistical conclusions depend on our choice of parameterization. This line of investigation led many researchers, e.g., Rao [74], Amari [2], James [49], Dawid [36], Efron [37], Akin [1], Kass [53], Skovgaard [78], Lauritzen [62] and Barndorff-Nielsen [14], to a rich theory of statistical manifolds, now known as Information Geometry.

Because differential geometry is concerned with characterizing how differential calculus depends on a given choice of parameterization, it is natural to apply tools from differential geometry to study parametric statistical models. In addition, differential geometry can help us to understand the relationship between the geometric structure of statistical models and their statistical properties [65].

**Example 2.2.1.** To motivate the need for differential geometric tools, consider a normal family on  $\mathbb{R}$ .

$$\left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) : \mu, \sigma^2 \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

The following is one of the many exponential representations of the normal family

$$\left\{ \exp \left( \theta^1 x^2 + \theta^2 x - K(\theta) \right) : \theta^1 = -\frac{1}{2\sigma^2}, \theta^2 = \frac{\mu}{\sigma^2}, \theta^1 < 0, \theta^2 \in \mathbb{R} \right\}$$

where  $K(\theta) = \frac{1}{2} \log \left( -\frac{\pi}{\theta^1} \right) - \frac{(\theta^2)^2}{4\theta^1}$ .

From the  $(\mu, \sigma^2)$  representation, it is typical to associate the normal family with the positive half plan  $\{\mathbb{R} \times (0, \infty)\}$ ; however, it is not clear whether or not the normal family has any of the Euclidean geometry of  $\mathbb{R}^2$ . In addition, it is logical to want the statistical conclusions of the parameters to be invariant between the  $(\mu, \sigma^2)$ , and the  $(\theta^1, \theta^2)$  representations above.

## 2.3 Affine geometry

To understand the geometry of any statistical model with an exponential family form, we first need to characterize its affine geometry. As we shall show, exponential families have one of the simplest geometries possible, i.e. they are convex subsets of affine spaces with affine coordinates [65]. To make this precise, we first need to recall the definition of an affine space.

**Definition 2.3.1.** [63] An affine space is a set  $X$ , a vector space  $V$ , and a transitive left group action  $\oplus$  of the additive group  $(V, +)$  of  $V$  on  $X$ , that is, the map  $\oplus$  satisfies the following conditions:

1. For all  $x \in X$ ,  $x \oplus 0 = x$ , where  $0$  is a zero vector in  $V$ .
2. For all  $x \in X$ , and  $v, u \in V$ ,  $(x \oplus v) \oplus u = x \oplus (v + u)$
3. For all  $x \in X$ , the mapping  $V \rightarrow X : v \mapsto x \oplus v$  is a bijection.

*Remark.* For any points  $x, y \in X$ , there exists a unique vector  $v_y \in V$  such that

$$y = x \oplus v_y$$

In other words, if one chooses any element of  $X$ , say  $x_0$ , to represent the zero element, then the action of the map  $\iota_{x_0} : V \rightarrow X, v \mapsto \iota_{x_0}(v) := (x_0 \oplus v)$  establishes a one-to-one correspondence between the vectors of  $V$  and the points of  $X$ .

Exponential families are defined relative to  $\nu_0$  which lies in the space of non-negative measures  $\mathfrak{M}$ . Such a measure is not uniquely defined; rather, it can be viewed relative to the equivalence relation  $\sim_{\text{Null}}$  that any two measures are equivalent if they agree on sets of measure zero. It is shown in [65] that any equivalence class  $[\nu] \in \mathfrak{M} / \sim_{\text{Null}}$ , treated as a family  $\mathcal{M}$  of non-negative measures on  $\Omega$  dominated by  $\nu$ , is an affine space when one considers the left action of the additive group of the vector space  $\mathcal{V}$  of all real measurable functions on  $\Omega$  (with the standard addition and scalar multiplication on  $\mathbb{R}$ ) using the map  $\nu \oplus f := e^f \nu$  where  $\nu \in \mathcal{M}$ ,  $f \in \mathcal{V}$ .

**Definition 2.3.2.** Define the map  $\oplus : \mathcal{M} \times \mathcal{V} \rightarrow \mathcal{M}$  translating  $\nu_0 \in \mathcal{M}$  by  $f \in \mathcal{V}$  to  $\nu \in \mathcal{M}$  using  $\nu_0 \oplus f = e^f \nu_0 = \nu$ . Observe that:

1. for  $\nu \in \mathcal{M}$   $f, g \in \mathcal{V}$

$$(\nu \oplus f) \oplus g = e^f(e^g \nu) = e^{f+g} \nu = \nu \oplus (f + g)$$

2. for any two measures  $\nu_0, \nu \in \mathcal{M}$ , there exists a unique positive measurable function  $f \in \mathcal{V}$  such that,  $\nu = \nu_0 \oplus f$ , which is precisely  $f = \log(\nu/\nu_0)$ .
3. the mapping  $\iota_{\nu_0} : \mathcal{V} \rightarrow \mathcal{M}$ , defined by

$$f \mapsto \iota_{\nu_0}(f) \triangleq \nu_0 \oplus f = e^f \nu_0 \quad (2.5)$$

is bijective, i.e. for all  $\nu \in \mathcal{M}$ , there exists unique  $g \in \mathcal{V}$ , precisely  $g = \log(\nu/\nu_0)$  that satisfy  $\nu = \nu_0 \oplus g$ .

The construction above shows that  $(\mathcal{M}, \mathcal{V}, \oplus)$  is an affine space [65].

Let  $\mathcal{P}$  be the set of all probability measures in  $\mathcal{M}$ . As a subset of  $\mathcal{M}$ ,  $\mathcal{P}$  cannot be an affine subspace in general since it is not closed under  $\oplus$ , i.e. translating any probability measure  $p \in \mathcal{P}$  by an arbitrary vector  $f \in \mathcal{V}$  (using the map  $\oplus$ ) is not a probability measure in general (i.e. the translated measure  $p \oplus f$  might not have a unit mass or even a finite mass).

A better identification of  $\mathcal{P}$  is to consider it as a subset of another set  $\mathcal{M}$ , the set of all non-negative measures in  $\mathcal{M}$  up to scaling. This identification is possible because any probability measure corresponds to some finite non-negative measure divided by its total mass.

The relation in which two measures are scaling of another is an equivalence relation, which implies that  $\mathcal{M}$  is a quotient set. To see this, consider two non-negative measures  $\nu_1, \nu_2 \in \mathcal{M}$ , and define the equivalence relation  $\nu_1 \sim_{\text{scale}} \nu_2$  if there exists a positive constant  $\lambda \in \mathbb{R}, \lambda > 0$  such that  $\nu_1 = \lambda \nu_2$ . This implies  $\mathcal{M} = \mathcal{M} / \sim_{\text{scale}}$ . In addition, let  $p \in \mathcal{P} \subset \mathcal{M}$ , then  $p$  corresponds to some equivalence class  $[\nu] \in \mathcal{M}$ . Suppose that  $\nu$  has a total mass  $a$ , then we can write

$$p = \frac{1}{a} \nu = e^{-\log(a)} \nu$$

Note that this scaling operation is a translation of  $\nu$  by the constant real measurable function  $-\log(a)$ , which is an element of  $\mathcal{V}$ . To keep the notation simple, we will use  $\nu \in \mathcal{M}$  to mean  $[\nu] \in \mathcal{M}$ .

This identification of  $\mathcal{P} \subset \mathcal{M}$  is important since, very similar to our construction of  $(\mathcal{M}, \mathcal{V}, \oplus)$ , it can be rigorously shown that the triple  $(\mathcal{M} / \sim_{\text{scale}}, \mathcal{V}, \oplus)$  is also an affine space (which is different from  $(\mathcal{M}, \mathcal{V}, \oplus)$ ). In addition, the set  $\mathcal{P}$  is a proper subset of  $\mathcal{M}$ .

This is true since  $\mathcal{M}$  also contains equivalence classes whose members are measures with an infinite mass, i.e. the equivalence relation  $\sim_{\text{scale}}$  is well defined even if the measures have infinite mass.

The set  $\mathcal{P}$  has another affine structure, the so-called dual structure.

**Definition 2.3.3.** [63] Let  $\nu_0$  be an arbitrary measure that dominates  $\mathcal{P}$ . Consider the triple  $(\widetilde{\mathcal{M}}, \mathscr{W}, +)$  defined as follows:

1.  $\widetilde{\mathcal{M}}$  is the set of all measurable functions on  $\Omega$  that agree with  $\nu_0$  on a set of measure zero, and integrate to 1, i.e.

$$\widetilde{\mathcal{M}} = \left\{ v \in \mathscr{V} : \int_{x \in \mathcal{X}} v(x) d\nu_0(x) = 1 \right\} \quad (2.6)$$

2.  $\mathscr{W}$  is the vector space of all measurable functions (with standard addition and scalar multiplication) on  $\Omega$  that integrate to zero (with respect to  $\nu_0$ ), i.e.

$$\mathscr{W} = \left\{ f \in \mathscr{V} : \int_{x \in \mathcal{X}} f(x) d\nu_0(x) = 0 \right\} \quad (2.7)$$

3. The map  $+ : \widetilde{\mathcal{M}} \times \mathscr{W} \rightarrow \widetilde{\mathcal{M}}$  is the standard addition of functions, i.e.

$$(v + f)(x) := v(x) + f(x), \quad x \in \mathcal{X}, v \in \widetilde{\mathcal{M}}, f \in \mathscr{W} \quad (2.8)$$

Note that  $\widetilde{\mathcal{M}}$ , and  $\mathscr{W}$  are well defined and independent of  $\nu_0$ . This is guaranteed by the Radon-Nikodym Theorem up to  $\nu_0$  measure zero. Since all elements agree with  $\nu_0$  on a set of measure zero, elements of  $\widetilde{\mathcal{M}}$ , and  $\mathscr{W}$  are well defined up to any measure that dominates.  $\mathcal{P}$ .

As was shown in [63], the triple  $(\widetilde{\mathcal{M}}, \mathscr{W}, +)$  is an affine space. Furthermore, the set  $\mathcal{P}$  of all probability measures in  $\widetilde{\mathcal{M}}$  is not all of this affine space since, for any probability measure  $p \in \widetilde{\mathcal{M}}$ , addition by any measurable functions  $f \in \mathscr{W}$  could generate a measurable function that is not strictly positive over  $\mathcal{X}$ .

This highlights a few very important points regarding the nature of the set  $\mathcal{P}$ .

1. The set  $\mathcal{P}$  is a proper subset of two different possibly infinite-dimensional affine spaces  $(\mathcal{M}, \mathscr{V}, \oplus)$ , and  $(\widetilde{\mathcal{M}}, \mathscr{W}, +)$ .

2. The boundary of the set  $\mathcal{P}$  is rather complicated and depends on which affine space we are looking at.

- In the case where  $\mathcal{P}$  is a subset of  $(\mathcal{M}, \mathcal{V}, \oplus)$ , the boundary is the case where the action of vectors  $f \in \mathcal{V}$  by the map  $\oplus$  violates the unit mass property of the probability measures.
- In the other case where  $\mathcal{P}$  is a subset of  $(\widetilde{\mathcal{M}}, \mathcal{W}, +)$ , the boundary is the case where the action of  $f \in \mathcal{W}$  by the map  $+$  violates the positivity property of probability measures.

To differentiate between these two affine spaces, we follow Amari [3] by referring to  $(\mathcal{M}, \mathcal{V}, \oplus)$  as the (+1) affine space and  $(\widetilde{\mathcal{M}}, \mathcal{W}, +)$  as the (-1) affine space.

A noteworthy special case arises when the collection of all probability measures is in a “finite” affine space.

**Definition 2.3.4.** Let  $(\mathcal{M}, V, \oplus)$  be an affine space. The affine space is said to be finite dimensional if the vector space generated by the action of the vector space  $V$  is finite dimensional; otherwise, the affine space is said to be infinite dimensional. Furthermore, if the vector space  $V$  is  $d$  dimensional, we say that the affine space  $(\mathcal{M}, V, \oplus)$  is  $d$  dimensional.

By choosing an arbitrary non-negative measure  $\nu_0 \in \mathcal{M}$ , and a  $d$ -dimensional vector subspace  $V$  of  $\mathcal{V}$  spanned by  $d$  independent vectors  $u^1, \dots, u^d$ , then there exists a unique measure  $\nu_f \in \mathcal{M}$  that satisfies

$$\begin{aligned} \nu_f &= \nu_0 \oplus f \\ &= \nu_0 \oplus (\theta^1 u^1 + \dots + \theta^d u^d) \\ &= \exp(\theta^1 u^1 + \dots + \theta^d u^d) \nu_0 \end{aligned}$$

Let  $m$  be the total mass of  $\nu_f$ , that is,  $m = \int_{\mathcal{X}} \exp(\theta^1 u^1 + \dots + \theta^d u^d) \nu_0$ . If that total mass is finite, we construct the probability measure associated with  $\nu_f$  by dividing through by  $m$ .

**Definition 2.3.5.** Formally, given a positive measure  $\nu_0$  and a set of linearly independent measurable functions  $u^1, \dots, u^d$  then there is a subset of the (+1)-affine subspace defined by probability measures in  $(\mathcal{M}, V, \oplus)$  that are exactly the elements represented by

$$\begin{aligned} p &= \nu_0 \oplus (f - K_f), \quad f \in V \\ &= \exp\left(\sum_{i=1}^d \theta^i u^i - K(\theta)\right) \nu_0 \end{aligned}$$

where

$$K(\theta) := \log \int_{\mathcal{X}} \exp(\theta^1 u^1 + \cdots + \theta^d u^d) d\nu_0 < \infty \quad (2.9)$$

In essence, the set of all probability measures in any finite-dimensional affine space is a set with an exponential family form. Furthermore, this construction provides a more intuitive explanation of why the different canonical parameterizations of any set in the exponential family form are affine transformations of one another. Lastly, it is now clear that the geometry of  $P$  is deeply connected to the affine space  $(\mathcal{M}, V, \oplus)$  that itself is connected to the geometry of the finite-dimensional vector space  $V$  acting on it.

**Theorem 2.3.1.** [13, Lemma 8.1, page 112] Let  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be a probability space. Suppose that  $P$  is a set of probability measures on  $\mathcal{B}(\mathcal{X})$  in the exponential family form with the following canonical representations:

1.  $\{p_\theta : \theta \in \Theta\}$  of  $P$ , with

$$\frac{dp}{d\nu}(x) = \exp\left(\sum_{i=1}^d \theta^i u^i - K(\theta)\right)$$

where,  $u = (u^1, \dots, u^d)$  are  $d$ -independent random variables on  $\mathcal{B}(\mathcal{X})$ ,  $K : \Theta \rightarrow \mathbb{R}$  is a normalizing function, and  $\nu$  is a finite measure on  $\mathcal{B}(\mathcal{X})$ .

2.  $\{p_\xi : \xi \in \Xi\}$  of  $P$ , with

$$\frac{dp}{d\tilde{\nu}}(x) = \exp\left(\sum_{i=1}^{\tilde{d}} \xi^i v^i - J(\xi)\right)$$

where,  $v = (v^1, \dots, v^{\tilde{d}})$  are  $\tilde{d}$ -independent random variables on  $\mathcal{B}(\mathcal{X})$ ,  $J : \Xi \rightarrow \mathbb{R}$  is a normalizing function, and  $\tilde{\nu}$  is a finite measure on  $\mathcal{B}(\mathcal{X})$

Then,

- $v = Au + B, \quad A \in \mathbb{R}^{\tilde{d} \times d}, B \in \mathbb{R}^{\tilde{d} \times 1}.$
- $\xi = \tilde{A}\theta + \tilde{B} \quad \tilde{A} \in \mathbb{R}^{\tilde{d} \times d}, \tilde{B} \in \mathbb{R}^{\tilde{d} \times 1}$
- $\tilde{d}, d \geq \text{order}(P)$

where  $A, \tilde{A}, B, \tilde{B}$  are constant non-singular matrices.

## 2.4 Convex structures

Consider again  $P$ , a set of probability measures on some probability space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$ , in an exponential family form with order  $d$ , parameter space  $\Theta \subseteq \mathbb{R}^d$ , and minimal canonical parameterization  $\{p_\theta \in P : \theta \in \Theta\}$ . As shown in Section 2.3,  $P$  is a subset of some  $d$  dimensional (+1) affine space  $(\mathcal{M}, V, \oplus)$ .  $P$  has two important convex geometries. The first convex structure is related to the map  $\oplus$  (2.5).

**Proposition 2.4.1.**  $P$  is a proper convex subset of the affine space  $(\mathcal{M}, V, \oplus)$  (with respect to the operator  $\oplus$ ).

*Proof.* Choose an arbitrary measure  $\nu_0 \in \mathcal{M}$ , and let  $p_f, p_g \in P$  be two probability measures associated with the two finite non-negative measures  $\nu_f, \nu_g \in \mathcal{M}$  with total masses  $e^{K_f}, e^{K_g}$  respectively, i.e.

$$e^{K_f} = \int_{\mathcal{X}} e^f \nu_0 < \infty \quad (2.10)$$

$$e^{K_g} = \int_{\mathcal{X}} e^g \nu_0 < \infty \quad (2.11)$$

As we have shown in Section 2.3, the map  $\nu_{\nu_0} := (\nu_0 \oplus \cdot)$  is a bijection between  $V$  and  $\mathcal{M}$ , and we can associate  $\nu_f, \nu_g$  with the two vectors  $f, g \in V$ .

Note that, for all  $t \in [0, 1]$ , by the properties of the vector space  $V$ ,  $tf + (1-t)g \in V$ . Let  $h$  be any vector in the line segment  $\{tf + (1-t)g \in V : t \in [0, 1]\} \subset V$ . By construction of the (+1) affine space, there exists a unique measure  $\nu_h$  such that  $\nu_h = \nu_0 \oplus h$ . i.e.

$$\begin{aligned} \nu_h &= \nu_0 \oplus (tf + (1-t)g), \quad \text{for some } t \in [0, 1] \\ &= \exp\{tf + (1-t)g\} \nu_0 \end{aligned}$$

In the trivial case  $t \in \{0, 1\}$ , it follows immediately that  $\nu_h$  is finite, that is,  $\nu_0 \oplus e^{K_h}$  is a probability measure in  $P$ . If  $t \in (0, 1)$ , by Hölder's inequality, we have

$$\begin{aligned} \int_{\mathcal{X}} \exp\{tf + (1-t)g\} \nu_0 &\leq \left( \int_{\mathcal{X}} (e^{tf})^{\frac{1}{t}} \nu_0 \right)^t \cdot \left( \int_{\mathcal{X}} (e^{(1-t)g})^{\frac{1}{1-t}} \nu_0 \right)^{1-t} \\ &\leq \left( \int_{\mathcal{X}} e^f \nu_0 \right)^t \cdot \left( \int_{\mathcal{X}} e^g \nu_0 \right)^{1-t} \\ &< \infty \end{aligned}$$

In other words, the set  $\{p_{tf+(1-t)g} = \nu_0 \oplus (tf + (1-t)g) \in \mathcal{M} : t \in [0, 1]\}$  is a subset of  $P$ .  $\square$

**Theorem 2.4.2.** [13, page 116] For any set in the exponential family form that is also full, any of the possible minimal canonical parameter domains is convex.

The set  $P$  has another convex structure that is related to the triple  $(\widetilde{\mathcal{M}}, \mathscr{W}, +)$  constructed in the previous section and is denoted by the  $(-1)$  affine geometry. Not only does this dual structure establish an additional notion of the convexity of  $P$ , it also connects Information Geometry to the notions of entropy, and other measures of information from Information Theory. In addition, understanding the relationship between the two convex structures and their related spaces can provide us with a deeper understanding of the relationship between the dualistic geometry of  $P$  and the important notion of divergence.

Within convex geometry, there is a notion of duality which is defined using the Legendre transform. Intuitively speaking, since the cumulant generating function  $K$  is a convex and smooth function over  $\Theta$  (or the restriction of  $K$  to the interior of  $\Theta$  in the case that  $P$  is not regular), the normal vector  $\nabla K$  of the tangent supporting hyperplane to  $K$  at a point  $\theta \in \Theta$  provides us with a unique one-to-one and a smooth mapping between  $\Theta$  and the range of  $\nabla K$  [5], where  $\nabla$  is the gradient operator.

An important question arises whether there exists a function  $K^*$  on the range of  $\nabla K$  whose gradient is the inverse of  $\nabla K$ , i.e.

$$(\nabla K^* \circ \nabla K)(\theta) = \theta, \quad \forall \theta \in \Theta \quad (2.12)$$

In convex analysis, the above requirement is exactly a defining characteristic of the Legendre transformation.

**Definition 2.4.1.** [75, page 256] Let  $\Theta \subseteq \mathbb{R}^d$  be open and let the map  $K : \Theta \rightarrow \mathbb{R}$  be differentiable. The Legendre conjugate of the pair  $(\Theta, K)$  is defined as the pair  $(\Theta^*, K^*)$ , where  $\Theta^*$  is an image of  $\Theta$  under the gradient mapping  $\nabla K$ , and  $K^*$  is a function on  $\Theta^*$  given by

$$K^*(\theta^*) = \sum_{i=1}^d (\nabla K)^{-1}(\theta^*)^i \cdot \theta^{*i} - (K \circ (\nabla K)^{-1})(\theta^*) \quad (2.13)$$

where  $\nabla$  is a gradient operator. If  $K^*$  is well defined (i.e.  $\nabla K$  is invertible), then passing from  $(\Theta, K)$  to  $(\Theta^*, K^*)$  is called a Legendre transformation.

*Remark.* (1) If  $K^*$  is well defined, then  $\nabla K$  is invertible. By the definition of  $\theta^*$ , we have

$$\theta = (\nabla K)^{-1}(\theta^*) \quad (2.14)$$

As a result, we can rewrite  $K^*$  as follows

$$K^*(\theta^*) = \sum_{i=1}^d \theta^i \cdot (\theta^*)^i - K(\theta) \quad (2.15)$$

*Remark.* (2) If  $K^*$  is well defined, we have

$$\nabla K^*(\theta^*) = (\nabla K)^{-1}(\theta^*)$$

By combining the above results with that of (2.14), we also have

$$\theta = \nabla K^*(\theta^*) \quad (2.16)$$

An important result from convex analysis guarantees that, if the set  $\Xi \subset \mathbb{R}^d$  is convex, and a map  $J : \Xi \rightarrow \mathbb{R}$  is a closed proper convex function that is also differentiable on  $\Xi$ , then the Legendre conjugate of  $(\Xi, J)$  is well defined [75, Theorem 26.4, page 256]. As a result, one can speak of the Legendre transformation  $(\Theta^*, K^*)$  of  $(\Theta, K)$ , where  $K$  is the cumulant generating function of the set  $P$ . In addition, since the map  $\nabla K$  is bijective, we can use the values  $\theta^* \in \Theta^*$  as another way to parameterize the set  $P = \{p_\theta \in P : \theta \in \Theta\}$ .

From the above discussion, it might seem that the set  $\Theta^*$  and the map  $K^*$  are mere mathematical by products with no statistical interpretations. In fact, the parameters  $\theta^* \in \Theta^*$  are well known in statistics and play a significant role in statistical inference.

**Theorem 2.4.3.** [13] Let  $\Omega = (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be a probability space. Suppose that  $P$  is a set of probability measures on  $\Omega$  in the exponential family form with canonical parameterization  $\{p_\theta \in P : \theta \in \Theta \subseteq \mathbb{R}^d\}$ . Let  $K$  be the cumulant generating function of  $P$ , and  $u = (u^1, \dots, u^d)$  be a canonical sufficient statistic. If  $P$  is full, then

$$\theta^* = \mathbb{E}_\theta[u], \quad \forall \theta^* \in \Theta^* \quad (2.17)$$

where  $\mathbb{E}_\theta[u^i] \triangleq \int_{x \in \mathcal{X}} u^i(x) dp_\theta(x)$ . Even more, if  $P$  is regular, then the above (2.17) equality holds for all  $\theta \in \Theta$ .

*Proof.* Let  $\nu_0$  be a finite carrying measure of  $P$ . Since  $K$  is a smooth function over  $\Theta$  (with restriction to the interior of  $\Theta$  if  $P$  is not regular). Then, by the definition of  $\theta^*$  we have

$$\begin{aligned}
(\theta^*)^i &= \frac{\partial}{\partial \theta^i} K(\theta) \\
&= \frac{1}{\int_{\mathcal{X}} \exp(\sum_{i=1}^d \theta^i u^i) d\nu_0} \int_{\mathcal{X}} u^i \exp(\sum_{i=1}^d \theta^i u^i) d\nu_0 \\
&= \int_{\mathcal{X}} u^i \exp(\sum_{i=1}^d \theta^i u^i - K(\theta)) d\nu_0 \\
&= \int_{x \in \mathcal{X}} u^i(x) dp_{\theta}(x) \\
&= \mathbb{E}_{\theta}[u^i]
\end{aligned}$$

□

Theorem 2.4.3 establishes that for every minimal canonical parameterization of  $P$ , there exists a dual parameterization (in the sense of convex analysis) of  $P$  where one is the Legendre transformation of the other. In addition, note that  $(u^1, \dots, u^d)$  are a basis for a vector subspace of the vector space  $\mathscr{W}$  of the affine space  $(\widetilde{\mathcal{M}}, \mathscr{W}, +)$ . This in fact establishes that the triple  $(\widetilde{\mathcal{M}}, W, +)$  where  $W$  is precisely the vector subspace spanned by  $(u^1, \dots, u^d)$  and establishes that  $(\widetilde{\mathcal{M}}, W, +)$  is a finite-dimensional affine space of the same dimension as  $W$ .

This establishes the dual convex nature of  $P$  which is a general property of statistical manifolds [2] characterized in the information geometry literature. Following Amari [3], we will denote the parameter space  $\Theta$  by the (+1) parameter space, and the parameter space  $\Theta^*$  by the (-1) parameter space [3].

Before we conclude this section, we turn our attention to the map  $K^*$  and discuss a strong connection between the dual affine space structure of  $P$  and the important notion of divergence.

The map  $K^*$  assigns for each parameter  $\theta^* \in \Theta^*$ , a well-known negative entropy of the probability measure  $p_{\theta} \in P$  [46], where  $\theta = \nabla K^*(\theta^*)$ .

**Theorem 2.4.4.** Suppose that  $P, K, \Theta$  are as stated in (2.4.3). Let  $(\Theta^*, K^*)$  be the Legendre transform of  $(\Theta, K)$ . Then  $K^*(\theta^*)$  is the negative entropy of the identity random variable  $I : \mathcal{X} \rightarrow \mathcal{X} : I(x) = x, \forall x \in \mathcal{X}$  with respect to the probability measure  $p_{\theta}$ ,  $\theta = \nabla K^*(\theta^*)$ .

*Proof.* Let  $\nu_0$  be a finite carrying measure on  $\Omega$  that dominated  $P$ . Using the form (2.15) of  $K^*$ , and by theorem (2.4.3), we have

$$\begin{aligned}
K^*(\theta^*) &= \sum_{i=1}^d \theta^i \cdot (\theta^*)^i - K(\theta) \\
&= \sum_{i=1}^d \theta^i \cdot \mathbb{E}_\theta[u^i] - K(\theta) \\
&= \mathbb{E}_\theta \left[ \sum_{i=1}^d \theta^i u^i - K(\theta) \right] \\
&= \int_{\mathcal{X}} \left( \sum_{i=1}^d \theta^i u^i - K(\theta) \right) \exp \left( \sum_{i=1}^d \theta^i u^i - K(\theta) \right) d\nu_0 \\
&= - \int_{x \in \mathcal{X}} p_\theta(x) \log \left( \frac{1}{p_\theta(x)} \right) d\nu_0(x)
\end{aligned}$$

for all  $\theta^* \in \Theta^*$ . □

The concept of divergence has a rich history in information theory and statistical inference with many applications in signal processing, optimization, machine learning, and quantum mechanics [15]. Intuitively speaking, divergence measures some degree of separation between two elements of a set. In optimization, we are given a set of signals  $M$ , and an element  $p \in M$  and we want to find the element  $q$  in  $M$  that minimizes some divergence of interest from  $p$  to  $q$ . In contrast to the notion of a distance (or metric), a divergence on a set is not required to satisfy the triangle inequality and is allowed to be asymmetric [15].

For a set of probability measures, an early and important divergence from one probability measure to another was introduced in 1951 by Kullback and Leibler [56].

**Definition 2.4.2.** [15] Let  $p, q$  be two probability measures on some probability space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  such that  $p$  and  $q$  agree on a set of measure zero. The divergence from  $q$  to  $p$  is defined by

$$D_{KL}(p, q) = \int_{\mathcal{X}} \log \left( \frac{dp}{dq} \right) dp \tag{2.18}$$

where  $\frac{dp}{dq}$  is the Radon–Nikodym derivative of  $p$  with respect to  $q$ .

*Remark.* (1) It is important to note that, the Kullback Leibler divergence between two probability measures is only defined if the two probability measures agree on a set of

measure zero. This is obvious from the definition, since the Radon-Nikodym derivative of one measure with respect to another is defined only if the two measures agree on a set of measure zero.

*Remark.* (2) In addition, note that, for an infinite set of probability measures,  $D_{KL}$  could be unbounded.

Kullback and Leibler [56] motivated the definition of  $D_{KL}$  by generalizing the concept of Shannon's information [76] and the similar concept of Wiener's information [82]. The definition of  $D_{KL}$  from  $q$  to  $p$  was motivated as a measure of the average information to discriminate between the hypothesis that an observation  $x \in \mathcal{X}$  is from a population with a true probability measure  $p$ , was from the true population, against the hypothesis that  $x$  was from a population with a probability measure  $q$  [56].

The divergence  $D_{KL}$  is important in statistical applications due to the link between maximum likelihood estimates and the Kullback-Leibler divergence from one probability measure to another in a family probability measures with an exponential family form.

**Definition 2.4.3.** Let  $\Omega = (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be a probability space and  $\{p_\theta(x_i), \theta \in \Theta\}$  be a parametric family of mutually continuous probability measures on  $\Omega$  with parameter space  $\Theta$ . Suppose that  $\mathcal{D} = \{x_1, \dots, x_n\}$  are independent and identically distributed observations drawn from the probability measure  $p_{\theta_{\text{true}}} \in P$ . The log-likelihood function given the data is a map  $\ell_{\mathcal{D}} : \Theta \rightarrow \mathbb{R}$  defined by

$$\ell_{\mathcal{D}}(\theta) = \log \left\{ \prod_{i=1}^n p_\theta(x_i) \right\}, \quad \theta \in \Theta \quad (2.19)$$

In addition, suppose that  $\ell_{\mathcal{D}}$  is well defined over  $\Theta$ , the maximum likelihood estimate  $\hat{\theta}_{\text{mle}}$  is the value of the parameter that maximizes  $\ell$  over  $\Theta$ .

$$\hat{\theta}_{\text{mle}} \triangleq \operatorname{argmax}_{\theta \in \Theta} \ell_{\mathcal{D}}(\theta) \quad (2.20)$$

**Proposition 2.4.5.** Suppose that  $P$  has an exponential family form (2.2.2). Maximizing the (log) likelihood function is equivalent to minimizing the Kullback-Leibler divergence over  $P$  relative to the true probability measure  $p$ .

This relationship establishes that we can focus our attention on minimizing the Kullback-Leibler divergence in applications. As we shall see later, minimizing the Kullback-Leibler divergence has strong geometric interpretations and will aid us in approximating an unknown probability measure  $p \in P$  using a subfamily  $S$  of  $P$ .

In the case where  $P$  has an exponential family form, the Kullback-Leibler divergence is related to the cumulant generating function (and its convex conjugate) of the family. To see this, we need to introduce the notion of Bregman divergence (in which  $D_{KL}$  is a special case).

In 1967, Bregman [17] introduced a more general class of divergences generated by convex functions to solve convex optimization problems.

**Definition 2.4.4.** [15] Let  $\mathcal{P}$  be a convex set of probability measures on some probability space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  such that any two elements  $p, q \in \mathcal{P}$  agree on a set of measure zero. For a continuously differentiable, strictly convex function  $\psi : \mathcal{P} \rightarrow \mathbb{R}$ , the Bregman divergence associated with  $\psi$  from  $q$  to  $p$  is the difference between the value of  $\psi$  at  $p$  and the value of the first order Taylor expansion of  $\psi$  around  $q$  evaluated at  $p$ :

$$D_\psi(p, q) = \int_{x \in \mathcal{X}} \psi(p(x)) - \left( \psi(q(x)) + (p(x) - q(x)) \cdot \nabla \psi(q(x)) \right) dp(x) \quad (2.21)$$

Intuitively, we can think of the Bregman divergence from  $q$  to  $p$  with respect to a convex function  $\psi$  as a measure of separation between  $p$  and  $q$  measured using the value of  $\psi$  at  $p$  and the value of the hyperplane tangent to  $\psi$  at  $q$  evaluated at  $p$  [5].

In convex analysis, Bregman divergences are of special interest because of a duality argument. Suppose that  $\psi$  is a strictly convex function defined on some convex set  $\Theta$ . Denote the convex conjugate of  $\psi$  by  $\psi^*$ , and let  $\Theta^*$  be the image of  $\Theta$  under  $\nabla \psi$ . The Bregman divergence derived from  $\psi$  is related to the Bregman divergence derived from the convex conjugate  $\psi^*$  of  $\psi$  by the following relation.

$$D_{\psi^*}(p^*, q^*) = D_\psi(q, p) \quad (2.22)$$

where  $p^* := \nabla \psi(p)$ ,  $q^* := \nabla \psi(q)$

For exponential families, it is clear that the convex maps  $K, K^*$  generate a Bregman divergence that is represented in the (+1) parameter space as the normalization constant with a dual in the (-1) parameter space of the Bregman divergence generated by  $K^*$ . One can easily show that the Bregman divergence  $D_{K^*}$  is the reverse Kullback-Leibler divergence on  $P$  [5]. As a result, we will be interested in minimizing the divergence relative to convex functions  $K$  (and its convex conjugate  $K^*$ ) which, as we have shown earlier, is related to maximum likelihood estimation methods.

## 2.5 Manifold structure

Let  $\Omega = (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be a probability space and  $P$  be a set of probability measures on  $\mathcal{B}(\mathcal{X})$  in the exponential family form with order  $d$ , and minimal canonical parameterization  $\{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ .

As we have shown above, the set  $P$  has dualistic affine structures. Let  $(\mathcal{M}, V, \oplus)$ , and  $(\widetilde{\mathcal{M}}, W, +)$  be the  $(+1)$  and  $(-1)$  affine spaces associated with  $P$  respectively. If we regard elements  $p \in P$  as points on a “smooth subset” of  $P$  embedded in the  $(+1)$  affine space, we can intuitively think of the  $(+1)$  parameters associated with every point  $p$  as a smooth function over  $P$  that varies over  $\Theta$ . This suggests that we should attempt to characterize  $P$  as a finite-dimensional smooth manifold. To make this argument rigorous, recall the definition of a topological  $d$  manifold, which is presented below.

**Definition 2.5.1.** [61] Let  $(P, \mathcal{O}(P))$  be a topological space, i.e.  $P$  is a set, and  $\mathcal{O}(P)$  is a topology on  $P$  (i.e. a collection of subsets of  $P$  satisfying the topology axioms). The pair  $(P, \mathcal{O}(P))$  is called a topological  $d$  manifold (or a  $d$ -dimensional topological manifold) if

1.  $(P, \mathcal{O}(P))$  is a Hausdorff topological space.
2.  $(P, \mathcal{O}(P))$  is a topological space in which every open cover has an open refinement that is locally finite (i.e., para-compact).
3. For every point  $p \in P$ , there exists a neighborhood  $\mathcal{U}_P \in \mathcal{O}(P)$  containing  $P$  such that  $\mathcal{U}_P$  is homeomorphic to an open subset of the Euclidean space  $\mathbb{R}^d$  (with respect to the standard topology  $\mathcal{O}(\mathbb{R}^d)$  on  $\mathbb{R}^d$ ).

One way to induce a topology on our set  $P$  is using any vector space topology of  $V$  (that is compatible with  $W$ , the vector space in the  $(-1)$  affine structure of  $P$ ).

**Definition 2.5.2.** Suppose that  $(V, \mathcal{O}(V))$  is a topology. Then

1. Fix any  $p_0 \in P$ .
2. Let  $O_V \in \mathcal{O}(V)$ . Define  $p_0 + O_V := \{p_0 \oplus f : f \in O_V\} \subseteq \mathcal{M}$ .
3. Define  $\mathcal{O}(P) := \{p_0 + O_V : O_V \in \mathcal{O}(V)\}$ , i.e. subsets of  $P$  associated with open sets of  $V$ .
4. Because the map  $f \mapsto p_0 \oplus f$  (2.5) is a bijection, it follows that  $(P, \mathcal{O}(P))$  is a topology.

Suppose that we endow  $P$  with a standard weak topology (induced from a weak vector space topology of  $\mathcal{V}$ ). Then, we can talk about continuous maps between  $(P, \mathcal{O}(P))$  and  $(\mathbb{R}^d, \mathcal{O}(\mathbb{R}^d))$ . The following theorem establishes that the map  $\alpha : P \rightarrow \mathbb{R}^d$  is a homeomorphism.

**Theorem 2.5.1.** [13, Theorem 8.3, page 120] Suppose that  $P$  is a set of probability measures on the probability space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P)$  in the exponential family form as per (2.2.1) that is also full (i.e., condition 2.4 holds). In addition, suppose that the canonical parameterization  $\{\mathbb{P}_\theta : \theta \in \Theta\}$  is minimal. We can endow  $\Theta$  with the usual Euclidean topology and  $P$  with the weak topology, then the following statement are true

1. The mapping  $\alpha^{-1} : \theta \mapsto P_\theta$  is continuous on the interior of  $\Theta \subset \mathbb{R}^d$ .
2. If there exists a minimal canonical statistic which is continuous, i.e.

$$u : \mathcal{X} \rightarrow \mathbb{R}^d, u = (u^1, \dots, u^d) \text{ is continuous.} \quad (2.23)$$

then  $\alpha$  is continuous on  $P$ .

This theorem outlines the important regularities required for the topological space  $(P, \mathcal{O}(P))$  to be regarded as a finite-dimensional topological manifold (with respect to the weak topology inherited from  $V$ ). In the case where  $P$  is regular and has a continuous minimal canonical statistic, then  $(P, \mathcal{O}(P))$  is precisely a  $d$ -dimensional topological manifold. On the other hand, if  $P$  is full but not regular (that is,  $\Theta$  is not open) by restricting ourselves to the interior of  $\Theta$ , then the interior of  $P$  (the set  $\check{P} \triangleq \{P_\theta : \theta \in \text{interior of } \Theta\}$ ) is topological  $d$  manifold. In differential geometry, the map  $\alpha$  is known as a coordinate map and the pair  $(\Theta, \alpha)$  is called a coordinate chart.

Going forward, whenever we consider differentiable structures on  $P$ , we will assume that  $P$  satisfies the above two conditions (2.4), and (2.23). In addition, we will abuse notation and use  $P$  instead of  $\check{P}$  whenever  $P$  is not regular, which should be clear from context.

Another important detail is a consequence of the fact that any canonical parameters are globally defined over  $P$ . In differential geometry, a coordinate map need not be defined over the entire manifold but rather over a local patch (which is an open set according to the chosen topology of the manifold). It is important to require many such local patches to cover the manifold and to require that the associated coordinate maps be compatible in some sense whenever we move from one local patch to another. For example, suppose that a topological manifold has two coordinate maps  $(\mathcal{U}_a, \alpha_a), (\mathcal{U}_b, \alpha_b)$  such that  $\mathcal{U}_a \cup \mathcal{U}_b$  covers

the manifold. It is important to restrict ourselves to structures that are compatible with any possible coordinate transition maps, i.e.  $\alpha_b \circ \alpha_a^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , or  $\alpha_a \circ \alpha_b^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  (where  $\circ$  denotes the composition of maps).

In our case of  $P$ , the coordinate maps  $(\Theta, \alpha)$  are defined over the entire space  $P$  (by restricting to its interior in the non-regular case). As a consequence, studying the geometry of  $P$  in a single coordinate chart is possible; however, care must be taken to ensure that the geometric structures on  $P$  do not depend on one chosen parameterization, i.e. we still need to restrict ourselves to geometric structures that are compatible with coordinate transition maps of  $P$ . In particular, if we are interested in studying smooth maps over  $P$ , we need to ensure that all parameterizations considered have smooth transition maps. Going forward, since all the charts of  $P$  are globally defined, we will avoid the need to define local neighborhoods for points and use  $P$  instead.

So far, we have only established the notion of continuity on  $(P, \mathcal{O}(P))$ . However, the notion of smoothness follows directly by restricting our coordinate maps to the parameterizations of  $P$  that are minimal canonical and their dual (in the sense presented in Section 2.4). This restriction implies smoothness of all chart transition maps because

1. the transition map between any two minimal canonical parameterization of  $P$  is an affine transformation, and it is also smooth.
2. the Legendre transformation from one minimal canonical parameterization to its convex dual is smooth by definition; and
3. the transition map from one minimal canonical parameterization to the convex dual of another is smooth, since the composition of smooth maps is smooth.

From the above results, we can conclude that all coordinate maps are diffeomorphisms between the manifold  $P$  and open subsets of the Euclidean space  $\mathbb{R}^d$ . In the language of differential geometry, we say that our manifold is endowed with a smooth structure or is a smooth  $d$  manifold.

## 2.6 Tangent spaces

We start the investigation of the tangent structure of the  $d$ -dimensional smooth manifold  $(P, \mathcal{O}(P))$ , constructed in the previous section, by characterizing the tangent space at an arbitrary point  $p \in P$ .

Before doing this, we need the notion of a curve on a manifold.

**Definition 2.6.1.** Let  $P$  be a  $d$ -dimensional smooth manifold. A continuous curve on  $P$  passing through a point  $p \in P$  is any continuous map  $\gamma : \mathbb{R} \rightarrow P$ . The curve is said to be centered at  $p$  if  $\gamma(0) = p$ . Suppose  $\theta$  is a chart map of  $P$ . The representation of the curve under the chart map  $\theta$  is the map  $\theta \circ \gamma : \mathbb{R} \rightarrow \mathbb{R}^d$ . The curve is said to be smooth if the representation of the curve under any chart of  $P$  is smooth in the traditional sense.

*Remark.* Since we can always center a curve on a manifold at a point by reparameterizing the curve, going forward we will assume that all curves passing through a point are centered at that point.

One of the most geometrically intuitive ways to characterize the tangent space at a point of any manifold is through the space of the velocities of all smooth curves passing through that point.

**Definition 2.6.2.** [61] Let  $P$  be a  $d$ -dimensional manifold. A map  $f : P \rightarrow \mathbb{R}^d$  is said to be smooth if, for any chart map  $\theta$  of  $P$ , the map

$$f \circ \theta^{-1} \rightarrow \mathbb{R}^d$$

is smooth in the traditional sense.

*Remark.* The set of all real smooth maps over a manifold  $P$ , denoted by  $C^\infty(P)$ , is a vector space when equipped with a point-wise addition and a scalar multiplication.

**Definition 2.6.3.** [61] Let  $(P, \mathcal{O}(P))$  be a smooth manifold. The velocity of a curve  $\gamma$  on  $P$  passing through  $p$  is the map  $v_{\gamma,p} : C^\infty(P) \rightarrow \mathbb{R}$  defined by

$$v_{\gamma,p}(f) = \frac{d}{dt}(f \circ \gamma)(0), \quad f \in C^\infty(P) \tag{2.24}$$

*Remark.* Since  $C^\infty(P)$  is a vector space, we can speak of linear maps from  $C^\infty(P)$  to  $\mathbb{R}$ . An important result from undergraduate differential geometry shows that, the velocity of a curve  $\gamma$  on  $P$  at a point  $p$  is a linear map [61]. As a result, we will adopt the notation  $v_{\gamma,p}f$  instead of  $v_{\gamma,p}(f)$ .

Denote the set of all smooth curves on  $P$  passing through  $p$  by  $\mathcal{R}_p(P)$ . By definition, we can identify the velocities at  $p$  with smooth curves passing through it. As a result, to uniquely identify velocities at  $p$  of  $P$ , we can use an equivalence relation on  $\mathcal{R}_p(P)$ . Consider the equivalence relation  $\sim_{\text{velocity}}$  that any two curves are equivalent if they have the same velocity. Then, the set of equivalence classes  $\mathcal{R}_p(P) / \sim_{\text{velocity}}$  contains all the velocities to  $P$  at  $p$ .

**Definition 2.6.4.** [61] Let  $(P, \mathcal{O}(P))$  be a  $d$ -dimensional smooth manifold. A tangent vector  $v_{\gamma,p}$  of  $P$  at a point  $p$  is an equivalence class  $[\gamma]$  of  $\mathcal{R}_p(P)/\sim_{\text{velocity}}$ , where  $\gamma$  is a smooth curve on  $P$  passing through  $p$ .

**Definition 2.6.5.** [61] A tangent space to a point  $p$  in a  $d$  dimensional smooth manifold  $P$ , denoted by  $T_pP$ , is the set of all tangent vectors of  $P$  at  $p$ .

*Remark.* In the above definition,  $T_pP$  is exactly the set  $\mathcal{R}_p(P)/\sim_{\text{velocity}}$ .

For our  $d$ -dimensional manifold  $P$  of probability measures on some probability space  $\Omega = (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  with an exponential family form, the tangent space is rather simple. This simplicity is a consequence of the fact that the tangent space of any finite-dimensional vector space  $V$  at any point  $v \in V$  is the vector space  $V$  itself.

Suppose that our manifold  $P$  of probability measures on  $\Omega$  has the  $(+1)$  affine space  $(\mathcal{M}, V, \oplus)$ . By fixing an arbitrary point  $p$  in the manifold  $P$ , the set  $\{p \oplus v : v \in V\}$  is a vector space isomorphic to  $V$  (where  $p$  plays the role of the zero vector). As a result, the tangent space of  $T_pP$  is exactly the vector space  $V$  (using the  $(+1)$  affine structure of  $P$ ). With a similar argument, it follows that  $P$  has another tangent space related to the  $(-1)$  affine structure of  $P$ . This highlights the following important points regarding the nature of the tangent space of  $P$  at any point  $p$ .

1.  $T_pP$  has two dual representations induced by the  $(+1)$  and  $(-1)$  affine structures of  $P$  [3].
2.  $T_pP$  is a  $d$ -dimensional vector space. Even though this is automatic from the theory of differential geometry (the tangent space of any  $d$ -dimensional (sufficiently) smooth manifold is a  $d$ -dimensional vector space), the above construction highlights the fact that, a basis of  $V$  is also a basis for  $T_pP$ .
3. Since our manifold is a constructed from a set of probability measures on a probability space,  $T_pP$  is a finite dimensional vector space of random variables. This is a direct consequence of the fact that  $V$  is a vector space of measurable functions on a probability space.

So far, our characterization of the tangent space to a manifold  $P$  at a point  $p$  is rather abstract. To make it more concrete, suppose that  $\theta$  is the coordinate map of  $P$  varying over an open subset  $\Theta$  of  $\mathbb{R}^d$ . Suppose  $v_{\gamma,p} \in T_pP$  is a tangent vector at  $p$ . By the definition

of  $v_{\gamma,p}$

$$\begin{aligned} v_{\gamma,p}f &= \frac{d}{dt}(f \circ \gamma)(0), \quad f \in C^\infty(P) \\ &= \frac{d}{dt}(f \circ \theta^{-1} \circ \theta \circ \gamma)(0) \end{aligned}$$

where  $\gamma$  is a smooth curve in  $P$  centered at  $p$ . By the multi-dimensional chain rule, we have

$$v_{\gamma,p}f = \sum_{i=1}^d \left\{ \left. \frac{\partial}{\partial \theta^i} (f \circ \theta^{-1}) \right|_{\theta(p)} \cdot \frac{d}{dt}(\theta \circ \gamma)^i(0) \right\}$$

where  $\theta(p)$  is the coordinate of the point  $p$ . To simplify notation, let  $v_\theta^i = \frac{d}{dt}(\theta \circ \gamma)^i$ , and define

$$\left( \frac{\partial}{\partial \theta^i} \right)_p (f) := \left. \frac{\partial}{\partial \theta^i} (f \circ \theta^{-1}) \right|_{\theta(p)} \quad (2.25)$$

As a result, we can write

$$v_{\gamma,p} = \sum_{i=1}^d v_\theta^i \cdot \left( \frac{\partial}{\partial \theta^i} \right)_p \quad (2.26)$$

The above (2.26) relation implies that the tangent vector of  $P$  at  $p$  has components maps  $\left( \frac{\partial}{\partial \theta^i} \right)_p$  with real coefficients  $v_\theta^i$  in the  $\theta$  coordinates.

**Definition 2.6.6.** We call  $v_\theta^i$  the  $i^{\text{th}}$  component of the velocity vector at point  $p \in P$  with respect to the coordinate chart map  $\theta$ .

**Definition 2.6.7.** We call  $\left( \frac{\partial}{\partial \theta^i} \right)_p$  the  $i^{\text{th}}$  basis element of  $T_pP$  with respect to the coordinate chart map  $\theta$ .

A general result in differential geometry shows that the set

$$\left\{ \left( \frac{\partial}{\partial \theta^1} \right)_p, \dots, \left( \frac{\partial}{\partial \theta^d} \right)_p \right\} \quad (2.27)$$

known as the  $\theta$ -chart induced basis for the tangent space  $T_pP$  is indeed a basis for the vector space  $T_pP$ . In addition, suppose that  $P$  has another coordinate map  $\xi$  that varies over  $\Xi$ . It directly follows that we can write any tangent vector  $v \in T_pP$  with respect to the  $\xi$ -chart induced basis as follows

$$v = \sum_{i=1}^d v_{\xi}^i \cdot \left( \frac{\partial}{\partial \xi^i} \right)_p$$

It is trivial to show that, changing from  $\partial/\partial\theta^i$  to  $\partial/\partial\xi^j$  we multiply by a jacobian matrix with entries  $\partial\theta^j/\partial\xi^i$ . In other words, we have the following relation between the different chart induced basis of tangent vectors in  $T_pP$

$$e_p^i = \sum_{j=1}^d (G(p))_{j,i} (e_p^*)^j, \quad e_p^i = \left( \frac{\partial}{\partial \theta^i} \right)_p, \quad (e_p^*)^j = \left( \frac{\partial}{\partial \xi^j} \right)_p \quad (2.28)$$

where  $G(p)$  is the Jacobian matrix  $\partial\xi^j/\partial\theta^i$  at the point  $p \in P$ . Similarly, one can easily show

$$(e_p^*)^j = \sum_{i=1}^d (G(p)^*)_{i,j} e_p^i \quad (2.29)$$

where  $G(p)^*$  is the Jacobian matrix  $\partial\theta^i/\partial\xi^j$  at the point  $p$ . Since  $G(p)$  is the Jacobian matrix of the coordinate representation of the change of chart diffeomorphism map  $\xi \circ \theta^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  (and its inverse  $G^*$  is a coordinate representation of the inverse map  $\theta \circ \xi^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ), the matrices  $G, G^*$  are inverses of each other.

In the case where the manifold  $P$  is our manifold of probability measures on  $\Omega$  with an exponential family form, if we are transforming from the  $(-1)$ -coordinate to the canonical parameters, the Jacobian matrix is the Fisher information matrix.

**Definition 2.6.8.** [13] Let  $P$  be a set of probability measures on some probability space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  with parameters  $\theta = (\theta^1, \dots, \theta^d) \in \Theta \subset \mathbb{R}$ . The score vector is defined as

$$S(\theta; X) := \left( \frac{\partial}{\partial \theta^1} \log(p(X; \theta)), \dots, \frac{\partial}{\partial \theta^d} \log(p(X; \theta)) \right) \quad (2.30)$$

*Remark.* (1) In our definition, we refer to both the random variable version of the score  $S(\theta; X)$ , and a specific realization version  $S(\theta; x)$  as score vectors.

*Remark.* (2) If the set  $P$  has an exponential family form with order  $d$ , the  $i$ -th component of the score vector is the zero mean random variable

$$S^i(\theta; X) = u^i(X) - \mathbb{E}_{\theta}[u^i(X)] \quad (2.31)$$

**Definition 2.6.9.** Let  $p(x; \theta)$  be a parametric probability measure on some probability space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$ . Let  $S(\theta; X)$  be the score vector. The Fisher information matrix is defined as the  $d \times d$  variance-covariance matrix with components

$$g_{ij}(\theta) := \int_{\mathcal{X}} S^i(\theta; X) S^j(\theta; X) dp_{\theta} \quad (2.32)$$

*Remark.* If  $P$  has an exponential family with canonical statistic  $u = (u^1, \dots, u^d)$ , we can write the Fisher information as

$$g_{ij}(\theta) = \int_{\mathcal{X}} (u(x) - \mathbb{E}_{\theta}[u(x)])^i \cdot (u(x) - \mathbb{E}_{\theta}[u(x)])^j dp_{\theta} \quad (2.33)$$

To see that, suppose  $\theta, \xi$  are two different coordinate maps of  $P$  where  $\theta$  is the (+1) coordinate map and  $\xi$  is its Legendre dual (i.e., the (-1)-coordinate map).

$$\begin{aligned} (G(p))_{i,j} &= \frac{\partial \xi^j}{\partial \theta^i} \\ &= \frac{\partial}{\partial \theta^i} \int_{\Omega} u^j dp \\ &= \frac{\partial}{\partial \theta^i} \int_{\Omega} u^j (u^i - \mathbb{E}_{\theta}[u^i]) dp \\ &= \mathbb{E}_{\theta}[(u^j - \mathbb{E}[u^j])(u^i - \mathbb{E}[u^i])] \end{aligned}$$

As a result of this relationship between  $\{e^1, \dots, e^d\}$ , and  $\{(e^*)^1, \dots, (e^*)^d\}$ , the two basis are referred to as dual [3].

For our smooth  $d$  manifold  $P$  of probability measures on  $\Omega$  with an exponential family form, another basis for  $T_p P$  has more statistical significance. Before discussing the nature of that basis, we need to first introduce the concept of a cotangent space and a gradient.

**Definition 2.6.10.** Let  $V$  be a vector space. The dual space  $V^*$  of  $V$  is a set of all linear maps over  $V$  with values in  $\mathbb{R}$ . Elements of  $V^*$  are called co-vectors.

**Definition 2.6.11.** Let  $P$  be a smooth  $d$ -manifold. A cotangent space to  $P$  at  $p$  is defined by

$$(T_p^* P) := (T_p P)^*$$

where  $(T_p P)^*$  is a dual space to  $T_p P$ .

**Definition 2.6.12.** Let  $f$  be a smooth map on  $C^\infty(P)$ . A gradient of  $f$  at the point  $p \in P$  is a map  $(df)_p : T_p P \rightarrow \mathbb{R}$  defined by

$$(df)_p(v) := v f, \quad v \in T_p P$$

*Remark.* The map  $(df)_p$  is a co-vector in the cotangent space  $T_p^* P$ .

*Remark.* From the theory of linear algebra, since  $T_p P$  is finite dimensional, the co-vector space  $T_p^* P$  is also a finite-dimensional vector space which has a dimension equal to that of the dimension of  $T_p P$ . Even more, the linear structure of  $T_p^* P$  is similar to that of  $T_p P$ .

Consider again our smooth  $d$  manifold of probability measures on  $\Omega$  with an exponential family form. An important real-valued function in statistics is the log-likelihood. To keep things concrete, we will otherwise consider the classical definition of the log-likelihood as a real map over  $\Theta$ .

Suppose that we have a set of independent and identically distributed observations  $\mathcal{D} \triangleq \{x_1, \dots, x_n\}, x_i \in \mathcal{X}$ . The log-likelihood map given  $\mathcal{D}$  is a map  $\ell_{\mathcal{D}} : \Theta \rightarrow \mathbb{R}$  defined by

$$\begin{aligned} \ell_{\mathcal{D}}(\theta_p) &= \log \left( \prod_{j=1}^n p_\theta(x_j) \right), \quad p_\theta \in P, \theta \in \Theta, x_i \in \mathcal{D} \\ &= \sum_{j=1}^n \left\{ \sum_{i=1}^d \theta_p^i u^i(x_j) - K(\theta) \right\} \\ &= \sum_{i=1}^d \theta_p^i s^i(\mathcal{D}) - nK(\theta_p), \quad s^i(\mathcal{D}) = \sum_{j=1}^n u^i(x_j), i = 1, \dots, d \end{aligned}$$

Treating  $\Theta$  as a submanifold of  $\mathbb{R}^d$ , clearly  $\ell_{\mathcal{D}} \in C^\infty(\Theta)$ . For any  $\theta_p \in \Theta$ , a gradient of  $\ell_{\mathcal{D}}$  is by definition an element of  $T_{\theta_p}^* \mathbb{R}^d$ . We can compute the components of the gradient  $((d\ell_{\mathcal{D}})_{\theta_p})_i$  by evaluating its action on the chart induced basis of  $T_p P$  (with respect to the coordinate map  $\theta$ ).

$$\begin{aligned} (d\ell_{\mathcal{D}})_{\theta_p} \left( \frac{\partial}{\partial \theta^i} \right)_p &= \left( \frac{\partial \ell}{\partial \theta^i} \right)_{\theta_p} \\ &= \frac{\partial}{\partial \theta^i} \left( \sum_{i=1}^d \theta_p^i \cdot s^i(\mathcal{D}) - nK(\theta_p) \right) \\ &= s^i(\mathcal{D}) - n \frac{\partial K}{\partial \theta_p^i} \\ &= s^i(\mathcal{D}) - \mathbb{E}_{\theta_p} [s^i] \end{aligned}$$

Intuitively speaking, the derivative of the likelihood function in the direction of one of the coordinate axis  $\theta^i$  can be thought of as a zero-mean random variable of the observations  $\mathcal{D}$ . As a result, we can think of the tangent space  $T_pP$  as being a vector space spanned by mean zero random variables on  $\Omega$ . An important basis for that representation of  $T_pP$  is that of the score vectors.

*Remark.* In the above characterization of the random variable representation of  $T_pP$  we followed an intuitive construction rather than a rigorous one. In fact, the characterization of  $T_pP$  as a vector space spanned by mean zero random variables is not precise. This representation is for a tangent space of another manifold  $\tilde{P} \triangleq \{\log(p) : p \in P\}$  using a log transformation of  $P$ ; however, since the log transformation preserves smooth manifold structures, their tangent spaces are isomorphic and have the same linear structure.

This tangent space characterization of  $T_pP$  provides another important geometric notion for the Fisher information matrix. The Fisher information can be thought of as a metric tensor on  $T_pP$ .

**Definition 2.6.13.** [61] Let  $P$  be a smooth  $d$  manifold. A metric tensor at  $p \in P$  is a  $(0, 2)$ -tensor  $g_p : T_pP \times T_pP \rightarrow \mathbb{R}$  such that the following conditions are satisfied:

1.  $g_p$  is a tensor (i.e. is a multi-linear map over the product space of vectors (and/or their co-vectors) with values in  $\mathbb{R}$ ).
2.  $g_p$  is symmetric, i.e. for all  $v, w \in T_pP$ , we have

$$g_p(v, w) = g_p(w, v) \tag{2.34}$$

3.  $g_p$  is non-degenerate, i.e. for all  $v \in T_pP, v \neq 0$ , there exists a  $w \in T_pP$  such that

$$w_p = g_p(v, w) \neq 0$$

It is clear from the definition 2.6.9 of the Fisher information matrix, that it is a coordinate representation (with respect to some basis of  $T_pP$ ) of an abstract Fisher information tensor metric on the tangent space to  $P$  at a point  $p$  (where we think of the tangent vectors as abstract entities). This highlights the fact that the Fisher information as an abstract metric tensor is invariant to the coordinate representation of our manifold  $P$  of probability measures on  $\Omega$  with an exponential family form.

## 2.7 Riemannian manifold structure

This geometric characterization also provides an important tool for applications. Similarly to how one can use the Euclidean inner product (on vectors in a Euclidean space) to decide on the length and angle between two Euclidean vectors, one can use the Fisher information metric tensor to decide the length and angle between two tangent vectors in the tangent space. In other words, the Fisher information metric tensor provides us with a higher-level geometry of the manifold.

**Definition 2.7.1.** Let  $P$  be a smooth  $d$ -manifold of probability measures on probability space  $\Omega$  with an exponential family form. Let  $T_p P$  be a tangent space to  $P$  at  $p$ , and  $g_p$  be the Fisher information metric on  $T_p P$ . Two tangent vectors  $v, w \in T_p P$  are said to be orthogonal if

$$g_p(v, w) = 0 \quad (2.35)$$

Now that we have defined a notion of orthogonality for any two vectors in a single tangent space, it is important to note that, in general, the Fisher information metric tensor of any two basis vectors is strictly neither 0 nor 1 (which is the case for the coordinate axis in a Euclidean space). In other words, the basis of the tangent space is not orthonormal in general. Statistically speaking, this is rather obvious since for the basis vectors of the tangent space to be orthonormal, we require that the covariance of the scores be either 0 or 1 for all points  $p \in P$ , which is not true in general for a family of probability measures in an exponential family form.

However, as discovered by Amari [3], any two dual basis for the tangent space at a point are mutually dual or reciprocal. To see this, let  $e_p^i$  be a basis vector in  $T_p P$ , and  $(e_p^*)^j$  is a vector in the dual basis containing  $e_p^i$ . Observe that

$$\begin{aligned} g_p(e_p^i, (e_p^*)^j) &= g_p(e_p^i, \sum_{k=1}^d (A(p)^*)_{j,k} e_p^k) \\ &= \sum_{k=1}^d (A(p)^*)_{j,k} \cdot g_p(e_p^i, e_p^k) \\ &= \frac{1}{\mathbb{E}_\theta[S^i(\theta; X) \cdot S^k(\theta, X)]} \mathbb{E}_\theta[S^i(\theta; X) \cdot S^k(\theta, X)] \\ &= \delta_k^i \end{aligned} \quad (2.36)$$

This relationship, referred to as complementary orthogonal [3], is rather new to the theory of differential geometry and is characteristic of manifolds in information geometry.

As we have noted above, the Fisher information metric tensor does not behave like a Euclidean metric tensor. This implies that  $P$  is not a Euclidean manifold and/or surface. To make this argument rigorous, we need to study the behavior of the Fisher information metric tensor on the tangent space to all points of the manifold. This requires the notion of a tangent bundle.

**Definition 2.7.2.** [61] Let  $P$  be a smooth  $d$ -manifold. Let  $T_pP$  be a tangent space to  $P$  at  $p \in P$ . The tangent bundle of  $P$  denoted by  $TP$  is defined by a disjoint join

$$TP = \dot{\bigcup}_{p \in P} T_pP \quad (2.37)$$

and we write an element of  $TP$  as the pair  $(p, v)$  where  $p \in P$ , and  $v \in T_pP$ .

Because we are interested in studying the smoothness properties of the metric tensor on the tangent space of a smooth manifold  $P$ , we need to equip the tangent bundle of  $P$  with a smooth manifold structure. A standard construction in differential geometry lifts the smooth structure from the base space  $P$  to the tangent bundle  $TP$  by using a canonical projection map that assigns to every element  $(p, v) \in TP$  a point  $p$  in  $P$ . Going forward, we will follow this standard construction. For more information, refer to [61].

Now that we have a smooth structure on the tangent bundle of a smooth manifold, we can introduce the notion of a Riemannian manifold and the orthogonal relation between the different representations of the tangent bundle.

**Definition 2.7.3.** [61] Let  $P$  be a smooth manifold. A Riemannian metric is a smooth map  $g$  on  $P$  that assigns to each  $p \in P$  a positive definite metric  $(0, 2)$ -tensor  $g_p : T_pP \times T_pP \rightarrow \mathbb{R}$ . The pair  $(P, g)$  is called a Riemannian manifold.

For our smooth  $d$ -manifold  $P$  of probability measures on  $\Omega$  with an exponential family form, it is easy to show that there exists a Riemannian metric that smoothly assigns to each point  $p \in P$  the Fisher information  $(0, 2)$ -tensor (2.6.9) which is by definition positive definite [3]. This metric is called a Fisher metric as one expects.

Suppose that a manifold  $P$  has a coordinate map  $\theta$ , the abstract notion of a tensor can be written as a combination of tensor components with respect to the vector basis on which it acts. In other words, suppose  $\left(\left(\frac{\partial \ell}{\partial \theta^i}\right)_p\right)_{i=1}^d$  is a basis for  $T_pP$ , such that, for all  $v \in T_pP$ , we have

$$v = \sum_{i=1}^d \left(\frac{\partial \ell}{\partial \theta^i}\right)_p \cdot v_\theta^i$$

by the multi-linearity of a tensor, we can write any  $(0, 2)$ -tensor  $T$  as follows

$$g_p(v_1, v_2) = \sum_{i=1}^d \sum_{j=1}^d T^{i,j} g_p \left[ \left( \frac{\partial \ell}{\partial \theta^i} \right)_p, \left( \frac{\partial \ell}{\partial \theta^j} \right)_p \right]$$

This establishes the structure of the tangent bundle on  $P$  which we can immediately use to study the projection of an unknown probability measure  $p \in P$  onto a subfamily of probability measures  $S$  in  $P$ .

As is the case in geometry, two intersecting curves are said to be orthogonal if their tangent vectors at the point of intersection are orthogonal (relative to some inner product on the tangent space at that point). More precisely, let  $\gamma_1, \gamma_2$  be different curves in our manifold  $P$  centered at  $p$  with tangent vectors  $v_{\gamma_1,p}, v_{\gamma_2,p}$  respectively at  $p$ . Then we can say that the two curves are orthogonal at  $p$  if

$$g_p(v_{\gamma_1,p}, v_{\gamma_2,p}) = 0 \tag{2.38}$$

Now, suppose that we are interested in estimating the parameters of an unknown probability measure  $p \in P$  using  $n$  i.i.d. observations  $\mathcal{D}$  drawn from  $p$ . Mathematically speaking, we are interested in finding the value of the parameter  $\hat{\theta}$  such that

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \ell_{\mathcal{D}}(\theta) \tag{2.39}$$

As we have shown in (2.19), the above objective is equivalent to

$$\hat{p} = \operatorname{argmin}_{q \in P} D_{\text{KL}}(p, q) \tag{2.40}$$

where  $\hat{p}$  is the point in  $P$  that correspond to  $\hat{\theta}$ .

Typically, we would like to compute the maximum likelihood estimate over the entire parameter space  $\Theta$  of  $P$ . If the dimension of the parameter space is large, that computation might be intractable. As a result, one might consider a subfamily  $S$  of  $P$  as a good candidate to approximate  $p$ . Using this new family, the above minimization problem (2.40) becomes the following restricted minimization problem over  $S$

$$\hat{p} = \operatorname{argmin}_{q \in S} D_{\text{KL}}(p, q) \tag{2.41}$$

The above minimization objective can be understood as finding the point  $\hat{p}$  in  $S$  such that the line segment connecting  $p$  to  $\hat{p}$  is the "shortest" for all  $q \in S$ . The word shortest is in quotation since we do not yet have a concept of distance on  $P$ .

Intuitively speaking, one can induce a distance on  $P$  using the Fisher metric  $g$  by using the tangent space at a point  $p$  as an approximation for the manifold near  $p$  where distances between points  $q$  near  $p$  are measured using the norm (induced from  $g_p$ ) of tangent vectors with their base  $p$  in the direction of  $q$ . We will take another approach by exploiting the fact that our manifold has a dually "flat" structure.

Recall that our manifold  $P$  has two collections of affine coordinate maps such that going from one collection to the other is through the Legendre transform. If we restrict our coordinate systems over  $P$  to either collection, then  $P$  is an affine manifold.

In an affine manifold, a *geodesic* between any two points has the following coordinate representation

$$\lambda(t) = a \cdot t + b, \quad t \in I \subset \mathbb{R} \tag{2.42}$$

where  $a, b$  are constants. As a result,  $P$  has two dual affine structures. Since all affine structures are flat, we can say  $P$  is dually flat, where duality here is in the convex sense of the Legendre transform.

In manifolds that are not affine, there is no reason to think that tangent vectors along different points in a curve have the same direction. Flat manifolds are rather special in that tangent vectors along geodesics do not change direction. However, if the metric on the manifold (flat or otherwise) is not Euclidean, parallelly transported vectors along any geodesic have different magnitudes (with respect to the norm induced from Riemannian metric).

## 2.8 Projections and the Pythagorean theorem

Dually flat manifolds are also rather special because there are two notions of a geodesic (relative to each affine structure). This permits the following generalized Pythagorean theorem due to Amari [3].

**Definition 2.8.1.** [5, Theorem 1.2, page 24] When a triangle  $pqr$  is orthogonal such that the dual geodesic connection  $p$  and  $q$  is orthogonal to the geodesic connecting  $q$  and  $r$ , the following generalized Pythagorean relation holds

$$D_K(r, p) = D_K(q, p) + D_K(r, q) \tag{2.43}$$

*Remark.* If the affine coordinate system is exactly the same as the convex dual affine coordinate system, a geodesic is a dual geodesic at the same time, and the generalized Pythagorean relation reduces to the Pythagorean relation in a Euclidean space.

This generalized Pythagorean relation provides a geometric interpretation to the minimization objective (2.40). One can show [5] that  $\hat{p}$  is a point in  $S$  such that the tangent vector of the geodesic from  $p$  to  $\hat{p}$  (at the point  $\hat{p}$ ) is orthogonal to any tangent vector to  $S$ .

**Definition 2.8.2.** [5, Definition 1.2, Page 26]  $\hat{p}_S$  is a geodesic projection of  $p$  to  $S$  when the geodesic connecting  $p$  and  $\hat{p}_S$  is orthogonal to  $S$ . Dually,  $\hat{p}_S^*$  is a dual geodesic projection of  $p$  to  $S$ , when the dual geodesic connecting  $p$  and  $\hat{p}_S^*$  is orthogonal to  $S$ .

This can give rise to the following projection theorem.

**Theorem 2.8.1.** [5, Theorem 1.4, page 26] Let  $p \in P$  and a smooth submanifold  $S \subset P$ . The point  $\hat{p}^*$  that minimizes the divergence  $D_K(p : q), q \in S$  is the dual geodesic projection of  $p$  to  $S$ . The point  $\hat{p}_S$  that minimizes the dual divergence  $D_{K^*}(p, q) \in S$  is a geodesic projection of  $p$  to  $S$ .

An important complication is related to the nature of the projection. One can have multiple points that satisfy the projection of  $p$  to  $S$ . However, a necessary conditions can be established if the submanifold  $S$  is also flat.

**Theorem 2.8.2.** [5, Theorem 1.5, Page 27] When  $S$  is a flat submanifold of a dually flat manifold  $P$ , the dual projection of  $p$  to  $S$  is unique and minimizes the divergence. Dually, when  $S$  is a dual flat submanifold of a dually flat manifold  $P$ , the projection of  $p$  to  $S$  is unique and minimizes the dual divergence.

This theorem establishes the importance of choosing a subfamily with an exponential form. This can easily be understood if we imagine a projection in a Euclidean space from a point to a curved surface. It is possible to have 2 points on the curved surface that minimize the divergence. If the surface is flat, the projection will be unique.

So far, we have established the notion of projection of a point onto a submanifold of  $P$ . Note that the above projection relied mainly on the tangent space at the point  $p$  and the restricted tangent bundle of  $P$  to  $S$ . This projection can be extended to a curve  $\gamma : \mathbb{R} \rightarrow P$  by considering a projection operator over the manifold  $P$  such that, at each point, the projection method outlined in the projection theorem (2.8.1). In the next section, we will review the theory of statistical bundles and formally introduce a projection operator over the infinite-dimensional manifold  $\mathcal{P} \supset P$  that can project the conditional density process curve solving some stochastic filtering problem onto a finite-dimensional subfamily.

# Chapter 3

## Infinite Dimensional Information Geometry

### 3.1 Introduction

In the previous chapter, we explored in detail the topological and geometric structure of finite-dimensional exponential families and models embedded in such families. This is the classical structure of Information Geometry as set out by Amari, [5]. There are two ways in which this has been extended in the literature; first by looking at the closure of exponential families, a so-called extended exponential family case, see [33] and second by looking at the infinite-dimensional extensions of exponential families, see [71]. We will consider examples of the extended exponential family case in our numerical examples where the boundary of models – typically simplex-based models rather than manifold-based ones – plays a critical role. In this chapter, we focus on the second of these approaches following the work of Pistone and others. This is necessary due to the fundamental infinite-dimensional nature of the general filtering problem.

Before looking at the details of the infinite-dimensional case we give a high level overview of the major challenges involved. Natural questions which arise starting with the underlying linear structures, for example, are they going to be Hilbert- or Banach-space-based? It turns out that Hilbert structures will be too restrictive and that we work with Banach manifolds. This means that we work with norms rather than inner products and so one of the critical notions of duality – linear duality, which plays an important role in finite-dimensional exponential families – will not have a direct equivalent in the infinite case. The second notation of duality, convex duality, extends more easily but still needs care.

The very rich nature of the set of all distributions means that the cumulant generating function and its derivatives, which basically encodes the Information Geometry in the finite case, does not exist for all possible centered random variables (tangent vectors). The tangent space at a particular distribution will have “well-behaved” directions where the cumulant generating function exists in an open neighborhood but also directions in which this does not happen. Further the space of ‘well-behaved’ directions is not the same for all distributions. Related to this, the Fisher information – in either its covariance or Hessian form – can fail to exist for similar reasons. We therefore have to explore a case in which we have the fundamental Pythagorean (dual) projection result when we are projecting from an infinite-dimensional Banach manifold to a finite-dimensional exponential family.

The following section breaks down these issues in detail showing which structure is preserved in the infinite case and which needs extra regularity results.

## 3.2 Overview

As we have discussed in the previous chapter, a set  $\mathcal{P}$  of all probability measures on some probability space  $\Omega = (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$ , dominated by an arbitrary measure  $\nu_0$  (not necessarily finite) on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  has two different affine structures.

1. A general (+1)-affine space  $(\mathcal{M}, \mathcal{V}, \oplus)$  where
  - (a)  $\mathcal{M}$  is a set of all non-negative measures on  $\Omega$ , dominated by  $\nu_0$  defined up to real scaling factor and sets of  $\nu_0$ -measure zero.
  - (b)  $\mathcal{V}$  is a real vector space of all measurable functions on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ , defined up to real scaling factor.
  - (c) A transitive left group action map  $\oplus$  of the additive group  $(\mathcal{V}, +)$  on  $\mathcal{M}$  defined by

$$\begin{aligned} \oplus : \mathcal{M} \times \mathcal{V} &\rightarrow \mathcal{M}, \\ (\nu, f) &\mapsto \nu \oplus f = e^f \cdot \nu \end{aligned} \tag{3.1}$$

2. A general (−1)-affine space  $(\widetilde{\mathcal{M}}, \mathcal{W}, +)$ 
  - (a)  $\widetilde{\mathcal{M}}$  is the set of all measurable functions on  $\Omega$  dominated by  $\nu_0$  whose total mass is 1.
  - (b)  $\mathcal{W}$  is the vector subspace of  $\mathcal{V}$  of measurable functions whose total mass is zero (with respect to  $\nu_0$ )

(c) A transitive left group action map  $+$  of the additive group  $(\mathscr{W}, +)$  on  $\widetilde{\mathcal{M}}$  defined by

$$\begin{aligned} + : \widetilde{\mathcal{M}} \times \mathscr{W} &\rightarrow \widetilde{\mathcal{M}}, \\ (v, f) &\mapsto (v + f)(x) \triangleq v(x) + f(x), \quad x \in \mathcal{X} \end{aligned} \tag{3.2}$$

The non-parametric geometry of  $\mathcal{P}$  has been studied since the early days of Information Geometry: see [36],[3], and [65]. Initially, it might not be clear why we study this infinite-dimensional geometry — rather than some finite-dimensional subfamily — is of practical importance for this thesis and not merely a mathematical curiosity. As was presented in the introductory chapter, we see from [32] that the solution to stochastic filtering problems is fundamentally infinite-dimensional because the evolution of the conditional density process solving an arbitrary stochastic filtering problem does not “stay” within any finite-dimensional family of probability measures unless very strict conditions are met.

A geometrically grounded method for approximating the evolution of an infinite-dimensional stochastic process is by projecting it on a well-chosen finite-dimensional subfamily of probability measures.

To understand the quality of the projection in the case that the state space is infinite, adequate geometry in the set  $\mathcal{P}$  is required. One of the challenges in constructing an infinite-dimensional geometry on  $\mathcal{P}$  is that, without additional regularity assumptions, it is not clear how one chooses a topology such that finite-dimensional exponential models as subsets of  $\mathcal{P}$  are themselves subsets of open sets in the chosen topology. This is a required topological feature because as we have seen, coordinate maps of the finite-dimensional exponential model are globally defined.

The topology on  $\mathcal{P}$  is not the only challenge. We also need to consider properties of this new geometry. In a perfect world, the new construction should reduce to the finite-dimensional geometry whenever we restrict our attention to any finite-dimensional subspace of  $\mathcal{P}$ .

It should be clear that, if we choose a subset  $P$  of  $\mathcal{P}$  such that  $p \in P$  satisfies additional regularity assumptions such as having compact support, square integrability, or smoothness of densities function, we would have a simpler geometric structure that is easy to understand and use. However, such regularity assumptions might only be viable in particular applications where they are justified. The difficulty arises in the fact that by excluding probability measures without theoretical justifications, we are also excluding practical probability models. For instance, excluding probability densities that are either unbounded or without compact support excludes the important Beta and Normal families.

The following are *desiderata* inspired by the characteristic features of the geometric structures in the finite-dimensional case that we would like the infinite-dimensional geometry of  $\mathcal{P}$  to have. In each case, we discuss the difference in the infinite-dimensional case with the finite case.

1. A collection of bijective maps  $\{s_i : \cdot\}$  from open sets  $\mathcal{U}_i \subset \mathcal{P}$  to an open sets  $\rightarrow B_i \subset \mathcal{V}$  of a some Banach space such that
  - 1.1. *The collection of subsets  $\{\mathcal{U}_i\}$  cover  $\mathcal{P}$ .* It is not clear if such a cover even exists. Even more, Since our manifold is infinite dimensional, the open sets of the linear space might not possess the properties which make the coordinate charts useful in applications.
  - 1.2. *The map  $s_i$  is continuous with respect to some topology on  $\mathcal{P}$  and a canonical Banach space topology on  $\mathcal{V}$ .* In particular, the topology should be weak enough to imply convergence with respect to the Lebesgue topology on  $L^p$  for all  $p > 0$ , otherwise the topology might be too strong for probability convergence theorems.
  - 1.3. *The map  $s_i$  should preserve the important convex structures of  $\mathcal{P}$ .* If that is not the case, many important geometric objects such as the Legendre dual or the cumulant generating function might not exist.
2. A cumulant generating functional map  $K : \mathcal{U}_i \subset \mathcal{V} \rightarrow \mathbb{R}$  such that
  - 2.1. *a generating function  $K$  is defined over a subset of  $\mathcal{V}$  whose image under  $s_i^{-1}$  is convex and open in the topology of  $\mathcal{P}$ .* If this does not hold, there is no guarantee that  $K$  itself is convex, and we might lose the duality between the “generalize” (+1) and (−1) parameters as is the case in finite dimensional.
  - 2.2.  *$K$  has a functional derivative.* Otherwise, the link between the mean and canonical parameters might be lost.
  - 2.3. *The derivative of  $K$  is an element of the dual space of its domain.* If that is not the case, the bidirectional Legendre transform might not hold.
  - 2.4. *The gradient of  $K$  is differentiable in a functional sense.* This is needed to relate tangent vectors in the different affine representation of  $\mathcal{P}$ . In addition, this directly impacts the degree of smoothness of any differentiable structure on  $\mathcal{P}$ .
  - 2.5. *There exists a dual map  $K^*$  on the linear dual of  $\mathcal{V}$  (i.e.  $\mathcal{V}^*$ ) whose gradient is the inverse of the gradient of  $K$ .* This is needed if the Legendre duality is to hold.

- 2.6. *For any probability measure  $p$  in an open set of  $\mathcal{P}$ , the gradient of  $K$  at the vector  $s_i(p)$  assigns the first and second cumulant of the distribution of  $p$ . One can find maps where the first derivative is continuous but not smooth, which will make it impossible to have a Fisher metric on the tangent space.*
3. The tangent space  $T_p\mathcal{P}$  to  $\mathcal{P}$  at any point  $p$  is well defined and:
  - 3.1.  *$T_p\mathcal{P}$  is spanned by centered random variables.* If the tangent space at a point is not spanned by a centered random variable, the score representation of the tangent space in finite-dimensional exponential models will be lost.
  - 3.2.  *$T_p\mathcal{P}$  has two representations that correspond to the two affine representations of  $\mathcal{P}$ , which are complementary orthogonal with respect to the Fisher metric.* Without an adequate orthogonal relationship, optimal projection of an arbitrary point onto a subset might be intractable.
4. There exists a well-defined  $(0,2)$ -tensor on  $T_p\mathcal{P} \times T_p\mathcal{P}$  at every point  $p \in \mathcal{P}$  such that
  - 4.1. *The  $(0,2)$ -tensor is also a metric tensor.* This requires that the tangent space is a Hilbert space which is unlikely. It is also possible that we have a bilinear form, that is degenerate which would prevent the notion of a Fisher information on the score vectors. Since second moments might not exist for all tangent vectors the tensor is only defined on a subspace.
  - 4.2. *The  $(0,2)$ -tensor is positive definite.* One can have a positive semi-definite or even a non-positive definite bilinear form, violating important properties of the variance of the scores.
5. *There exists a smooth map on  $\mathcal{P}$  with values in the tangent bundle  $T\mathcal{P}$  of  $\mathcal{P}$  that assigns to every point  $p \in \mathcal{P}$  a positive definite metric tensor.* Many cases might prevent such a smooth map from existing, such as when the manifold is not smooth enough to allow for a smooth assignment of a bilinear form on the tangent space.
6. *A likelihood map over  $\mathcal{P}$  with values in  $\mathbb{R}$  that is differentiable.* A likelihood map might exist but not be continuous or differentiable making it impossible to use likelihood methods or the strong tools of calculus.

From the list above, one could conclude that generalizing the finite-dimensional exponential family geometry might be rather difficult. The following are the three main sources of difficulties:

1. Extending finite-dimensional vector spaces into infinite-dimensional Banach spaces yields many sources of problems.
  - (a) In finite dimensions, vector spaces are always isomorphic to their duals, which is not the case in infinite dimensions.
  - (b) In finite dimensions, it is always possible to split the vector space into two complementary vector subspaces, which is not generally the case in infinite dimensions (unless the linear space has a Hilbert structure).
  - (c) Linear maps between finite-dimensional vector spaces are always continuous, which is not generally the case in infinite dimensions.
2. The set of probability measures  $\mathcal{P}$  is quite complicated.  $\mathcal{P}$  might contain probability measures whose densities are unbounded, piece wise functions, or nowhere differentiable. As a result, our set  $\mathcal{P}$  might be split into subfamilies with complex boundaries.
3. Statistically important maps (such as a cumulant generating function) over a “generalized” space of parameters might not be well defined or possess the desirable qualities, which is a byproduct of the generalization from finite to infinite dimensions. For instance, the directional derivative operator in infinite dimensions is rather sensitive to the choice of topology which is not the case if the manifold is finite dimensional.

### 3.3 Definitions

In the finite-dimensional case, the geometry was rather “revealed” by connecting ideas from the theory of exponential families, affine geometry, convex analysis and differential geometry. This is in stark contrast to the important technical choices to be made in infinite-dimensional settings.

The first rigorous treatment of a statistical manifold geometry on  $\mathcal{P}$  was due to Pistone and Sempi [71] where they constructed a Banach manifold on  $\mathcal{P}$  with a reasonable tangent bundle structure where the fibers in the bundle are Banach subspaces spanned by exponentially integrable centered random variables. Following this seminal work, many researchers, Gibilisco [43], Pistone and Rogantin [70], Grasselli [44, 45], Cena and Pistone [27], extended the construction to include a dual mean parameterization and presented detailed tangent and cotangent bundle structures equipped with a generalized multilinear  $n$ -form (including a 2-form playing the role of the Fisher metric tensor). It is important

to note that the infinite-dimensional geometry on  $\mathcal{P}$ , known as the theory of statistical bundles, is not considered complete [68] and is rather still an area of research.

Many infinite-dimensional manifolds have been studied in classical differential geometry, including Hilbert manifolds, Banach manifolds, and Fréchet manifolds [61]. The following is a list of some the important differences between finite and infinite dimensional manifolds

1. In finite dimensional manifolds, all coordinate systems are continuous with respect to open sets in the same Euclidean space. In infinite-dimensional theory, open sets need not be subspaces of the same linear space. It is only required that overlapping open sets in the manifold have isomorphic images, which could be in different linear spaces.
2. Since we need to do calculus on the manifold, we require a generalized derivative on function spaces, i.e. Fréchet derivative (a generalization of the derivative on normed spaces) and the Gateaux differential (a generalization of the directional derivative).

We start by formally recalling the notion of a functional derivative from functional analysis, after which we introduce the notion of an atlas which is required to make the infinite-dimensional geometry rigorous.

**Definition 3.3.1.** (Fréchet derivative)[50] Let  $V, W$  be two normed vector spaces. Suppose that  $U \subset V$  is open. A map  $f : U \rightarrow W$  is said to be Fréchet differentiable at  $u \in U$  if there exists a bounded linear operator  $A : V \rightarrow W$  such that

$$\lim_{\|h\|_V \rightarrow 0} \frac{\|f(u+h) - f(u) - Ah\|_W}{\|h\|_V} = 0 \quad (3.3)$$

where the limit is in the sense of a limit of a function on a metric space

**Definition 3.3.2.** (Gateaux derivative)[50] Let  $V, W$  be two locally convex topological vector spaces. Suppose that  $U \subset V$  is open and  $f : U \rightarrow W$ . The Gateaux differential of  $f$  at  $u \in U$  in the direction  $v \in V$  is defined as

$$\lim_{t \rightarrow 0} \frac{f(u+tv) - f(u)}{t} = \left. \frac{d}{dt} f(u+tv) \right|_{t=0} \quad (3.4)$$

where the limit is taken with respect to the topology of  $W$ . If the limit exists for all  $v \in V$ , then one says that  $f$  is Gateaux differentiable at  $u$ .

In the presentation of the finite-dimensional exponential model, we avoided the definition of an atlas because the finite-dimensional manifold of an exponential model has a globally defined coordinate system. This is not the general case in differential geometry, and there is no reason to suspect that the infinite-dimensional space  $\mathcal{P}$  has such a trivial atlas.

**Definition 3.3.3.** [61] Let  $\mathcal{P}$  be a set, an atlas of class  $C^k, k > 0$ , on  $\mathcal{P}$  is a collection of charts  $(\mathcal{U}_i, s_i), i \in I$ , where  $I$  is an arbitrary indexing set such that

1. Each  $\mathcal{U}_i$  is a subset of  $\mathcal{P}$ , and the union  $\bigcup_{i \in I} \mathcal{U}_i$  is the whole set  $\mathcal{P}$ .
2. Each  $s_i$  is a bijection from  $\mathcal{U}_i$  onto an open subset  $B_i = s_i(\mathcal{U}_i)$  of some Banach space
3. For any  $i, j \in I$  if  $\mathcal{U}_i \cap \mathcal{U}_j \neq \emptyset$ , then  $\mathcal{U}_i$  is isomorphic to  $\mathcal{U}_j$ .
4. The transition map  $s_j \circ s_i^{-1} : s_i(\mathcal{U}_i \cap \mathcal{U}_j) \rightarrow s_j(\mathcal{U}_i \cap \mathcal{U}_j)$  is  $r$ -times Fréchet differentiable (i.e. the  $r^{\text{th}}$  Fréchet derivative exists and is a continuous function with respect to  $B_i$ -norm topology on subsets of  $B_i$  and the operator norm topology on space of linear maps between  $B_i^r$  and  $B_j$ ).

### 3.4 Generalized parameter space

The vector space  $\mathcal{V}$  generating the two affine spaces related to  $\mathcal{P}$  is not canonically equipped with a non-trivial norm that can induce a useful statistical manifold structure. Pistone and Sempi [71] addressed this challenge using a subset of  $\mathcal{V}$  modeled based on ideas from Orlicz space theory. This subset is a well behaved set of centered, exponentially integrable random variables.

**Definition 3.4.1.** (Cramér class)[70, Definition 2, Page 724] For each probability measure  $p \in \mathcal{P}$ , the Cramér class at  $p$  is the set  $\mathcal{C}_p$  of all random variables  $u$  on  $\Omega$  such that the moment generating function of  $u$  with respect to  $p$  given by

$$MGF_p(t) = \int_{\mathcal{X}} e^{tu} dp = \mathbb{E}_p[e^{tu}], \quad t \in \mathbb{R} \quad (3.5)$$

is finite in a neighborhood of the origin 0. If moreover the expectation of  $u$  is zero, then we shall call the set  $B_p \subset \mathcal{C}_p$  the centered Cramér class at  $p$ .

Note that, for any  $p \in \mathcal{P}$ , the canonical statistic of any finite-dimensional exponential model containing  $p$  is in the Cramér class  $B_p$ . As we shall see later,  $B_p$  will play a central role in the topology and tangent structure on  $\mathcal{P}$ .

The use of Orlicz spaces in probability theory dates back to 1980 where some bounds were clearer using norms on Orlicz spaces than with tail probabilities [73]. Orlicz spaces were introduced as a generalization of Lebesgue spaces  $L^\alpha$ ,  $\alpha > 0$  where the power function is replaced by a general convex function  $\Phi$  known as a *Young function*.

We will start the discussion of the topology introduced in [71], by first recalling a few important definitions from Orlicz theory.

**Definition 3.4.2.** (Young function)[69] Let  $\phi \in C[0, +\infty[$  satisfies

1.  $\phi(0) = 0$ .
2.  $\phi$  is strictly increasing, and
3.  $\lim_{u \rightarrow +\infty} \phi(u) = +\infty$ ,

then its primitive function

$$\Phi(u) = \int_0^x \phi(u) du, \quad x \geq 0$$

is strictly convex. The function  $\Phi$  is extended to  $\mathbb{R}$  by symmetry  $\Phi(u) = \Phi(|u|)$ , and is called a Young function.

**Definition 3.4.3.** (Orlicz space)[69] Let  $\Phi$  be a Young function, and  $L^{\dagger(\Phi)}(\nu_0)$  be a set of all real measurable functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  on some measurable space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  such that

$$\int_{\mathcal{X}} \Phi(|f|) d\nu_0 < \infty \tag{3.6}$$

for an arbitrary measures  $\nu_0$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ . An Orlicz space  $L^\Phi(\nu_0)$  is the smallest vector space spanned by  $L^{\dagger(\Phi)}(\nu_0)$ , i.e.,

$$L^\Phi(\nu_0) := \left\{ f : \exists c \in \mathbb{R}, c > 0, \int_{\mathcal{X}} \Phi(c \cdot |f|) d\nu_0 < \infty \right\} \tag{3.7}$$

An Orlicz space is typically equipped with a Luxemburg norm derived from the Young function generating it. In most cases, this Young function is chosen such that the generated Orlicz space is also a Banach space (i.e. Cauchy sequences always converge in the space).

For every  $p \in \mathcal{P}$ , Pistone and Sempì [71] modeled a generalized parameter over  $p$  using the Orlicz space  $L^\Phi(p)$  with the following consideration in mind

1. The Orlicz space  $L^\Phi(p)$  should contain the centered Cramér class  $B_p$  as a subset.
2. Induce an Orlicz norm on  $L^\Phi(p)$  such that the centered Cramér class  $B_p$  is a closed Banach subspace. The closed part is a rather important technical point for the convergence of sequences in  $B_p$ . The Banach subspace part is important since we want the vector space acting on the probability measures near  $p$  to be complete.
3. Introduce a generalized cumulant generating functional whose domain is a subset of  $B_p$  and use these subsets to define a continuous mapping from the generalized parameter space to subsets of  $\mathcal{P}$

Formally, Let  $p$  be a probability measure in  $\mathcal{P}$ . Construct an Orlicz space  $L^\Phi(p)$  as follows

- use the young function

$$\cosh(u) - 1, \quad u \in \mathcal{V} \quad (3.8)$$

to generate the Orlicz space

$$L^{\cosh-1}(p) \triangleq \{u \in \mathcal{V} : \exists c > 0, \int_{\mathcal{X}} (\cosh(\frac{1}{c} \cdot u) - 1) dp < \infty\} \quad (3.9)$$

- equip the Orlicz space  $L^{\cosh-1}(p)$  with the norm

$$\|u\|_{\cosh-1,p} \triangleq \inf \left\{ c > 0, \left( \int_{\mathcal{X}} (\cosh(\frac{1}{c} \cdot u) - 1) dp \right) \leq 1 \right\}, \quad u \in L^{\cosh-1}(p) \quad (3.10)$$

- Denote the open unit ball of  $L^{\cosh-1}(p)$  by

$$B_p^{\cosh-1} \triangleq \{u \in L^{\cosh-1}(p) : \|u\|_{\cosh-1,p} < 1\} \quad (3.11)$$

The space  $L^{\cosh-1}(p)$  is not a natural choice from the point of view of functional analysis and manifold theory [27]; however, it is natural for statistics due to the link between  $L^{\cosh-1}(p)$  and the Cramér class of the random variables  $\mathcal{C}_p$  [27], where the latter is rather important in statistical theory and more importantly for exponential models. The flaws of the space is the fact that the bounded random variables are not dense in the space, and  $L^{\cosh-1}(p)$  is not separable unless  $\mathcal{X}$  is finite. However, there exists an equivalent convex function with  $u \mapsto \cosh(u) - 1$  such that the generated Orlicz space is separable [27], as we will see later.

By the restriction to the centered Cramér class of random variables, we will rescue having the set of bounded random variables being dense in the restricted space, although it would restrict the scope of the definition of the manifold to ensure that it contains all possible statistical models that agree with an arbitrary probability measure  $p \in \mathcal{P}$  on a set of measures zero [27]. This restriction would not impact our applications since we are mostly interested in exponential models to approximate solutions of filtering problems. This restriction highlights that the construction is not yet complete as a general manifold of exponential models [27, Page 29].

The Luxemburg norm  $\|u\|_{\cosh-1,p}$  is rather special because it induces a topology on  $L^{\cosh-1}(p)$  such that the convex balanced set  $\{u : \mathbb{E}_p[\cosh(u) - 1] \leq 1\}$  is an open ball in the topology [71]. In addition, one can show [71, Page 1549] that  $\|u\|_{\cosh-1,p}$  is unique and  $(L^{\cosh-1}(p), \|u\|_{\cosh-1,p})$  is a Banach space with respect to  $\|u\|_{\cosh-1,p}$ .

Even more, the Cramér class  $\mathcal{C}_p$  at  $p$  is a Banach space with the Orlicz norm  $\|u\|_{\cosh-1,p}$ , and the centered Cramér class  $B_p$  is a closed subspace of  $(\mathcal{C}_p, \|u\|_{\cosh-1,p})$  [70, Proposition 3, page 724]. As a result,  $B_p$  is a closed Banach subspace of the normed Orlicz space  $L^{\cosh-1}(p)$  [71, Proposition 2.3].

This topology is rather interesting because, for any  $\alpha > 0$ ,  $\|\cdot\|_{\cosh-1,p}$  reduces to the standard Lebesgue  $L^\alpha$  norm whenever we restrict our attention to subsets of  $\mathcal{P}$  that are  $\alpha$ -integrable in the standard Lebesgue  $L^\alpha(p)$  sense [71, Page 1553].

Denote the intersection of  $B_p$  and the open unit Orlicz ball  $B_p^{\cosh-1}$  by  $O_p$ , i.e.

$$O_p := \{u \in B_p : \|u\|_{\cosh-1,p} < 1\} \quad (3.12)$$

To identify good candidates of open sets to use as generalized parameters for subsets of  $\mathcal{P}$ , we first need to identify which subsets of  $B_p$  admit a map that generalizes the cumulant generating function. As it turns out,  $O_p$  are exactly these sets. Following [70], we will define the cumulant generating functional by the log transformation of a generalized moment generating functional, and show that the generalized cumulant generating functional does indeed have some of the characteristic features of cumulant generating functions in finite-dimensional exponential models.

**Definition 3.4.4.** [71, Definition 2.6] Let  $\Omega = (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$  be some probability space with a probability measure  $p$  on  $\Omega$ . Suppose  $L^{\cosh-1}(p)$  is the Orlicz space (3.9). Define

1. The moment generating functional  $M_p : L^{\cosh-1}(p) \rightarrow \bar{\mathbb{R}}_+$

$$M_p(u) := \mathbb{E}_p[e^u], \quad u \in L^{\cosh-1}(p) \quad (3.13)$$

and denote its domain by

$$\mathcal{D}(M_p) = \{u \in L^{\cosh^{-1}}(p) : M_p(u) < \infty\} \quad (3.14)$$

2. The cumulant generating functional  $K_p : L^{\cosh^{-1}}(p) \rightarrow \mathbb{R}$

$$K_p(u) := \log(M_p(u)), \quad u \in L^{\cosh^{-1}}(p) \quad (3.15)$$

3. Denote the proper domain of  $K_p$  by

$$\mathcal{D}(K_p) := \{u \in L^{\cosh^{-1}}(p) : K_p(u) < \infty\} \quad (3.16)$$

4. The symmetrized cumulant generating functional  $\tilde{K}_p : \mathcal{D}(K_p) \rightarrow \mathbb{R}$

$$\begin{aligned} \tilde{K}_p &:= \frac{1}{2} \left[ K_p(u) + K_p(-u) \right] \\ &= \log \left[ K_p(u) \cdot K_p(-u) \right]^{1/2} \end{aligned} \quad (3.17)$$

*Remark.* In Definition 3.13, we are thinking of  $u \in L^{\cosh^{-1}}(p)$  as directions, hence the Definition 3.4.4 does not exactly match Definition 3.13.

The following proposition, shows that the moment generating functional defined above 3.13, indeed generalizes the moment generating function in the sense that it satisfies the characteristic features of moment generating functions (see [81]), i.e.

1. Every moment generating function is convex, and lower semi-continuous
2. Every moment generating function is analytic in the interior of its proper domain.
3. Every moment generating function has a derivative which is obtained by differentiating under the integral sign.

**Proposition 3.4.1.** [27, Theorem 10, Page 35] The moment generating functional (3.13) satisfy the following properties:

1.  $M_p(0) = 1$ ; otherwise, for each centered random variable  $u \neq 0$ ,  $M_p(u) > 1$ .
2.  $M_p$  is convex and lower semi-continuous, and its proper domain is a convex set which contains  $B_p^{\cosh^{-1}}$ ; in particular the interior of such a domain is a non empty convex set

3.  $M_p$  is infinitely Gâteaux-differentiable in the interior of its proper domain; the  $n^{\text{th}}$  derivative in the direction of  $v$  is given by

$$v \mapsto \mathbb{E}_p[v^n \exp(u)], \quad u \in B_p^{\cosh^{-1}}, v \in O_p \quad (3.18)$$

where  $\mathbb{E}_p$  is the expectation with respect to  $p$ .

4.  $M_p$  is bounded and infinitely Fréchet-differentiable and analytic on  $B_p^{\cosh^{-1}}$ .

Observe the rather subtle point that, the proper domain of the cumulant generating functional  $K_p$  satisfy

$$O_p = B_p \cap B_p^{\cosh^{-1}} \subset \mathcal{D}(K_p) = \mathcal{D}(M_p) \cap B_p \quad (3.19)$$

The following proposition establishes that, if we restrict our attention to the centered Cramér class at  $p$  (to ensure the positivity of  $K_p$ ), and use vectors in  $B_p$  as directions of differentiation, the cumulant generating functional  $K_p$  would be a reasonable generalization of the finite-dimensional cumulant generating function. More importantly, if we exponentially tilt the vector  $u \in O_p$  (with the base of tilt at  $p$ ), we end up with a probability measure in  $\mathcal{P}$ , hence we could induce a topology on  $\mathcal{P}$  from that of  $L^{\cosh^{-1}}(p)$ ,  $K_p$ , and the action of the exponential tilt map.

**Proposition 3.4.2.** [27, Theorem 12, Page 37] For  $u \in B_p$ , the cumulant generating functional  $K_p$  defined in (3.15) has the following properties

1.  $K_p$  is null at 0, elsewhere it is strictly positive, convex, lower semicontinuous, and its proper domain is a convex set which contains  $O_p$ ; in particular, the interior of such a domain is a nonempty convex set.
2.  $K_p$  is infinitely Gâteaux-differentiable in the interior of its proper domain
3.  $K_p$  is bounded, infinitely Fréchet-differentiable and analytic on  $O_p$ .
4. For all  $u \in O_p$ ,  $q = \exp(u - K_p(u))$  is a probability measure in  $\mathcal{P}$  and the value of the  $n$ th-differential of  $K_p$  at  $u$  in the direction of  $v \in B_p$  is the  $n$ th cumulant of  $v$  under the probability measure  $p_u = e^{u - K_p(u)} p$ , that is

$$d_u^n K_p(v^1, \dots, v^n) = \left. \frac{d^n}{dt^n} \log \mathbb{E}_{p_u} [\exp(tv)] \right|_{t=0} \quad (3.20)$$

Following a similar construction to finite-dimensional exponential models, we define a maximal exponential model using the cumulant generating functional over its proper domain. As we will see later, this is a rather characteristic feature of this topology on  $\mathcal{P}$

**Definition 3.4.5.** (Maximal exponential model)[27, Definition 20] Let  $p$  be a point in  $\mathcal{P}$ . Denote a topological interior of the proper domain of the cumulant generating functional  $K_p$  (3.15) by  $\mathcal{D}(K)_p$ . The maximum exponential model at  $p$  is

$$\mathcal{E}(p) \triangleq \{\exp(u - K_p(u)) \cdot p : u \in \mathcal{D}(K)_p\} \quad (3.21)$$

In the proposition (3.4.2) above, the last two properties of  $K_p$  are rather quite important. They establish that  $K_p$  is differentiable over its proper domain, and has a smooth differential in any direction  $v$  in the centered Cramér class  $B_p$ . This is a required property if we are to establish a notion of Legendre duality later on.

We immediately use that information and define the exponential tilt map

$$\begin{aligned} e_p : O_p &\rightarrow \mathcal{U}_p \subset \mathcal{P} \\ u &\mapsto e_p(u) = e^{u - K_p(u)} p \end{aligned} \quad (3.22)$$

where  $\mathcal{U}_p$  is the image of  $O_p$  under  $e_p$ . Note that, all probability measures  $q \in \mathcal{U}_p$  have the form  $q = e^{u - K_p(u)} p$  for some  $u \in O_p$

To use this map to induce a topology on  $\mathcal{P}$ ,  $e_p$  must be one-to-one. In fact, this is the case [71, Page 1553]. Observe that, for any  $u, v \in O_p$ , if  $e_p(u) = e_p(v)$ , we have that  $u - v$  equals a constant. Since the only constant in  $O_p$  is a zero function (since by definition it contains a subset of the centered Cramér class at  $p$ ), it follows that  $u = v$ , and  $e_p$  is one-to-one.

In this new topology, the open sets are images of  $O_p$  under the mapping  $e_p$ . This indeed looks similar to the finite-dimensional case discussed in (2.3), where one could induce a topology on a finite-dimensional exponential model  $P \subset \mathcal{P}$  using the vector space acting on elements of  $P$  using an exponential map with the exception that open sets in the topology on  $\mathcal{P}$  do not cover it fully, which was the case in finite dimension.

Following the publication of Pistone and Sempi [71], other topologies on  $\mathcal{P}$  were proposed, such as a topology associated with an infinite-dimensional Reproducing Kernel Hilbert Manifold [40], and a topology modeled using Sobolv spaces [66]. However, both topologies are stronger than the topology induced from  $\|\cdot\|_{\cosh-1,p}$ . In addition, as we will see later, the tangent bundle structure in the case of this Banach topology is more similar

(but not exact) to the geometry of the finite-dimensional exponential models rather than that of Fukumizu and Newton.

The above topological construction on  $\mathcal{P}$  achieves the topologically relevant desiderata outlined earlier in this chapter; however, a few important key questions have not yet been rigorously studied.

1. *What is the nature of the boundary of  $\mathcal{P}$  with respect to its two affine structures?* As we outlined in the previous chapter,  $\mathcal{P}$  has two different boundary behaviors in different affine structures (i.e. total positivity in the  $(-1)$ -affine space and unit mass in the  $(+1)$ -affine space).
2. *What is the nature of the boundary of open sets that represent finite-dimensional exponential models?* It is not clear how the open Banach balls map to the finite dimensional  $(+1)$ -parameter spaces  $\mathcal{P}$  especially in the case where the finite-dimensional exponential model is not regular.

This concludes our discussion of the topology on  $\mathcal{P}$  which we denote by  $\mathcal{O}(\mathcal{P})$ , and a candidate generalized parameter space, which we refer to as a  $(+1)$ -generalized parameter space. We now shift our discussion to the manifold structure on  $\mathcal{P}$ .

### 3.5 Manifold structure

In the same monograph [71] where Pistone and Semi introduced the topology  $\mathcal{O}(\mathcal{P})$ , they also introduced a smooth atlas on the topological space  $(\mathcal{P}, \mathcal{O}(\mathcal{P}))$ . The theory of infinite-dimensional statistical manifolds has been rigorously developed [70, 44, 27, 44] since then with an introduction of a convex dual affine parameterization between the generalized  $(+1)$  parameterization and a new notion of mean parameters, which happens to reduce to the right notion in finite dimensions.

Since the topology  $\mathcal{O}(\mathcal{P})$  was induced using the one-to-one map  $e_p$ , for all  $p \in \mathcal{P}$  (3.22), one can immediately proceed by defining coordinate charts on a neighborhood of an arbitrary point  $p_0 \in \mathcal{P}$  using the inverse map  $e_{p_0}^{-1}$ . We can speak of the inverse map  $e_{p_0}^{-1}$  because by the restriction of open sets in  $\mathcal{O}(\mathcal{P})$  to the image of  $O_{p_0}$  under  $e_{p_0}$ ,  $e_{p_0}$  becomes a bijection.

Formally, Let  $p$  be any point in  $\mathcal{P}$ , and  $\mathcal{U}_{p_0} \in \mathcal{O}(\mathcal{P})$  be a neighborhood of  $p$ . Define a coordinate map  $s_{p_0}$  on  $\mathcal{U}_{p_0}$  by the inverse of  $e_{p_0}$ , i.e.

$$\begin{aligned} s_{p_0} : \mathcal{U}_{p_0} &\rightarrow O_{p_0} \subset B_{p_0}, \\ p &\mapsto s_{p_0}(p) = e_{p_0}^{-1}(p) = \log\left(\frac{p}{p_0}\right) - \mathbb{E}_{p_0}\left[\log\left(\frac{p}{p_0}\right)\right] \end{aligned} \quad (3.23)$$

An interesting point regarding the map  $s_{p_0}$  is its relation to the Kullback-Leibler divergence. To see that, note that the value of map at  $p_0$  is

$$\begin{aligned} s_{p_0}(p_0) &= \log\left(\frac{p_0}{p_0}\right) - \mathbb{E}_{p_0}\left[\log\left(\frac{p_0}{p_0}\right)\right] \\ &= 0 \end{aligned}$$

As we move away from  $p_0$ , say to  $p \in O_{p_0}$ , we have

$$\begin{aligned} s_{p_0}(p) &= \log\left(\frac{p}{p_0}\right) - \mathbb{E}_{p_0}\left[\log\left(\frac{p}{p_0}\right)\right] \\ &= \log\left(\frac{p}{p_0}\right) + \mathbb{E}_{p_0}\left[\log\left(\frac{p_0}{p}\right)\right] \\ &= \log\left(\frac{p}{p_0}\right) + D_{KL}(p_0, p) \end{aligned}$$

where  $D_{KL}(p_0, p)$  is the Kullback-Leibler divergence from  $p$  to  $p_0$ , and  $\log\left(\frac{p}{p_0}\right)$  is the log ratio of the two densities.

To continue the construction of the manifold, we need to show that different open neighborhoods of  $p$  are isomorphic. It suffices to show, for any two probability measures  $p, q \in \mathcal{P}$  connected by a 1-dimensional exponential model we have an isomorphism between the Orlicz spaces  $L^{\cosh^{-1}}(q), L^{\cosh^{-1}}(p)$ . This is a sufficient condition because being connected by a 1-dimensional exponential model is an equivalence relation. In other words, let  $p_0, p_1, p_2$  be probability measures in  $\mathcal{P}$ , if one can show that  $p_0 \sim p_1$  (where  $\sim$  is the equivalence relation that they are connected by a 1-dimensional exponential model) implies that  $L^{\cosh^{-1}}(p_0) \cong L^{\cosh^{-1}}(p_1)$  then, if  $p_1 \sim p_2$ , it follows that  $p_0 \sim p_2$ , and  $L^{\cosh^{-1}}(p_0) \cong L^{\cosh^{-1}}(p_2)$ .

This relationship is a defining feature of the topology  $\mathcal{O}(\mathcal{P})$  and establishes the fact that a maximal exponential model at  $p$  is indeed maximal. In other words, this construction

ensures that any finite-dimensional exponential family containing an arbitrary point  $p \in \mathcal{P}$  is contained in a "maximal" exponential family [71]. In fact, as can be shown [71], a finite-dimensional exponential model containing  $p$  is fully covered by some open neighborhood of  $p$  in the topology [71]. This ensures that coordinate systems on finite-dimensional exponential models are globally defined which was a desired feature of the new geometry. For technical details, see [70, Proposition 5, Page 727].

The following theorem establishes this isomorphism notion and connects it with a few important key facts about the topology.

**Theorem 3.5.1.** [27, Theorem 19 and 21] The following statements are equivalent

1.  $p, q \in \mathcal{P}$  are connected by an exponential arc.
2.  $q \in \mathcal{E}(p)$ .
3.  $\mathcal{E}(q) \cong \mathcal{E}(p)$ .
4.  $\log\left(\frac{q}{p}\right)$  belongs to both  $L^{\cosh^{-1}}(q)$  and  $L^{\cosh^{-1}}(p)$ .
5.  $L^{\cosh^{-1}}(q)$  and  $L^{\cosh^{-1}}(p)$  are equal as vector spaces, and their norms are equivalent.

The above Theorem 3.5.1 establishes the missing condition to rigorously show that  $(\mathcal{P}, \mathcal{O}(\mathcal{P}))$  is a Banach manifold (i.e. it satisfies Conditions 1,2, and 3 in Definition 3.3.3 of an atlas.

Now, an important question should come to mind, what kind of an atlas do we have? is it smooth? is it affine? As it turns out, the answer to all these questions is yes [71].

**Theorem 3.5.2.** [71, Theorem 3.6, Page 1556] Let  $(\mathcal{P}, \mathcal{O}(\mathcal{P}))$  be the topological space constructed above. Suppose that  $(\mathcal{U}_\alpha, s_\alpha)$ , and  $(\mathcal{U}_\beta, s_\beta)$  are two coordinate charts in  $\mathcal{A}(\mathcal{P})$  such that  $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \emptyset$ . The coordinate transition map is a composition map

$$s_\beta \circ s_\alpha^{-1} : s_\alpha(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \rightarrow s_\beta(\mathcal{U}_\alpha \cap \mathcal{U}_\beta)$$

which simplifies to

$$(s_\beta \circ s_\alpha^{-1})(u) = u + \log \frac{\alpha}{\beta} - \mathbb{E}_\beta \left[ u + \log \frac{\alpha}{\beta} \right] \quad (3.24)$$

Define the atlas

$$\mathcal{A}(\mathcal{P}) \triangleq \{(O_p, s_p) : p \in \mathcal{P}\} \quad (3.25)$$

Then

- $\mathcal{A}(\mathcal{P})$  is a  $C^\infty$ -isomorphism.
- $\mathcal{A}(\mathcal{P})$  is an affine atlas (i.e., all transition maps in the atlas are affine maps).

One can consider a coarser topology on  $\mathcal{P}$  which would produce another atlas that is equivalent to 3.25 [27]; however, this later atlas will emphasize the impact of having exponential arcs and will be important later for developing the dual coordinate system.

**Definition 3.5.1.** [27, Definition 23, Page 43] Let  $(\mathcal{P}, \mathcal{O}(\mathcal{P}))$  be the topological space constructed above. Define the atlas

$$\mathcal{A}_{\mathcal{E}}(\mathcal{P}) \triangleq \{(\mathcal{E}(p), s_p) : p \in \mathcal{P}\} \quad (3.26)$$

**Proposition 3.5.3.** [27, Theorem 25, Page 44] The atlas  $\mathcal{A}(\mathcal{P})$  defined in (3.25), and the atlas  $\mathcal{A}_{\mathcal{E}}(\mathcal{P})$  defined in (3.26) are equivalent.

Observe that the modified atlas  $\mathcal{A}_{\mathcal{E}}(\mathcal{P})$  emphasizes the fact that  $\mathcal{E}(p)$  are the connected components of the topology. In addition, we can think of the coordinate maps in this new atlas as being globally defined on  $\mathcal{E}(p)$  similar to how coordinate systems were globally defined on a finite-dimensional exponential model.

So far, our discussion has been focused on the exponential geometry  $\mathcal{P}$ , what about the (dual) geometry related to the generalized  $(-1)$  affine structure? In [70] another parameterization of  $\mathcal{P}$  is provided; however, as we will see, the parameterization is not a coordinate system [27]. In fact, introducing a dual coordinate system for the  $(-1)$  affine structure of  $\mathcal{P}$  requires the construction of another manifold, a mixture manifold, that is related to the exponential manifold geometry of  $\mathcal{P}$  that we constructed above.

Recall that, in the finite dimensional case, the dual affine representation of  $\mathcal{P}$  was connected through a few important links:

- The transition maps in the atlas in the finite dimensional case, were non-linear functions and are exactly the Legendre transform of the cumulant generating function, and its convex dual, the relative entropy.
- In the finite-dimensional case, the  $(-1)$  parameter space was spanned by the mean parameters.

Consider the following convex functions

- $\Phi_1 : u \mapsto \exp(|u|) - |u| - 1$

- $\Phi_2 : u \mapsto (1 + |u|)\log(1 + |u|) - |u|$

One can show [70], that the function  $u \mapsto \cosh(|u|) - 1$  is equivalent to  $\Phi_1$  in the sense that the norms that they induce on  $B_p$  are equivalent. In addition, because all these functions (along with  $u \mapsto \cosh(|u|) - 1$ ) are strictly convex and differentiable, we can talk about the inverse of their derivatives (i.e.  $[\Phi']^{-1}$ ), and define a convex conjugate  $\Phi^*$  of  $\Phi$  such that

$$[\Phi^*]' \circ [\Phi'] : u \mapsto u \quad (3.27)$$

In Banach theory, an important consequence of this convex duality is the Banach duality. More precisely, if  $\Phi^*$  is a convex conjugate to  $\Phi$ , for any  $p \in B_p$ , the following bi-linear form

$$L^\Phi(p) \times L^{\Phi^*}(p) \ni (u, {}^*u) \mapsto \int_{\mathcal{X}} (u \cdot {}^*u) dp \in \mathbb{R} \quad (3.28)$$

is continuous [70]. This is a rather important result for our manifold  $\mathcal{P}$ . If we knew the convex conjugate for  $u \mapsto \cosh(u) - 1$ , we might be able to establish a Banach duality between  $B_p$  and  $L^{(\cosh-1)^*}(p)$  (or one of its subsets).

As it turns out, [70],  $\Phi_1$  and  $\Phi_2$  are convex conjugate. Combining this result with the fact that  $u \mapsto \cosh(u) - 1$  is equivalent to  $\Phi_1$ , we should consider the Orlicz space generated by the Young function  $\Phi_2$ .

Even with this Banach duality, we are not done. In general, Banach duality does not imply equality between the spaces involved (which is automatic if the vector space is finite-dimensional). In other words, even though  $L^{\cosh-1}(p)$  and  $L^{\Phi_2}(p)$  are Banach dual, they are not equal.

This fact complicates the identification of the Banach dual  $(B_p)^*$  in  $L^{\Phi_2}(p)$  to  $B_p$ . We now need to find a way to map  $B_p$  onto  $L^{\Phi_2}(p)$ . A reasonable candidate is a so-called  $x \log x$  class.

**Definition 3.5.2.** ( $x \log x$  class)[70, Definition 6, Page 728] Let  $p \in \mathcal{P}$ . Denote the Banach space of centered random variables in  $L^{\Phi_2}(p)$  by  ${}^*B_p$ , a so-called  $x \log x$  class.

The following proposition establishes a link between the space  ${}^*B_p$  and  $p$ -integrability, which is important for the notion of mean parameters.

**Proposition 3.5.4.** [70, Proposition 6] Let  $p \in \mathcal{P}$ . A random variable  $u$  on  $\Omega$  belongs to  ${}^*B_p$  if and only if it is centered and  $\Phi_2(u)$  is  $p$ -integrable.

By the definition of  ${}^*B_p$ , it is identified with a subset of  $(B_p)^* \subseteq L^{\Phi_2}(p)$ . In fact [70],  ${}^*B_p$  is a proper subset of  $(B_p)^*$ . In addition, one can show that  $B_p$  is isomorphic to  $({}^*B_p)^*$  [70, Proposition 8, Page 729]. As a result  ${}^*B_p$  is the "dual" generalized coordinate system of  $\mathcal{P}$ .

It is important to note that even though the space  ${}^*B_p$  was induced from the convex dual to  $\cosh(|u|) - 1$ , it was not induced directly from the cumulant generating functional  $K_p$ . However, as we will see below, the differential of  $K_p$  at  $u \in O_p$  in any direction in  $B_p$  maps into  ${}^*B_p$ , and the mapping is one-to-one.

Recall that for any coordinate chart  $(\mathcal{U}_p, s_p)$  in the atlas, the cumulant generating functional is infinitely differentiable at any parameter  $u \in O_p$  in any direction  $v \in B_p$ . In particular, the first order differential of  $K_p$  at  $u \in O_p$  in any direction  $v \in B_p$  is

$$\begin{aligned}
(d_u^1 K_p)(v) &= \left. \frac{d}{dt} K_p(u + tv) \right|_{t=0} \\
&= \left. \frac{d}{dt} \log(\mathbb{E}_p(e^{u+tv})) \right|_{t=0} \\
&= \left. \frac{1}{M_p(u + tv)} \int_{\mathcal{X}} v e^{u+tv} dp \right|_{t=0} \\
&= \frac{1}{M_p(u)} \int_{\mathcal{X}} v e^u dp \\
&= \frac{1}{e^{\log(\mathbb{E}_p(e^u))}} \int_{\mathcal{X}} v e^u dp \\
&= \int_{\mathcal{X}} v e^{u - K_p(u)} dp \\
&= \mathbb{E}_{p_u}[v]
\end{aligned}$$

where  $p_u$  is the point in the manifold at the parameter  $u$  under the coordinate map  $s_p$  (i.e.  $p_u = e^{u - K_p(u)}$ ).

To see how is this related to the Banach dual  ${}^*B_p$ , observe that

$$(d_u^1 K_p)(v) = \mathbb{E}_{p_u}[v] = \mathbb{E}_p \left[ \left( \frac{p_u}{p} - 1 \right) v \right]$$

where  $\frac{p_u}{p} - 1$  is a random variable which we can think of as an element of  ${}^*B_p$  and denote it by  ${}^*u$ . In other words

$$\nabla K_p(u) = e^{u - K_p(u)} - 1 = \frac{p_u}{p} - 1 = {}^*u \in {}^*B_p \quad (3.29)$$

In a more compact notation, we will write

$$(d_u^1 K_p)(v) = \mathbb{E}_p[*uv] \quad (3.30)$$

The following proposition establishes important properties regarding the mapping  $B_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$ .

**Proposition 3.5.5.** [27] Let  $p \in \mathcal{P}$ , and  $K_p$  be the cumulant generating functional at  $p$ . Then

1. the mapping  $B_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$  is monotonic, and in particular one-to-one.
2. the weak derivative of the map  $B_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$  at  $u$  applied to  $w \in B_p$  is given by

$$D(\nabla K_p(u))w = \frac{p_u}{p}(w - \mathbb{E}_{p_u}[w])$$

and it is one-to-one at each point.

By differentiating  $K_p$ , we can map from the generalized (+1) coordinate system to its Banach dual. However, it is not clear whether  $\nabla K_p$  could provide an alternative coordinate system to  $\mathcal{P}$ . Before we can present that, let us first see what happens when we consider a finite-dimensional subfamily model  $P \subset \mathcal{E}(p)$  (where  $\subset$  does not mean a submanifold).

Suppose that  $P$  is a finite-dimensional exponential model containing  $p \in \mathcal{P}$  where  $(s^1, \dots, s^d)$  is a sufficient statistic of  $P$ . By definition,  $s^i \in B_p$ . Define

$$\Theta \triangleq \left\{ \theta = (\theta^1, \dots, \theta^d) \in \mathbb{R}^d : \sum_{i=1}^d \theta^i s^i \in \mathcal{D}(K_p) \right\} \quad (3.31)$$

Let  $\psi(\theta) \triangleq K_p(\sum_{i=1}^d \theta^i s^i)$ . Observe that all probability measures in  $P$  have the following exponential form

$$p_\theta = \exp\left(\sum_{i=1}^d \theta^i s^i - \psi(\theta)\right)p, \quad \theta \in \Theta$$

In other words, any finite-dimensional exponential model  $P$  is a subset of the maximal exponential model at  $p$  [70], i.e.  $P \subset \mathcal{E}(p)$ .

Recall that the linear space spanned by the canonical statistic  $s = (s^1, \dots, s^d)$  uniquely characterizes finite-dimensional exponential models. Observe that

$$\frac{\partial}{\partial \theta^j} \psi(\theta) = \mathbb{E}_{p_\theta}[u^j] = (d_{u=\sum_{i=1}^d (\theta^i s^i)} K_p)(u^j) = \mathbb{E}_p \left[ \left( \frac{p_\theta}{p} - 1 \right) u^j \right] = \mathbb{E}_p[{}^*u_\theta u^j]$$

In other words, the duality between  ${}^*B_p$  and  $B_p$  reduces to the Legendre transform in the case of  $P$ . In addition, this duality leads to the mean parameterization of  $P$  as desired.

Now, let us look closely at the linear mapping

$$\begin{aligned} (\nabla K_{p_0} \circ s_{p_0}) : \mathcal{P} \ni \mathcal{U}_{p_0} &\rightarrow {}^*B_{p_0} \\ p &\mapsto \frac{p}{p_0} - 1 \end{aligned} \tag{3.32}$$

This map cannot be a coordinate chart on  $\mathcal{P}$  because it is bounded below by  $-1$  [27]. As a result, we cannot directly use the cumulant-generating functional to move from the coordinate system associated with the exponential geometry of  $\mathcal{P}$  to the coordinate system associated with the mixture geometry of  $\mathcal{P}$  (i.e., the coordinate system associated with the  $(-1)$  affine structure of  $\mathcal{P}$ ). However, as we will see later, we can still define a general notion of convex duality between these two distinct manifold structures of  $\mathcal{P}$ .

Recall that, in the finite-dimensional exponential model, the convex dual of the cumulant generating function is the reverse Kullback-Leibler divergence. If we are to have some notion of a duality that reduces to the right convex duality in the finite-dimensional case, then one should establish a reverse Kullback-Leibler between points in  $\mathcal{P}$  and another dual manifold that is somewhat related to  $\mathcal{P}$ . An intuitive choice is the set of all non-negative measurable functions with unit mass in the generalized  $(-1)$  affine structure of  $\mathcal{P}$ , that is,  $(\widetilde{M}, \mathcal{W}, +)$

Recall that  $\widetilde{M}$  is the set of all measurable functions on  $\Omega$  with a total mass of one dominated by an arbitrary measure  $\nu$  on  $\Omega$ . Denote the set of all non-negative measurable functions on  $\Omega$  dominated by  $\nu$  by  $\mathcal{P}_\geq$ . Note that  $\mathcal{P} \subseteq \mathcal{P}_\geq \subseteq \widetilde{M}$  and, for any  $\mathfrak{q} \in \widetilde{M}$ , we can normalize it by  $\tilde{\mathfrak{q}} = \frac{|\mathfrak{q}|}{\int |\mathfrak{q}| d\nu_0}$ . Let  $p \in \mathcal{P}$ , and consider the set

$$\{\mathfrak{q} \in \widetilde{M} : D_{KL}(\tilde{\mathfrak{q}}, p) < \infty\} \subset \widetilde{M} \tag{3.33}$$

As it turns out, this set is related to a Banach subset of the Orlicz space  $L^\Psi(p)$  (which is the convex dual of  $L^{\cosh^{-1}}(p)$ ).

**Proposition 3.5.6.** [27, Proposition 31 and 32, Page 48] Let  $p \in \mathcal{P}$  be given. For each  $\mathfrak{q} \in \widetilde{M}$  denote the probability density associated with  $\mathfrak{q}$  by  $\tilde{\mathfrak{q}} = \frac{|\mathfrak{q}|}{\int |\mathfrak{q}| d\nu}$ .

1. The Kullback-Leibler divergence  $D_{KL}(\tilde{\mathfrak{q}}, p)$  is finite if and only if  $\mathfrak{q} \in {}^*\mathcal{U}_p$ :

$$D_{KL}(\tilde{\mathfrak{q}}, p) = \mathbb{E}_p \left[ \frac{\tilde{\mathfrak{q}}}{p} \log \left( \frac{\tilde{\mathfrak{q}}}{p} \right) \right] < \infty$$

where

$${}^*\mathcal{U}_p \triangleq \left\{ \mathfrak{q} \in \widetilde{M} : \frac{\mathfrak{q}}{p} \in L^\Psi(p) \right\} \quad (3.34)$$

2. In addition, letting  $\mathcal{U}_p$  be the set defined in (3.22), we have  $\mathcal{U}_p \subset {}^*\mathcal{U}_p$ .

An important result regarding the collection  $\{{}^*\mathcal{U}_q : q \in \mathcal{U}_p\}$ , they are all isomorphic [27, Proposition 34]. In addition,  $\mathcal{U}_p \subset {}^*\mathcal{U}_p$  [27, Proposition 32].

This suggests that we use as a dual manifold to  $\mathcal{P}$  in the following union

$$\bigcup_{p \in \mathcal{P}} {}^*\mathcal{U}_p \quad (3.35)$$

which covers  $\mathcal{P}$  (also covers  $\widetilde{M}$  [27, Page 48]). To finish the construction, we need to find a bijective map  $\eta_p : {}^*\mathcal{U}_p \rightarrow {}^*B_p$ , and show that the transition maps are indeed affine.

Let  $\mathfrak{p} \in \widetilde{M}$ , and consider the map  $\eta_{\mathfrak{p}} : {}^*\mathcal{U}_{\mathfrak{p}} \rightarrow {}^*B_{\mathfrak{p}}$  defined by

$$p \mapsto \frac{p}{\mathfrak{p}} - 1 \quad (3.36)$$

Note the following:

- For any  $q \in \mathcal{E}(p)$ ,  $\eta_p(q)$  is an element of  ${}^*B_p$  which is identified with  $\mathbb{E}_q[v]$  (where  $v \in B_p$ ) and could be called an expectation parameter [27].

- $\eta_{\mathfrak{p}}$  is bijective.

- The inverse of  $\eta_{\mathfrak{p}}$  is

$$\eta_{\mathfrak{p}}^{-1} : {}^*B_{\mathfrak{p}} \ni u \mapsto (u + 1)\mathfrak{p} \in {}^*\mathcal{U}_{\mathfrak{p}} \quad (3.37)$$

- For each pair  $\alpha, \beta \in \mathcal{E}(p)$ , the composition map

$$\begin{aligned} \eta_\beta \circ \eta_\alpha^{-1} : {}^*B_\alpha &\rightarrow {}^*B_\beta \\ {}^*u &\mapsto {}^*u \frac{\alpha}{\beta} + \frac{\alpha}{\beta} - 1 \end{aligned} \quad (3.38)$$

is  $C^\infty$ -affine.

We now have all the tools to define the affine manifold that is dual to  $\mathcal{P}$ .

**Theorem 3.5.7.** [27, Theorem 36, Page 50] Let  $p \in \mathcal{P}$  be fixed. Define  ${}^*\mathcal{E}(p)$  as the subset of  $\mathcal{P}$ :

$${}^*\mathcal{E}(p) \triangleq \{q \in \mathcal{P} : \frac{q}{p} \in L^\Psi(p)\} \quad (3.39)$$

Then, the collection of charts

$$\{({}^*\mathcal{U}_{p_\alpha}, \eta_{p_\alpha}) : p_\alpha \in \mathcal{E}(p)\}$$

is an affine  $C^\infty$ -atlas on  ${}^*\mathcal{E}(p)$

We denote the manifold of the mixture geometry of  $\mathcal{P}$  by  ${}^*\mathcal{P}$ . The following proposition establishes the important relation between  $\mathcal{E}(p)$  and  ${}^*\mathcal{E}(p)$ .

**Proposition 3.5.8.** [27, Proposition 38 and 39] For each probability measure  $p \in \mathcal{P}$ , we have

- the injection  $\alpha$  of  ${}^*B_p$  into  $(B_p)^*$

$$\alpha : {}^*B_p \ni {}^*u \mapsto \mathbb{E}_p[{}^*u] \in (B_p)^* \quad (3.40)$$

has a closed range, i.e. image of  $\alpha$ , therefore image of  $\alpha$  is a subspace of  $(B_p)^*$  and  $\alpha$  is a isomorphism of  ${}^*B_p$  into its range in  $(B_p)^*$ .

- the inclusion  $\iota_p : \mathcal{E}(p) \hookrightarrow {}^*\mathcal{E}(p)$  is of class  $C^\infty$ .

This construction highlights a rather interesting feature of the dual affine structures of  $\mathcal{P}$ . As we highlighted before, the boundary  $\mathcal{P}$  is rather different in the different affine structures.

It is important to note that this problem in duality is due to the insistence that the topology of the generalized (+1) parameterization includes the well-behaved exponentially

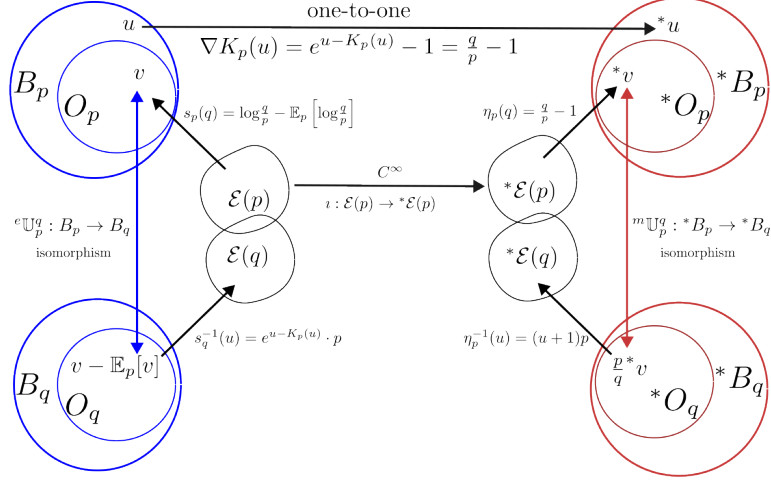


Figure 3.1: Infinite-dimensional manifold structure of a set of probability measures that agree on a set of measure zero

integrable random variables, which is a standard practice in statistics. It is an open area of research to try and simplify this construction. For more information, see the excellent reviews by [30, 68].

Figure 3.1 summarizes the dual manifold structures of  $\mathcal{P}$  highlighting the exponential and mixture Banach manifolds and the associated parallel transports.

This concludes our presentation of the manifold portion of the theory of statistical bundles on  $\mathcal{P}$ . We continue the presentation of the theory and cover the all important tangent bundle structure.

### 3.6 Tangent spaces

Similar to the tangent space construction in the finite-dimensional case (2.6.5), we can characterize the tangent space to  $\mathcal{P}$  at a point by the Quotient set of all regular curves passing through the point modulo the equivalence relation that any two curves are equivalent if their tangent vectors at the point are in the same direction.

In a neighborhood of  $p \in \mathcal{P}$ , a curve  $\gamma(t)$  centered at  $p$  has a coordinate representation  $u(t) \in B_\alpha$  under a chart  $s_\alpha : \mathcal{U}_\alpha \rightarrow B_q$ . As a result, we can write the curve  $\gamma(t) =$

$$e^{u(t)-K_\alpha(u(t))} \cdot \alpha$$

The tangent vector  $\dot{u}(0)$  is connected to the score function of  $\gamma(t)$  with respect to  $p$  as follows.

$$\begin{aligned} \dot{\gamma}(t) \Big|_{t=0} &= \frac{d}{dt} \left[ u(t) - K_\alpha(u(t)) \right] \cdot \gamma(t) \Big|_{t=0} \\ &\Downarrow \\ \dot{u}(t) - K_\alpha(\dot{u}(0)) &= \frac{\dot{\gamma}(t)}{\gamma(t)} \Big|_{t=0} \\ &= \frac{\dot{\gamma}(t)}{p} \frac{p}{\gamma(t)} \Big|_{t=0} \\ &= \frac{d}{dt} \left( \log \frac{\gamma(t)}{p} \right) \Big|_{t=0} \end{aligned}$$

If the chart is centered (i.e.  $\alpha = p$ ), we can write

$$\dot{u}(0) = \frac{d}{dt} \left( \log \frac{\gamma(t)}{p} \right) \quad (3.41)$$

In other words, the score function  $\gamma(t)$  with respect to  $p$  (in a chart centered at  $p$ ) are elements in the tangent space  $T_p\mathcal{P}$ . In addition to the identification of  $T_p\mathcal{P}$ , it should be clear that the tangent space  $T_p\mathcal{P}$  inherits the vector space structure and topology from  $B_p$  [71] where we can identify tangent vectors at  $p$  by vectors in  $B_p$ .

With that identification, one can easily use the derivative of the cumulant generating functional to define a bilinear form on the product space  $B_p \times B_p$  and establish a notion of orthogonality between vectors in  $B_p$ . In addition, we can use the Banach duality between  $B_p$  and its predual  ${}^*B_p$  to establish a notion of orthogonality on  $B_p \times {}^*B_p$  using the relation (3.30)

**Definition 3.6.1.** [70, Definition 10, Page 730] Let  $p$  be a point in  $\mathcal{P}$ .

- Define the scalar product  $\langle \cdot, \cdot \rangle_p : B_p \times B_p \rightarrow \mathbb{R}$  by the second derivative of the cumulant generating functional  $K_p$  (3.15) in the direction  $u, v \in B_p$ , i.e

$$\begin{aligned} \langle u, v \rangle_p &= (d^2 K_p)(u, v) \\ &= \mathbb{E}_p[u \cdot v] - \mathbb{E}_p[u] \mathbb{E}_p[v] \end{aligned} \quad (3.42)$$

- We say that  $u$  and  $v$  are orthogonal if  $\langle u, v \rangle_p = 0$
- Define the bilinear form  $\langle \cdot, \cdot \rangle_{*,p} : B_p^* \times B_p \rightarrow \mathbb{R}$  by

$$\langle *u, v \rangle_{*,p} = \mathbb{E}_p[*uv] \quad (3.43)$$

- We say that  $*u \in *B_p$  and  $v \in B_p$  are orthogonal if  $\langle *u, v \rangle_{*,p} = 0$

The above bilinear forms  $\langle \cdot, \cdot \rangle_p, \langle *u, v \rangle_{*,p}$  can be thought of as a non-parametric generalization of the Fisher metric tensor in the finite-dimensional case.

The set  $\mathcal{P}$  has another geometry, the mixture manifold geometry  $*\mathcal{P}$  we outlined in (3.5.7). In the infinite-dimensional case, because Banach spaces are not reflexive, we must distinguish between the tangent space of the exponential manifold of  $\mathcal{P}$  and its predual tangent space [27]. In fact, the predual tangent space of  $\mathcal{P}$  is the tangent space of the mixture manifold geometry  $*\mathcal{P}$  [68].

Before presenting the tangent bundle and predual tangent bundles of  $\mathcal{P}$ , we need to define the nature of the transformation over the tangent space as we move between different points in  $\mathcal{E}(p)$  (or  $*\mathcal{E}(p)$ ).

For the exponential manifold, it is trivial to show that the derivative of the transition maps (3.24) between overlapping coordinate charts  $s_\alpha : \mathcal{U}_\alpha \rightarrow B_\alpha$  and  $s_\beta : \mathcal{U}_\beta \rightarrow B_\beta$  is an isomorphism, i.e.

$$(d_v s_\beta \circ s_\alpha^{-1})(u) = v - \mathbb{E}_\beta[v], \quad u \in B_\alpha, v \in B_\beta \quad (3.44)$$

is an isomorphism between  $B_\alpha$  and  $B_\beta$  which we denote by  ${}^e\mathbb{U}_\alpha^\beta v$ .

For the mixture manifold  $*\mathcal{P}$ , one can show that the mapping

$${}^m\mathbb{U}_\alpha^\beta : *B_\alpha \ni *u \mapsto \frac{\alpha}{\beta} *u \in *B_\beta \quad (3.45)$$

is an isomorphism of the tangent spaces  $*B_\alpha$  onto  $*B_\beta$  to a point  $p \in *\mathcal{P}$  such that  $p \in *B_\alpha \cap *B_\beta$ .

**Definition 3.6.2.** (Tangent and Pretangent Bundles)[68, Definition 12]

- The set  $T\mathcal{P} \triangleq \{(p, u) : p \in \mathcal{P}, u \in B_p\}$  together with the charts  $\{(q, u) : q \in \mathcal{E}(p), u \in B_p\} \rightarrow B_p \times B_p$  defined by

$$(q, v) \mapsto (s_p(q), {}^e\mathbb{U}_p^q v)$$

define a tangent bundle  $T\mathcal{P}$  of  $\mathcal{P}$ .

- The set  $\{(q, {}^*u) : q \in \mathcal{P}, {}^*u \in {}^*B_p\}$  together with the charts  $\{(q, u) : q \in \mathcal{E}(p), {}^*u \in {}^*B_p \rightarrow B_p \times {}^*B_p\}$  defined by

$$(q, v) \mapsto (s_p(q), {}^m\mathbb{U}_p^q)$$

is a pretangent bundle  ${}^*T\mathcal{P}$ .

The maps  ${}^e\mathbb{U}_p^q : B_p \rightarrow B_q$  (3.44), and  ${}^m\mathbb{U}_p^q : {}^*B_p \rightarrow {}^*B_q$  (3.45) play an important role in the two geometries of  $\mathcal{P}$ . In the pretangent bundle of  $\mathcal{P}$ , the action of the dual  $({}^e\mathbb{U}_p^q)^*$  is identified by  ${}^m\mathbb{U}_p^q$  [68]. More precisely,  ${}^e\mathbb{U}_p^q$  and  ${}^m\mathbb{U}_p^q$  are dual semigroups, i.e.

$$\langle {}^e\mathbb{U}_p^q u, v \rangle_q = \langle u, {}^m\mathbb{U}_p^q v \rangle_q, \quad \langle w, v \rangle_q = \langle {}^e\mathbb{U}_p^q w, {}^e\mathbb{U}_p^q v \rangle_p \quad (3.46)$$

In fact, they are the parallel transports in the exponential and mixture tangent bundles (or the pretangent bundle), respectively [68]. Intuitively speaking, a parallel transport of tangent vectors ensures that as we transport them along curves in the manifold, the tangent vectors stay parallel. To formally define these notions, the introduction of the notion of connections and covariant derivatives is required. For more details, see [30].

One can couple the tangent and pretangent bundles of  $\mathcal{P}$  to produce a new bundle

$$({}^*T \times T)\mathcal{P} = \{(p, v, w) : p \in \mathcal{P}, v \in B_p, w \in {}^*B_p\}$$

with the duality coupling

$$({}^*T \times T)\mathcal{P} \ni (p, v, w) \mapsto \langle v, w \rangle_p = \mathbb{E}_p[uv] = \mathbb{E}_q[{}^m\mathbb{U}_p^q v {}^e\mathbb{U}_p^q w]$$

This concludes our discussion on tangent spaces. So far, we have not presented the general notion of having a complementary orthogonal notion between the tangent bundle and the pretangent bundle. In the next section, we present the modern approach to the problem and develop a projection operator for approximating the conditional density process of a stochastic filtering problem.

### 3.7 Duality and Projections

Recall that in the finite-dimensional case, the basis of the tangent space of the (+1) representation is complementarily orthogonal to the basis of the tangent vector of the (-1) representation (2.36). In the finite-dimensional case this was rather trivial because the

manifold was equipped with a Riemannian metric, allowing us to establish an orthogonal notion between the tangent vectors in the different representations. In infinite-dimensional settings, our manifold  $\mathcal{P}$  is not Riemannian and does not have a Hilbert structure. As a result, a splitting of the tangent spaces is required if we are to establish a similar notion between the tangent and pretangent spaces at a point.

Let  $p$  be a probability measure in  $\mathcal{P}$ . A splitting of  $B_p$  is a pair  $(V_p^1, V_p^2)$  such that

$$B_p = V_p^1 + V_p^2, \quad V_p^1 \cap V_p^2 = \{0\}$$

where  $V_p^1, V_p^2$  are two closed subspaces of  $B_p$ . In other words, any element  $u \in B_p$  can be written uniquely as the sum of two elements  $u_p^1 \in V_p^1$  and  $u_p^2 \in V_p^2$ . As a result, we can define a projection map  $pr_i : B_p \rightarrow V_p^i, i = 1, 2$  defined by the splitting  $pr_i(u) = u_i$ . Define the annihilating subspaces of  $B_p^*$ , and  ${}^*B_p$  as follows.

$$\begin{aligned} (V_p^i)^0 &= \{u^* \in B_p^* : \langle u^*, u_j \rangle_{*,p} = 0, \forall u_j \in V_p^j, j \neq i\} \\ {}^0(V_p^i) &= \{{}^*u \in {}^*B_p : \mathbb{E}_p[{}^*uu_j] = 0, \forall u_j \in V_p^j, j \neq i\} \end{aligned}$$

with  $i, j \in \{1, 2\}$

*Remark.*  ${}^0(V_p^1), {}^0(V_p^2)$  are closed in  ${}^*B_p$  and  ${}^0(V_p^1) \cap {}^0(V_p^2) = \{0\}$ .

**Proposition 3.7.1.** [70, Proposition 28, Page 744]  $(V_p^1)^0$  and  $(V_p^2)^0$  split in  $B_p^*$  and any element of  $B_p^*$  can be written as  $u^* = u_1^* + u_2^*$ , with  $u_i^* \in (V_p^i)^0, i = 1, 2$ . If  $v \in B_p, v = v_1 + v_2, v_i \in V_p^i$ , then  $u_i^* = u^* \circ Pr_i(v)$ ; so  $u_i^* = u^* \circ Pr_i$ , for  $i = 1, 2$ .

One way to characterize  $u_i^*$  is by the partial derivative

$$u_i^* = \left. \frac{d}{dt} K_p(u + pr_i(tv)) \right|_{t=0}$$

To establish a notion of orthogonal duality between a splitting of  $B_p$  and that of  ${}^*B_p$ , we first need to map components in the split of  $B_p^*$  onto components in the split of  ${}^*B_p$ ; however, in general  $u_i^* \in V_p^i$  is not in  ${}^*B_p$  even if  ${}^*u$  is [70]. In its full generality, this is a large obstacle in establishing orthogonal duality between the dual geometries of  $\mathcal{P}$ . Note that this is still an open problem [27].

For the purpose of this thesis, a particular subbundle to the maximal exponential bundle was introduced by restricting the space of parameters to the so-called Orlicz-Sobolev space. For the time being, let us assume that such a splitting exists and proceed to define the notion of orthogonal projection.

Suppose that  $V_p^1 + V_p^2 = B_p$ , for some  $p \in \mathcal{E}(p)$ . We can think of the cumulant generating functional as a function of two variables, i.e.

$$\begin{aligned} K_p(\cdot, \cdot) : V_p^1 \times V_p^2 &\rightarrow \mathbb{R} \\ (w_1, w_2) &\mapsto K_p(w_1, w_2) = K_p(w_1 + w_2) \end{aligned}$$

We can take partial derivatives by composing the relevant projection splitting maps  $pr_i : u \mapsto u_i$ .

**Definition 3.7.1.** (Fisher information operator)[70, Definition 29] The value at  $q = e_p(u)$  of the Hessian linear operator from  $B_p$  to  $B_p^*$  of  $K_p$  at  $u$  will be denoted by  $G(p, q)$ :

$$\langle G(p, q)w, v \rangle_{*,p} = (d_u^2 K_p)(w, v), \quad w, v \in B_p \quad (3.47)$$

and it is called a Fisher information operator.

Given a splitting  $v = v_1 + v_2$  we can consider the partitioned operators  $G_{ij}$ , with  $i, j \in \{1, 2\}$ , restricted from  $V_p^j$  to  $(V_p^k)^o$ ,  $k \neq i$ , such that

$$\langle G_{ij}(p, q)w_j, v_i \rangle_{*,p} = \langle G(p, q)w_j, v_i \rangle_{*,p}.$$

By the definition of the derivative of the cumulant generating functional, we have

$$\text{Cov}_q[w, v] = \langle w - \mathbb{E}_q(w), v - \mathbb{E}_q(v) \rangle_q = \langle G(p, q)w, v \rangle_{*,p} \quad (3.48)$$

In the infinite-dimensional manifold, a local Pythagorean-type relation can be admitted.

**Proposition 3.7.2.** [27] Let  $\alpha \in \mathcal{P}$ ,  $(\mathcal{U}_\alpha, s_\alpha)$  and  $({}^*\mathcal{U}_\alpha, \eta_\alpha)$  be two charts of  $\mathcal{E}(\alpha)$  and  ${}^*\mathcal{E}(\alpha)$  respectively such that  $p \in \mathcal{U}_\alpha$ ,  $u \in s_\alpha(p)$ , and  $0 \leq \mathfrak{q} \in {}^*\mathcal{U}_\alpha$ .

If  $\langle \eta_\alpha(\mathfrak{q}), s_\alpha(\mathfrak{q}) \rangle = 0$ , then we have

$$D_{KL}(\mathfrak{q}, p) = D_{KL}(\mathfrak{q}, \alpha) + D_{KL}(\alpha, p) \quad (3.49)$$

where  $D_{KL}$  is a Kullback-Leibler divergence.

An important feature of the statistical bundles is the possibility to define Orlicz-Sobolev for the tangent spaces and use that to solve projection problems. These spaces were

introduced for the study of partial differential equations in the statistical bundle [23]. In particular, consider the parabolic equation

$$\frac{\partial}{\partial t} \pi_t(x) = \mathcal{L} \pi_t(x)$$

which one can write as

$$\frac{1}{\pi_t(x)} \frac{d}{dt} \pi_t(x) = \frac{1}{\pi_t(x)} \mathcal{L} \pi_t(x) \quad (3.50)$$

Note that the left side of equation (3.50), is the score of the solution curve  $t \mapsto \pi_t$ , and the right-hand side is a vector field in some statistical bundle. If one can construct a subbundle to the maximal exponential bundle (and the pretangent bundle) such that an appropriate splitting is achieved, one can use the partitioned operator  $G_{ij}$  above to project a curve in the infinite-dimensional space  $\mathcal{E}(p)$  to a finite-dimensional submanifold.

Consider the following restriction to the space of coordinates

**Definition 3.7.2.** [23, Definition 3.9]

- Let  $p \in \mathcal{P}$ , and suppose that  $\mathcal{E}(p)$  is a maximal exponential model centered at  $p$ . The exponential Orlicz-Sobolev spaces of  $\mathcal{E}(p)$  are the vector spaces

$$\begin{aligned} W_{\cosh-1}^1 &= \{u \in L^{\cosh-1}(p) \mid \partial_j u \in L^{\cosh-1}(p), j = 1, \dots, d\} \\ W_{(\cosh-1)^*}^1 &= \{u \in L^{(\cosh-1)^*}(p) \mid \partial_j u \in L^{(\cosh-1)^*}(p), j = 1, \dots, d\} \end{aligned}$$

where  $\partial_i$  refers the  $i^{\text{th}}$  derivative.

- The  $W_{\cosh-1}^1$ -exponential family at  $p$  is

$$\mathcal{E}_1(p) = \{e^{u-K_p(u)} \cdot p \mid u \in B_p \cap W_{\cosh-1}^1\}$$

the set  $V_p^1 = B_p \cap W_{\cosh-1}^1$  is a convex open set

$$V_p^1 \subset \{u \in W_{\cosh-1}^1 : \mathbb{E}_p[u] = 0\}$$

it contains all coordinate functions  $v_i$  and polynomials of order two.

It can be shown that these spaces split  $B_p^*$  and characterizes a complementary orthogonal relation between the splitting of  $B_p$  and the corresponding splitting in  ${}^*B_p$

**Definition 3.7.3.** [23, Definition 5, Page 233] Let  $\alpha \in \mathcal{P}$ .

1. A statistical differentiable exponential bundle is a manifold defined on the set

$$T\mathcal{E}_1(\alpha) \triangleq \{(p, u) : p \in \mathcal{E}_1(\alpha), u \in B_p^1\} \quad (3.51)$$

by the affine atlas of global charts

$$T\mathcal{E}_1(\alpha) \ni (q, u) \mapsto (s_p(q), {}^e\mathbb{U}_q^p u) \in B_p^1 \times B_p^1$$

where  $p \in \mathcal{E}_1(\alpha)$

2. A statistical differentiable predual bundle is the manifold defined on the set

$${}^*T\mathcal{E}_1(\alpha) \triangleq \{(p, {}^*u) : p \in \mathcal{E}_1(\alpha), {}^*u \in {}^*B_p^1\}$$

by the affine atlas of global charts

$${}^*T\mathcal{E}_1(\alpha) \ni (q, {}^*u) \mapsto (s_p(q), {}^m\mathbb{U}_q^p {}^*u)$$

If we are to narrow our attention to the differentiable exponential bundle, the following regularities are guaranteed.

**Proposition 3.7.3.** [23, Proposition 3.10, Page 16] Assume  $u \in V_\alpha^1$ ,  $p = e^{u-K_\alpha(u)}.\alpha \in \mathcal{E}_1(\alpha)$ , and  $f \in W_{\cosh-1}^1$

- It follows that  $f e^{u-K_p(u)} \in W_{(\cosh-1)^*}^1$  and  $f\alpha \in W_{(\cosh-1)^*}^1$
- $\nabla e^{u-K_p(u)} = \nabla u e^{u-K_p(u)}$ , and  $\nabla(e^{u-K_p(u)}\alpha) = (\nabla u - v) e^{u-K_p(u)}\alpha$
- If  $f \in W_{(\cosh-1)^*}^1$  and  $g \in W_{\cosh-1}^1$ , then

$$\langle f, \partial_j g \rangle_p = \langle v_j f - \partial_j f, g \rangle_p$$

To use the differentiable exponential bundles, one must ensure that the stochastic process is guaranteed to evolve in the space. For this to be the case, more regularity assumptions are required on the linear operator associated with the stochastic process. Going forward, we will assume that the solution to a stochastic filtering problem for a given class of test functions does have a solution curve in the base space of the statistical differentiable bundle without justification. The type of regularities required is still an open research area.

As we defined in (1.4.3) a projection filter requires a subspace to project on. In classical applications, one typically chooses a subspace that is also a submanifold on its own. This is not the case for statistical bundles due to the challenge of establishing a well-defined splitting of the tangent spaces. As a result, the following axioms are required to ensure that our definition of a submodel has a well-defined splitting of the tangent bundles.

**Definition 3.7.4.** [70, Definition 27] Let  $\alpha$  be an arbitrary probability measure on some probability space  $\Omega$ . Suppose that  $\mathcal{S}$  is a subset of  $\mathcal{E}(p) \subset \mathcal{P}$ , and for each probability measure  $p \in \mathcal{S}$ , let  $V_p^1$  be a closed subspace of  $B_p$  and  $V_p^2$  be a closed subspace of  ${}^*B_p$  such that  $V_p^1 \cap V_p^2 = \{0\}$  with continuous immersions  $B_p \hookrightarrow V_p^1 + V_p^2 \hookrightarrow {}^*B_p$ .

Let  $\alpha_p : \mathcal{W}(p) \rightarrow O_p^1 \times O_p^2 \subset V_p^1 \times V_p^2$  be a diffeomorphism of a neighborhood  $\mathcal{W}(p)$  of  $p$  that maps  $\mathcal{S} \cap \mathcal{W}_p$  onto  $O_p^1 \times \{0\}$ . Assume that there exists an atlas  $\mathcal{A}(\mathcal{S})$  of such mappings  $\alpha_p$  covers  $\mathcal{S}$ .

1. It follows that  $\mathcal{S}$  is a manifold with a chart  $\alpha_{p,|\mathcal{S}}, \alpha \in \mathcal{A}(\mathcal{S})$  with a tangent space  $T_p\mathcal{S}$  isomorphic to  $V_p, p \in \mathcal{S}$ . We say that such a manifold is a submodel of  $\mathcal{E}(\alpha)$ .
2. If the space  $V_p^2$  is a closed subspace of  $B_p$ , that is  $V_p^1$  splits in  $B_p$ , then  $\mathcal{S}$  is a submanifold of  $\mathcal{E}(\alpha)$ .

In the thesis we are interested only in the projection onto a finite dimensional exponential models; however, other submodels are possible, for more information, see [23]. In order to ensure that the finite-dimensional family is indeed a submodel, one can check the nature of the split of the tangent space of the model into the maximal exponential model  $\mathcal{E}(p)$ .

It can be shown that there exists a splitting chart that provides an immersion of an exponential family model into the maximal exponential family together with a complementary model given by another infinite-dimensional exponential family. Intuitively speaking, suppose that  $\gamma(t)$  is a curve in the maximal exponential model  $\mathcal{E}(\alpha)$ , for some point  $\alpha \in \mathcal{P}$ . The role of the projection operator is to parallelly transport the tangent vector to curve  $\gamma(t)$  at a point  $\gamma(t_p) = 0$  onto the tangent space of a finite-dimensional submodel of  $\mathcal{E}(p)$ .

The following example will provide an intuitive explanation of how the projection filter works in the case that we have a statistical submodel  $P \subset \mathcal{E}(p)$  where  $P$  is a finite-dimensional exponential model with an exponential family form. Let  $\alpha$  be an arbitrary probability measure on some probability space  $\Omega$ . Suppose  $p$  is a probability measure in  $W_{\cosh-1}^1$ -exponential family, and  $p \in \mathcal{E}(\alpha)$ . i.e.  $p = e^{u-K_\alpha(u) \cdot \alpha}$  where  $u \in O_\alpha^1 = O_\alpha \cap B_\alpha \cap W_{\cosh-1}^1$ .

Let  $Ap$  be a nonlinear differential operator  $\frac{1}{p}\mathcal{L}^*p$ , where  $\mathcal{L}$  is a differential operator for some stochastic process  $\pi = \{\pi_t : t \geq 0\}$  with the right regularity assumptions. i.e.,

- the operator  $Ap$  has a vector field in the differentiable mixture bundle  ${}^*\mathcal{E}(\alpha)$ , i.e.  $Ap \in {}^*B_p^1$ ;
- the domain of the operator is a subset of  $\mathcal{E}_1(\alpha)$

Recall that the differentiable predual bundle  ${}^*T\mathcal{E}(\alpha)$  has a chart centered at  $p$ , i.e

$$(q, v) \mapsto (s_p(q), {}^m\mathbb{U}_q^p) \in B_p^1 \times {}^*B_p^1$$

where  $s_p$  is the exponential chart and  ${}^m\mathbb{U}_q^p$  is the mixture parallel transport  $v \mapsto \frac{q}{p}v$ . The expression of the operator  $A$  in a chart centered at  $\alpha$  is  $u \mapsto \hat{A}_\alpha(u)$ , i.e.

$$\begin{aligned} \hat{A}_\alpha(u) &= e^{u-K_\alpha(u)} A(e^{u-K_\alpha(u)}.\alpha) \\ &= \frac{1}{\alpha} \mathcal{L}^*(e^{u-K_\alpha(u)}.\alpha) \end{aligned}$$

Note that for each  $v \in B_p^1$ ,  $\langle Ap, v \rangle_p$  is well defined. Suppose that, by an application of the exponential parallel transport  ${}^e\mathbb{U}_p^q v$ , we have finite  $n$ -dimension vector subspaces  $V_n(p)$  of the tangent space  $B_p^1$  (assuming  $V_n \in B_\alpha^1$ ). Because the exponential transport has no effect on the partial derivatives, we have  $u, v \in B_\alpha^1$

$$\langle A, {}^e\mathbb{U}_\alpha^p v \rangle_p$$

If we have a basis for  $V_n$ , say  $w_1, \dots, w_n$  such that  $w_i - \mathbb{E}_p[w_i], i = 1, \dots, n$ , we can write

$$u = \sum_{i=1}^n \theta_i w_i, \quad v = \sum_{j=1}^n \alpha_j w_j$$

which we can use to rewrite

$$\begin{aligned} \langle Ap, {}^e\mathbb{U}_\alpha^p v \rangle_p &= \langle Ap, {}^e\mathbb{U}_\alpha^p \sum_{j=1}^n \alpha_j w_j \rangle_p \\ &= \sum_{j=1}^n \alpha_j \langle Ap, {}^e\mathbb{U}_\alpha^p w_j \rangle_p \end{aligned}$$

In other words, we look for a curve  $t \mapsto \gamma(t)$ , whose score  $D\gamma(t)$  is such that

$$\langle D\gamma(t) - A\gamma(t), {}^e\mathbb{U}_\alpha^{\gamma(t)} w_k \rangle_{\gamma(t)} = 0, \quad j = 1, \dots, n \quad (3.52)$$

Since by assumption, the curve  $t \mapsto (\gamma(t), D\gamma(t) - A\gamma(t))$  belongs to the statistical bundle, the score can be written as

$$D\gamma(t) = \frac{\dot{\gamma}t}{\gamma(t)} = \sum_{j=1}^n \dot{\theta}_j(t) {}^e\mathbb{U}_\alpha^{\gamma(t)} w_j \quad (3.53)$$

it follows that

$$\begin{aligned} \langle D\gamma(t), {}^e\mathbb{U}_\alpha^p w_j \rangle_{\gamma(t)} &= \sum_{j=1}^n \dot{\theta}_i(t) \langle {}^e\mathbb{U}_\alpha^{\gamma(t)} w_i, {}^e\mathbb{U}_\alpha^p w_j \rangle_{\gamma(t)} \\ &= \sum_{j=1}^n g_{ij}(t) \dot{\theta}_i(t) \end{aligned} \quad (3.54)$$

where  $g_{ij}$  is the Fisher information matrix of the finite dimensional model, i.e.

$$g(\theta) = \left[ \left\langle {}^e\mathbb{U}_\alpha^{p_\theta} w_i, {}^e\mathbb{U}_\alpha^{p_\theta} w_j \right\rangle_{p_\theta} \right]_{i,j} = \left[ \text{Cov}_{p_\theta}(w_i, w_j) \right]_{i,j}$$

As a result, we can multiply the projection equation (3.53) by the inverse of the Fisher information matrix and sumer of  $j$  to get a system of non-linear differential equations, which intuitively speaking solve a projection problem for a flow on the finite-dimensional submanifold.

It is important to note that this projection operator is optimal in the sense of the Kullback-Leibler divergence with respect to the finite-dimensional family given a well-chosen finite-dimensional subfamily.

**Theorem 3.7.4.** [23, Theorem 6.6, Page 38] The vector field projection approach leading to 3.54 provides the best possible approximation of the Fokker-Planck-Kolmogorov equation solution in Kullback Leibler in a finite-dimensional exponential family, provided that the sufficient statistics of the family are chosen among the eigenfunctions of the adjoint operator  $\mathcal{L}$  of the original Fokker-Planck-Equation 1.2.3, and provided that the family is an exponential family when using the eigenfunctions. In other words, under such conditions the Fisher Rao projected equation 3.54 provides an exact maximum likelihood estimator for the solution for the Fokker-Planck-Kolmogorov equation in the related exponential family.

It is important to note that currently there are no convergence theorems that would guarantee that if the projected evolution is in a manifold without boundary, the projected equation converges to the true mean of the optimal filter [23]. This is another open research problem.

In the next chapter, we will motivate this projection method in a  $K$ -dimensional simplex to develop more intuition regarding how the projection operator works.

# Chapter 4

## Examples

This chapter is devoted to examples that illustrate the ideas and properties of the projection operator introduced in [23]. We can through numerical examples develop an intuition to the idea of projection by examining it using standard arguments. In addition, since most solutions are implemented numerically, understanding the key strength and limitation of the projection idea could further develop our intuition for how it works in practice.

We are generally interested in understanding how well the projection method works for large and infinite state spaces. The following are some of the key ideas and properties that we would like to explore.

- (a) How sensitive is the approximation to different choices of exponential models?
- (b) How well does the approximation behave near the boundary of the space of solutions.
- (c) Under what conditions is the approximation optimal, if any?
- (d) Is the approximation universal for a given class of problems? i.e. how well does it work for different choices of test functions?
- (e) What are the general guidelines for using the projection filter?
- (f) What are the computational challenges that might arise in applications of the projection method?

Typically, we would like to explore these properties theoretically, but due to time constraints, we will start this exploration with numerical experiments and analytical examples.

We start by formulating the projection method in a discrete state space.

## 4.1 Projection in the discrete state space

In order to explore the idea of approximating solutions of SDEs and related filtering problems using projections onto low-dimensional exponential families, we can run numerical experiments in the environment of the finite-state case. Hopefully, this setting allows us to develop intuition about the infinite-dimensional case and understand the numerical strengths and limitations of the projection idea, since all numerical computational solutions are fundamentally finite.

We begin by formulating the projection idea in the discrete state space.

**Definition 4.1.1.** (Continuous time Markov process)

- (a) A square matrix  $A(x, y)$  is called a Markov generator on a finite state space  $\mathcal{X}$  if:
- (i) for all  $x \in \mathcal{X}$  we have  $\sum_{y \in \mathcal{X}} A(x, y) = 0$  and
  - (ii)  $A(x, y) \geq 0$  for all  $x \neq y \in \mathcal{X}$ .
- (b) Let  $P$  be a  $N \times N$  stochastic matrix and set  $A := \lambda(P - I_{N \times N})$  for  $\lambda > 0$ . This defines the semigroup  $P_t := e^{tA}$  with an infinitesimal generator  $A$  using the exponential function on the matrices.
- (c) The corresponding continuous-time Markov process has i.i.d. exponentially distributed jumping times with rate  $\lambda > 0$  and we denote by  $X_t^x$  the value of the process at time  $t$  starting at  $x$  at  $t = 0$ .
- (d) For a (test) function  $\varphi(x)$  we define

$$P_t \varphi(x) = \mathbb{E}[\varphi(X_t) | X_0 = x] = e^{tA} \varphi(x).$$

Consider then the differential equation

$$\frac{\partial}{\partial t} P_t = \frac{\partial}{\partial t} e^{tA} = A e^{tA}$$

with an initial distribution  $\pi_0(x)$ . We have the distribution of  $X_t^x$  being defined by

$$\pi_t(x) = P_t \pi_0 = e^{tA} \pi_0(x). \tag{4.1}$$

and

$$\frac{\partial}{\partial t} \pi_t(x) = \frac{\partial}{\partial t} P_t \pi_0(x) = A e^{tA} \pi_0(x) = A \pi_t(x) \tag{4.2}$$

which corresponds to a vector field on the space of all probability measures on  $\mathcal{X}$  which is  $\Delta^{(K)}$ , the closed  $K$ -simplex corresponding to the general  $K$ -dimensional extended multinomial model [33] where  $K = N - 1$ .

We start with a classical setting using a standard limiting argument on finite-state processes.

**Example 4.1.1.** Consider a random walk model on  $\mathcal{X} = \{1, \dots, N\}$  such that the one step transition matrix is defined by

$$P_{j,k} := \mathbb{P}[X_{m+1} = k | X_m = j] = \begin{cases} \theta_k, & k = j + 1 \\ 1 - \theta_k - \phi_k & k = j \\ \phi_k & k = j - 1 \\ 0, & \text{otherwise} \end{cases}$$

for  $x \in \{2, \dots, N - 1\}$  and we have absorbing boundary conditions

$$\mathbb{P}[X_1 = 1 | X_0 = 1] = \mathbb{P}[X_1 = N | X_0 = N] = 1.$$

so if  $N = 5, \theta = 0.3, \phi = 0.3, \lambda = 2$  gives

$$P = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.4 & 0.3 & 0.0 & 0.0 \\ d & 0.0 & 0.3 & 0.4 & 0.3 \\ 0.0 & 0.0 & 0.3 & 0.4 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{pmatrix}, A = \begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.6 & -1.2 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.6 & -1.2 & 0.6 & 0.0 \\ 0.0 & 0.0 & 0.6 & -1.2 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

Consider a low dimensional exponential family subfamily of  $\Delta^{(K)}$ , say  $\mathcal{M}$ , defined by

$$\pi_k(\theta) = \pi_k^0 \exp \left[ \sum_{q=1}^Q S_k^{(q)} \theta_q - \psi(\theta) \right]. \quad (4.3)$$

where  $\psi$  is the cumulant generating function. By standard arguments this has mixture flat orthogonal fibers

$$F(\mu_1, \dots, \mu_Q) := \left\{ \rho \in \Delta^{(K)} : \sum_{k=0}^K S_k^{(1)} \rho_k = \mu_1, \dots, \sum_{k=0}^K S_k^{(Q)} \rho_k = \mu_Q \right\},$$

and for all  $\rho \in F(\mu_1, \dots, \mu_Q)$ . The Kullback-Leibler projection of  $\rho$  on to  $\mathcal{M}$  has mean parameters  $(\mu_1, \dots, \mu_Q)$ .

The key intuition from this definition is that the complete simplex can be partitioned into mixture flat orthogonal fibers with each fiber being defined by having the same Kullback-Leibler projection onto  $\mathcal{M}$  (i.e. they are all orthogonal to  $\mathcal{M}$  in the sense of the Fisher metric). As we have shown in Chapter 3, the generalization of this finite-dimensional space to infinite dimensions has considerable technical problems. As a result, to build intuition, we will focus on the finite-dimensional, but closed, case of the finite simplex.

Within the closed finite-dimensional simplex we can explore the projection method by considering a differential operator  $A$  which defines a one-dimensional family of probability mass functions  $\rho : [0, \infty) \ni t \mapsto (\rho_0(t), \dots, \rho_K(t)) \in \Delta^{(K)}$  by a set of differential equations

$$\frac{d\rho_k}{dt}(t) = (A\rho)_k \quad (4.4)$$

for all  $t \geq 0$ , all  $k$ , and an initial conditions  $\rho(0) = (\rho_k(0))_{k=0}^K$ . We are interested in tracking the evolution of the expected value of test functions  $\varphi : \mathcal{X} \rightarrow \mathbb{R}$  under (4.4) i.e to compute

$$\mu_\varphi(t) := \sum_{k=0}^K \varphi(k) \rho_k(t), \quad t \in [0, \infty)$$

Rather than directly solving these  $K+1$ -equations to get  $\rho(t)$ , we look at the Kullback-Leibler projection of  $\rho(t)$  onto  $\mathcal{M}$  and see if there exists a lower dimension differential equation on  $\mathcal{M}$  which would also generate the required evolution at least for a subset of possible test functions.

**Definition 4.1.2.** Let  $\rho(t)$  be a solution of Equation 4.4 and  $\mathcal{M}$  be the  $Q$ -dimensional exponential family defined in (4.3). Denote the Kullback-Leibler projection of  $\rho(t)$  onto  $\mathcal{M}$  by  $\hat{\theta}(t) = (\hat{\theta}_1(t), \dots, \hat{\theta}_Q(t))$ . For the test function  $\varphi$  we have the identity

$$\mu_\varphi(t) := \sum_{k=0}^K \varphi(k) \rho_k(t) = \sum_{k=0}^K \varphi(k) \pi_k \left( \hat{\theta}[\mu(t)] \right) \quad (4.5)$$

for all  $t \geq 0$ , where  $\mu(t) := \sum_{k=1}^K k \rho_k(t)$ . This follows immediately from the definition 4.3 since  $\rho(t)$  and  $\pi(\hat{\theta}[\mu(t)])$  always lie on the same fiber  $F(\mu_0^{\rho(t)}, \dots, \mu_Q^{\rho(t)})$  (that is, they have the same values of the mean parameters).

To make Definition 4.1.2 operationally useful, we require that there is a differential equation of lower dimension than  $K$  on  $\mathcal{M}$  defined by (4.5) which can be solved *without*

knowing the curve  $\rho(t)$ . One way to understand the form of this equation is that we parallelly transport the tangent vectors  $A\rho_t$  along the mixture orthogonal fibers  $F(\mu(t))$ , thus we have the following necessary condition

$$\sum_{k=0}^K \varphi(k) \left( \rho_k(t) - \pi_k(\widehat{\theta}(\mu(t))) \right) = 0$$

which follows directly from the identity (4.5). Observe that by differentiating both sides we have

$$\begin{aligned} 0 &= \sum_{k=0}^K \varphi(k) \left( \frac{d}{dt} \rho_k(t) - \frac{d}{dt} \pi_k(\widehat{\theta}(\mu(t))) \right) \\ &= \sum_{k=0}^K \varphi(k) \left( (A\rho)_k - \sum_{r=1}^Q \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) \frac{\partial \theta_r(t)}{dt} \right) \\ &= \langle \varphi, A\rho \rangle - \sum_{k=0}^K \varphi(k) \sum_{r=1}^Q \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) \frac{\partial \theta_r(t)}{dt} \\ &= \langle \varphi, A\rho \rangle - \sum_{k=0}^K \varphi(k) \sum_{r=1}^Q \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) \sum_{s=1}^Q \frac{\partial \theta_r(t)}{\partial \mu_s} \frac{d\mu_s}{dt}(t) \\ &= \langle \varphi, A\rho \rangle - \sum_{r,s=1}^Q \left[ \sum_{k=0}^K \varphi(k) \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) \right] I^{rs}(\theta(t)) \frac{d\mu_s}{dt}(t) \end{aligned}$$

where  $I^{rs}(\mu)$  is the inverse of the Fisher matrix

$$I_{rs}(\mu) = \text{Cov}(S^{(r)}, S^{(s)})$$

If we assume that  $\varphi$  lies in the span of the canonical statistic  $S$  of  $\mathcal{M}$  such that

$$\varphi(k) = \sum_{q=1}^Q \alpha_q S_k^{(q)}$$

we get

$$\begin{aligned}
\sum_{r=0}^Q \left[ \sum_{k=0}^K \varphi(k) \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) \right] &= \sum_{q,r=1}^Q \alpha_q \left[ \sum_{k=0}^K \left( S_k^{(q)} - \mu_r(t) \right) \pi_k(\theta(t)) \right] \\
&= \sum_{q,r=1}^Q \alpha_q \left[ \sum_{k=0}^K \left( S_k^{(q)} - \mu_q(t) \right) \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) \right] \\
&= \sum_{q,r=1}^Q \alpha_q I_{qr}(\theta(t))
\end{aligned}$$

giving

$$\langle \varphi, A\rho \rangle = \sum_{q=1}^Q \alpha_q \frac{d\mu_q}{dt}(t)$$

which is the projection Equation 3.54 derived in Chapter 3 (albeit in the finite-dimensional simplex).

Thus we have the necessary condition that the flow on  $\mathcal{M}$  should only depend on the fiber, i.e.

$$\langle \varphi, A(\rho + w) \rangle = \sum_{q=1}^Q \alpha_q \frac{d}{dt} \mu_q(t)$$

for any  $w = (w_0, \dots, w_K)$  such that

$$\sum_{k=0}^K w_k = \sum_{k=0}^K S_k^{(1)} w_k = \dots = \sum_{k=0}^K S_k^{(Q)} w_k = 0$$

or

$$\left\langle \sum_{q=1}^Q \alpha_q S_k^{(q)}, A(\rho + w) \right\rangle = \sum_{q=1}^Q \alpha_q \frac{d}{dt} \mu_q(t)$$

In the information geometric form, this can be written as

$$\left\langle \sum_{q=1}^Q \alpha_q (S_k^{(q)} - \mu_q), Aw \right\rangle = 0 \tag{4.6}$$

or for all  $q$

$$\langle S_k^{(q)}, Aw \rangle = 0$$

which would hold if  $S_k$  is a left eigenvector of  $A$ . This highlights the importance of knowing the eigenfunction of the operator that generates the flow of the differential equation of interest, which is rather difficult in the general case.

We are now ready to formulate the projection idea in the discrete state space.

**Definition 4.1.3.** (KL Projection algorithm) We want, for a given generator  $A$  and choice of test function  $\varphi(x)$ , to construct a  $Q$ -dimensional exponential family so that there exists a flow on the exponential family which agrees with  $P_t\varphi(x)$  and corresponds to the Kullback-Leibler projection of  $P_t\varphi(x)$  onto the low-dimensional exponential family.

We use the fiber  $F(\mu_1, \dots, \mu_Q)$  which is defined by vectors  $w$  which satisfy

$$\langle 1_N, w \rangle = \langle S^q, w \rangle = 0,$$

for  $q = 1, \dots, Q$ . We also require the unbiasedness condition i.e.

$$\langle S^{(q)}, Aw \rangle = 0$$

for all  $q$ .

The following construction satisfies the conditions for  $\varphi$  in the space spanned by  $\{S^{(1)}, \dots, S^{(Q)}\}$  which are all (left) eigenfunctions of the infinitesimal operator  $A$ . We solve the  $Q$ -dimensional equation in the mean parameter space of  $\mathcal{M}$

$$\frac{d}{dt}\mu_q(t) = \sum_{k=0}^K \left( S_k^{(q)} - \mu_q(t) \right) (A\pi(\mu(t)))_k \quad (4.7)$$

when  $\varphi$  lies in the span of the sufficient statistics via

$$\varphi(k) = \sum_{q=1}^Q \alpha_q S_k^{(q)}$$

we have the flow of the test function under  $A$

$$\sum_{q=1}^Q \alpha_q \mu_q(t)$$

where  $(\mu_1(t), \dots, \mu_Q(t))$  solves Equations (4.7).

*Remark.* Note that the above unbiasedness condition is the same as the condition 3.52 in the infinite-dimensional statistical bundle of Chapter 3 with the simplification of being in a finite-dimensional simplex.

Intuitively speaking, by parallelly transporting the tangent vector  $A\varphi$  along the mixture fibre  $F(\mu_\varphi(t))$ , we end up with a differential equation  $\frac{d}{dt}\mu_\varphi(t)$  on  $\mathcal{M}$  that approximates the mean of the true evolution.

In addition, as was established by Theorem 3.7.4, this approximation is optimal in the Kullback Leibler sense for any lower-dimension exponential family.

## 4.2 Numerical experiments

We first explore the projection idea by solving all the equations directly in the closed simplex  $\Delta^{(K)}$  in cases where  $K$  is small. This, of course, is not the case of interest, i.e.  $K$  being large. However, having a computable representation of the  $Q$ -dimensional exponential family  $\mathcal{M}$  could give insight into the properties of the projection idea and allow us to visualize the projected solution. Later, we will consider the case where  $K$  large.

We can write the  $Q$ -dimensional exponential family (4.3) in terms of  $K + 1$  equations

$$\begin{aligned} \frac{d\pi_k}{d\mu_q}(\theta) &= \sum_{s=1}^Q \frac{d\pi_k}{d\theta_s} \frac{\partial\theta_s}{\partial\mu_q} = \left[ \sum_{s=1}^Q \left( S_k^{(s)} - \mu_s(t) \right) \pi_k(\theta(t)) \frac{\partial\theta_s(t)}{\partial\mu_q} \right] \\ &= \sum_{r,s=1}^Q \left( S_k^{(r)} - \mu_r(t) \right) \pi_k(\theta(t)) I^{sq}(\mu) \end{aligned}$$

which defines the  $Q$ -dimensional mean parameter space.

### 4.2.1 Boundary and initial condition effects

**Example 4.2.1.** In order to clearly visualise the projection, consider the case  $K = 2$ ,  $Q = 1$  and  $S_k = k$  (identity mapping). In this case, the  $Q$ -dimensional exponential family is a solution of

$$\frac{d\pi_k(\theta)}{d\mu} = \frac{(S_k - \mu)}{\text{Var}_\mu(S)} \pi_k(\theta(\mu)).$$

For the unbiasedness condition to hold, we write

$$\frac{d\rho_k(t)}{dt} = A(\rho_t)_k = \frac{(S_k - \mu)}{\text{Var}_\mu(S)} \pi_k(\theta(\mu)) + w_k$$

where

$$\sum_{k=0}^K w_k = \sum_{k=0}^K k w_k = 0.$$

Figure 4.1 shows the solutions to the above equations, where the blue curve in the simplex is the evolution in  $\mathcal{M}$ , the low-dimensional exponential family, while the red curve is the true solution of the original differential equation. The agreement of the mean values is shown in the right panel. Note that both solutions can reach the boundary of the simplex at which point the solution to the corresponding equations vanishes (it does not make sense to speak of tangent vectors at the boundary). In addition, while there is agreement on the value of  $\mu(t)$  in the relative interior of the simplex, there is no agreement on the range of  $\mu$ .

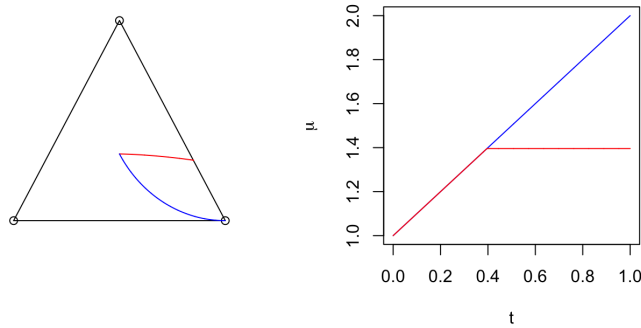


Figure 4.1: Evolution of the true equation in red, and its projection in the exponential family in blue with the same initial condition. Right hand panel shows the calculated mean value  $\mu(t)$  for both.

To further study this behavior, we examined the degree of agreement between the true solution and the projected one under different initial conditions. Figure 4.2 shows the solutions to the same equations, but under two different initial conditions. Notice that the projected solution is rather optimal before the true solution approaches the boundary, similar to the case when the initial condition was in the middle of the simplex.

The projected simulation shows a remarkable result. Regardless of the nature of the evolution, as long as the true evolution is not close to the boundary, the mean statistics match exactly.

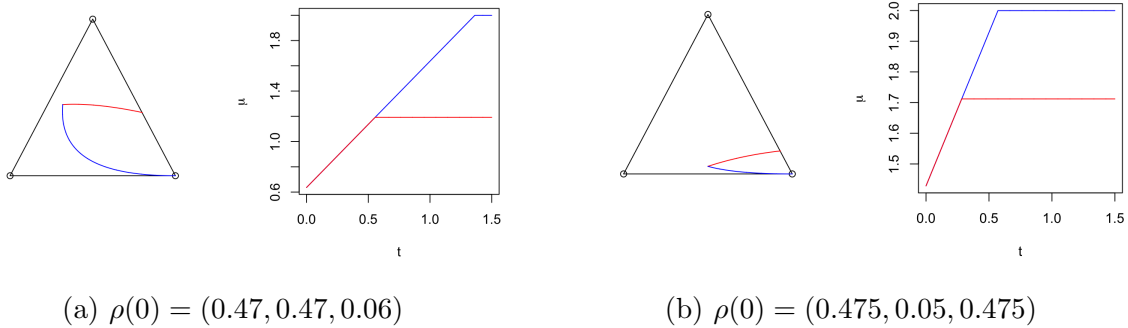


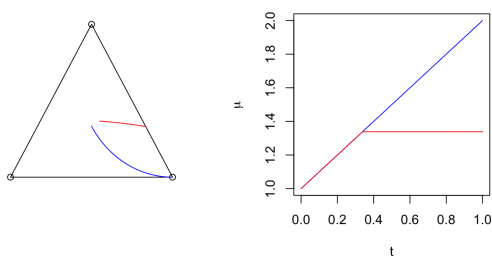
Figure 4.2: A visualization of the solution to a differential equation in red, its projection on a 2-dimensional exponential family in blue with the same initial conditions, and the calculated mean value for both.

Figures 4.2 and 4.1 above highlight the rather less understood behavior of the projection filter in the entire infinite-dimensional space. Boundary effects have, as expected, a rather large effect on the quality of the approximation. This also emphasizes that one should study the nature of the domain of the operator  $A$  and whether or not its domain has any boundaries.

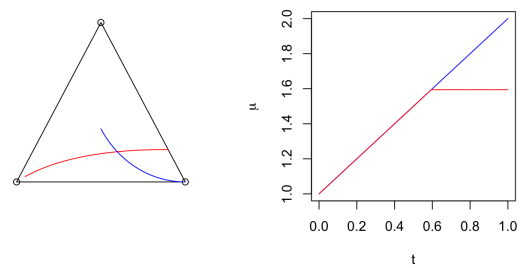
Another property that is rather less clear is the sensitivity of the projected evolution to the initial condition being an element in the lower-dimensional manifold. This puts an additional restriction on the choice of the lower-dimensional manifold.

Figure 4.3 shows the impact of the initial condition not being an element of the lower-dimensional manifold. In panel (a), and (b) note that changes in the initial conditions do not impact the mean estimates while in panel (c), and (d) they do. This can be explained intuitively because different initial conditions imply different differential equation solutions; however, the level of change might depend on the degree of shift in the projected vector field on the lower-dimensional manifold.

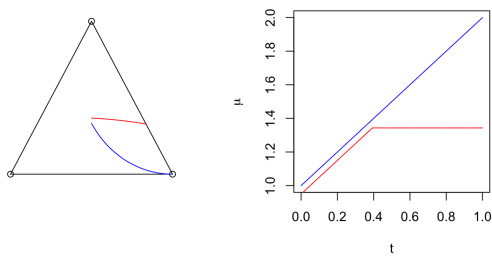
Although this example is extremely simple, it does demonstrate a surprising property of the proposed projection method. The solutions in the simplex and the lower dimensional exponential family agree when they exist; however, the solution on the exponential family



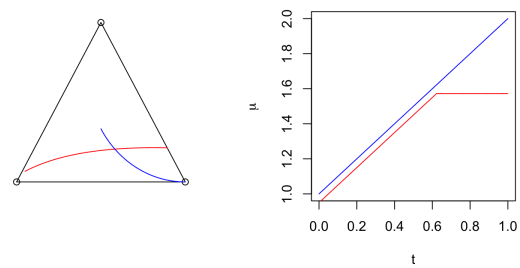
(a)  $\rho(0) = (0.365, 0.27, 0.365)$



(b)  $\rho(0) = (0.034, 0.932, 0.034)$



(c)  $\rho(0) = (0.366, 0.317, 0.317)$



(d)  $\rho(0) = ((0.067, 0.916, 0.017)$

Figure 4.3: A visualization of the solution to a differential equation in red, its projection on a 2-dimensional exponential family in blue not containing the initial condition, and the calculated mean value for both.

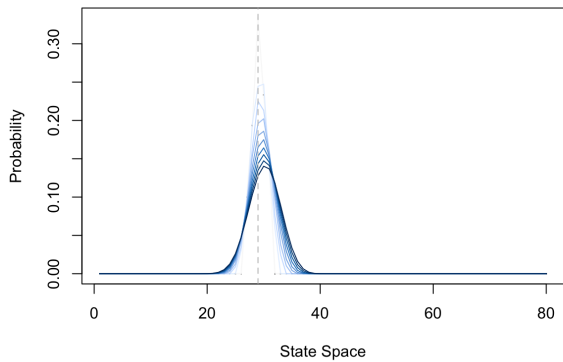
has a different range and does not predict the range of the true solution  $\rho(t)$ . This property needs care and is not explicitly discussed in [23].

## 4.2.2 Visualizing the flow

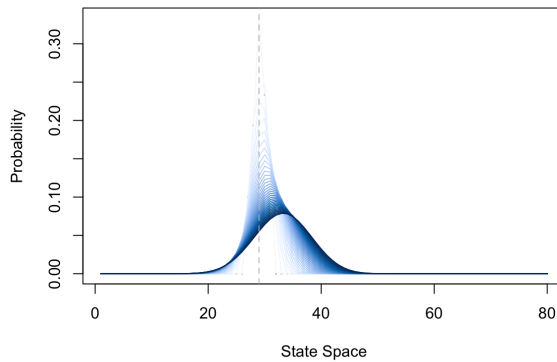
**Example 4.2.2.** By letting the state space  $K$  to grow, i.e. a high dimension simplex, we can visualize the evolution of the projected equation in a continuous like form and use computational methods to understand the projection method deeply. Consider again Example 4.1.1. Figure 4.4 shows the evaluation of the random walk model 4.1.1 with parameters  $\theta = 0.5, \phi = 0.3$  starting from a singular distribution. In this plot, the finite boundaries have little effect over the time scale of interest by design.

Observe that even though we are starting from a singular distribution (which is the case of the deterministic initial condition in stochastic filtering), the evolution approaches a Gaussian distribution, as is expected from the Brownian motion process. In addition, as expected, the process drifts due to differences between  $\theta$  and  $\phi$ .

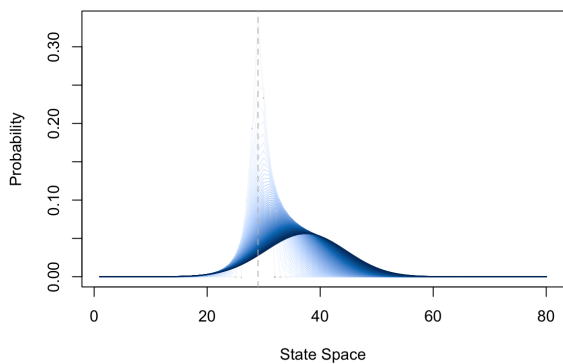
Figure 4.5 shows the same process on a larger simplex, with a longer simulation time. This is the recommended approach for simulation experiments to ensure that the boundary of the simplex does not impact the conclusions.



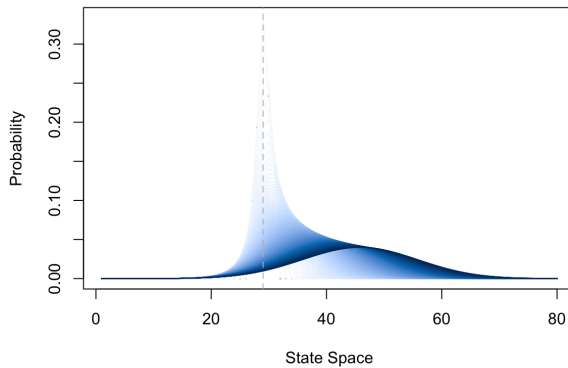
(a) 10 steps



(b) 35 steps

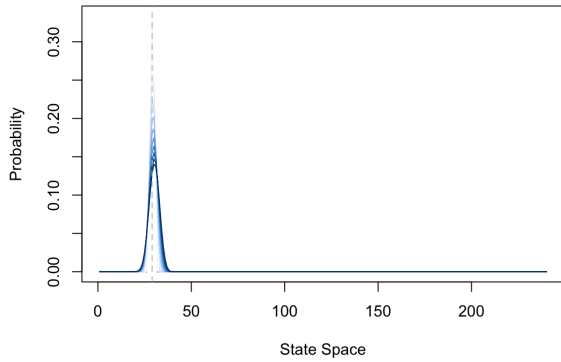


(c) 70 steps

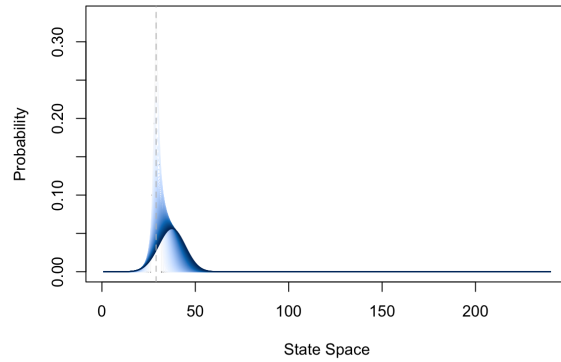


(d) 140 steps

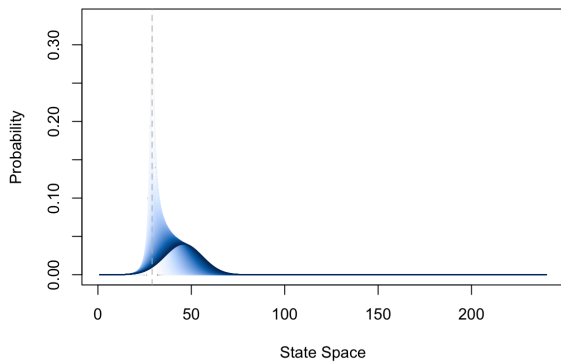
Figure 4.4: The flow of a non-stationary random walk with parameters  $\theta = 0.5, \phi = 0.3$  on the closed 80-simplex at different time steps. The vertical gray dashed line is marker for the singular initial distribution



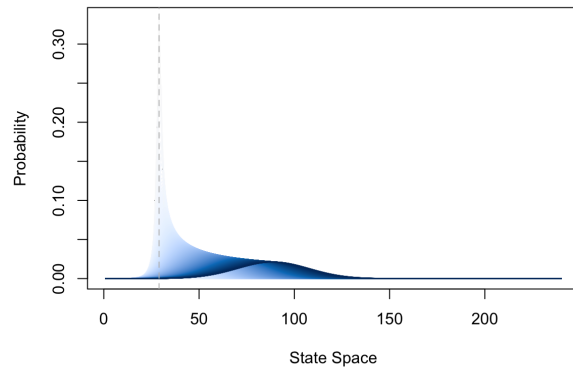
(a) 10 steps



(b) 35 steps



(c) 140 steps



(d) 500 steps

Figure 4.5: The flow of a non-stationary random walk with parameters  $\theta = 0.5, \phi = 0.3$  on the closed 240-simplex at different time steps. The vertical gray dashed line is marker for the singular initial distribution

### 4.2.3 Optimal projection of a stationary random walk

**Example 4.2.3.** To use the projection method, we need to select an exponential family in the closed simplex such that the unbiasedness condition 4.6 holds. In addition, for the projected evolution to provide an optimal (in the sense of Kullback-Leibler divergence) we need to select the sufficient statistics of the family based on the eigenfunctions of the operator generating the true solution.

Figure 4.6 shows the first two computed eigenfunctions of the operator in the random walk visualized earlier in Figure 4.5

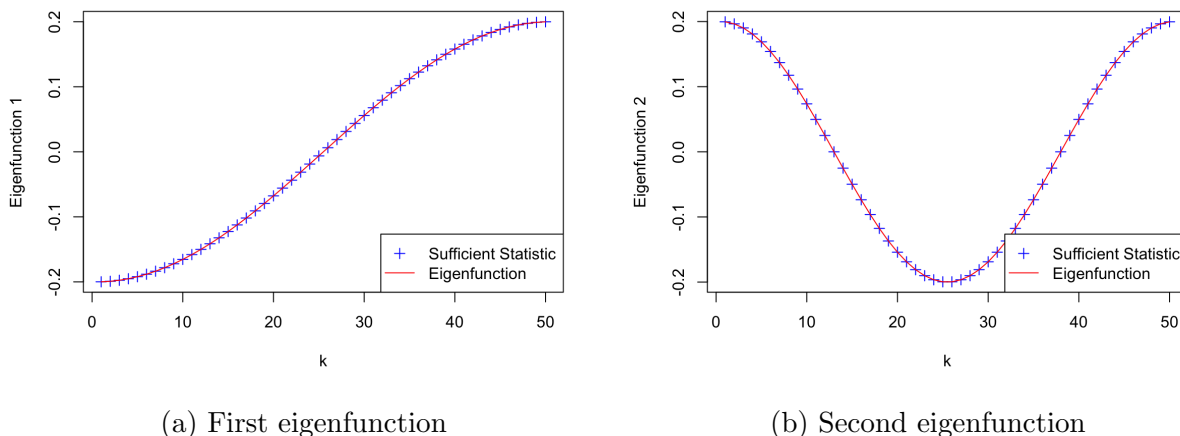
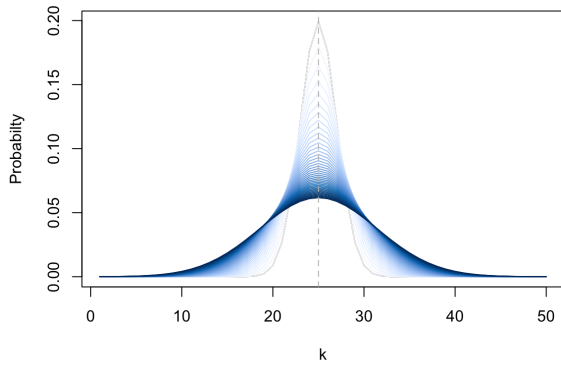


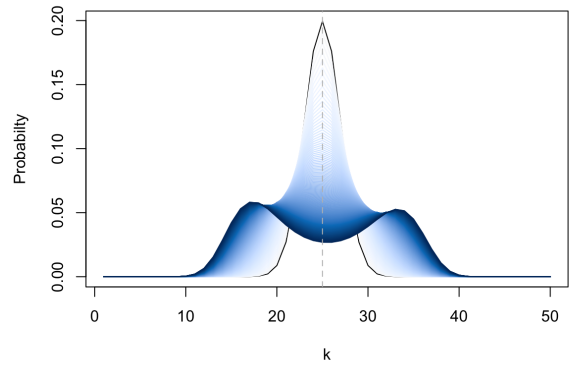
Figure 4.6: Two computed eigenfunctions of the operator associated with a stationary random walk model 4.1.1 with parameters  $\theta = 0.35$  and  $\phi = 0.35$

Using these two eigenfunctions as a sufficient statistic for a low-dimensional exponential family, we computed the flow for both the true solution and its projection on the 2-dimensional exponential family.

Figure 4.7 shows the flow of the true solution and its projection. Note the differences in the shape of the probability mass functions between the true solution and its projection. This emphasizes the comments in [18] regarding the fact that we can think of the projected solution as the evolution of other stochastic differential equations. However, the impact of this on applications is unclear.



(a) true solution



(b) projected solution

Figure 4.7: Flow of the mean of a stationary random walk on a 50-simplex with parameters  $\theta = \phi = 0.35$  and its projection on a 2-dimensional exponential family with a sufficient statistic matching the first two eigenfunctions of the infinitesimal generator of the random walk. The gray dashed vertical line highlights the position of the singular initial condition. Note the differences in the shape of the probability mass functions between the random walk, and its approximation.

To determine the quality of the projection, we computed the mean of the true solution and its projection on the lower-dimensional exponential family. Figure ?? shows the mean of the true solution and the mean of the projected solution. Note that the estimate is optimal, emphasizing the importance of decomposing the operator of the true evolution into its eigenfunctions to aid the choice of the lower-dimensional exponential family.

In addition, note the level of generality of the projection method. As long as the operator of the solution exists and is in the tangent bundle of our total space (which is always the case for discrete state spaces), the projection method could yield optimal results (in the sense of maximum likelihood or Kullback-Leibler divergence) if the choice of a lower-dimensional exponential family is aided by the eigenfunction decomposition of the operator.

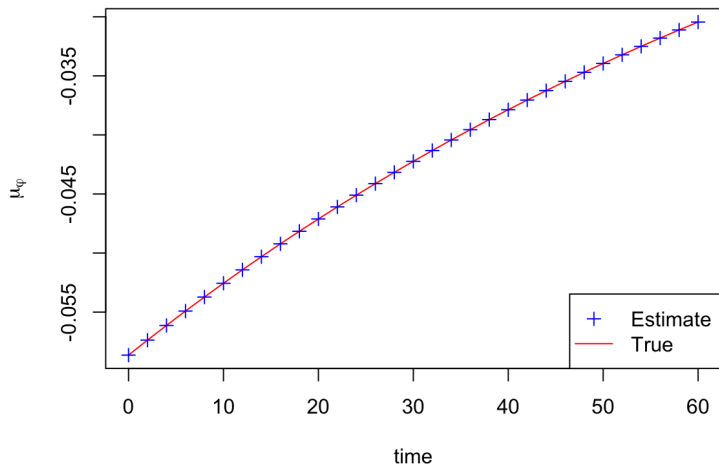
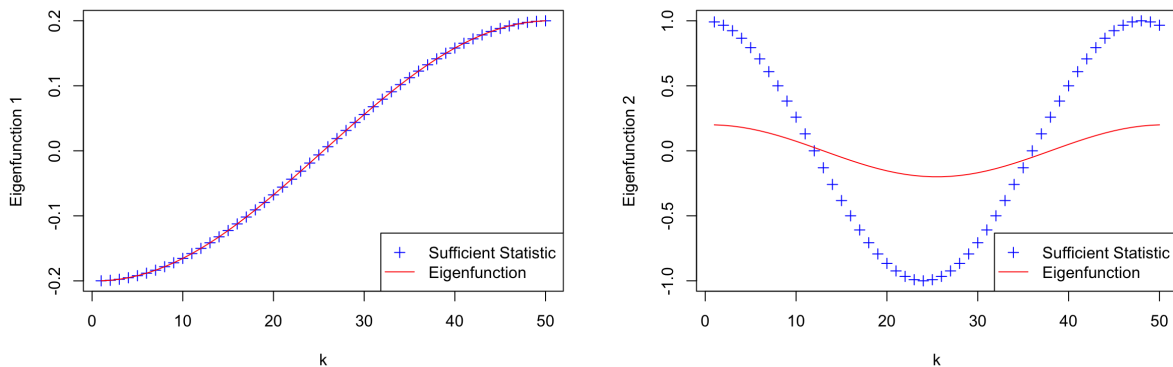


Figure 4.8: Computed mean of the true solution of a stationary random walk on a 50-simplex with parameters  $\theta = \phi = 0.35$  and its projection on a 2-dimensional exponential family with sufficient statistics matching the eigenfunctions of the infinitesimal generator of the random walk.

## 4.2.4 Suboptimal projection of a stationary random walk

**Example 4.2.4.** In order to demonstrate the impact of choosing the sufficient statistics not among the eigenfunctions of the infinitesimal generator, Figure 4.9 shows the eigenfunctions of the operator and the sufficient statistic of the low-dimensional manifold.

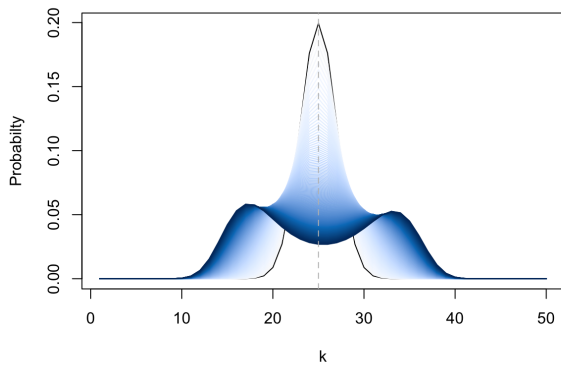


(a) First eigenfunction

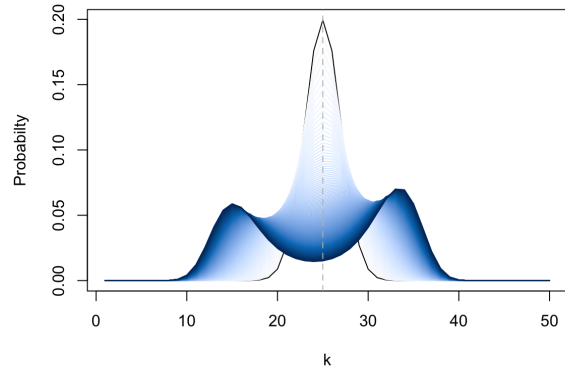
(b) Second eigenfunction

Figure 4.9: Two computed eigenfunctions of the operator associated with a stationary random walk model 4.1.1 and the two sufficient statistics of the low-dimensional exponential family

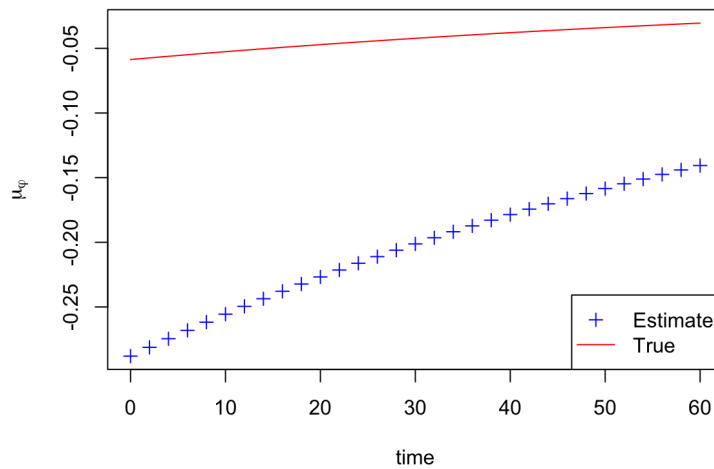
As a result of the nonoptimal choice of the finite-dimensional family, Figure 4.10 shows the difference in the projected mass functions between the two low-dimensional model choices and the corresponding change in estimates. Note that the change in the mass function is rather subtle, whereas the change in the estimate is relatively large.



(a) Optimal sufficient statistics



(b) Non-optimal sufficient statistics



(c) true mean vs estimate

Figure 4.10: Evolution of probability mass functions of a stationary random walk with parameters  $\theta = \phi = 0.35$  on 50-simplex and its projection two different low-dimensional manifolds. (a) shows the evolution of the mass function when the sufficient statistics are exactly the eigenfunctions of the infinitesimal generator. (b) shows the evolution of the mass function when the second sufficient statistic is a nonlinear function of the second eigenfunction. (c) Computed mean of the true solution and its approximation using a projection on a non-optimal 2-dimensional exponential family.

# Chapter 5

## Conclusion

To understand the projection method, we reviewed Pistone's extension to finite-dimensional information geometry and presented the material in the context of its relationship to the geometry of finite-dimensional models, making the material approachable for applications, and building a so-called bridge between these two geometries, which are treated separately in the literature.

We outlined the substantive differences between infinite- and finite-dimensional models that are related to projection methods. We emphasized the challenges in establishing a dually flat affine manifold structure due to the Banach manifold nature of Pistone's extensions impacting topology and convex duality, which trivially hold in finite-dimensional statistical manifolds. Also, we highlighted the need for separately considering the tangent and cotangent representations of the tangent bundle structures due to reflexivity issues in infinite dimensions, which is not required in finite-dimensional statistical models due to the fact that any finite vector space is isometric to its dual. We closed the review by discussing the difficulties in establishing a general split of the tangent spaces in infinite dimensions, which holds trivially in the finite-dimensional case.

In Chapter 4, we demonstrated that using standard arguments one can explore properties of the projection idea in the discrete-state case with a focus on visualizing solutions and building intuition. For example, we numerically demonstrated the impact of choosing the sufficient statistics among the left eigenfunctions of the infinitesimal generator, which is required for the optimality of the approximation with respect to the chosen family [23, Theorem 6.6, page 38].

In addition, we have established two novel results that have not been discussed in the literature. First, we qualitatively showed that near the boundary of a set of probability

measures, the state approximation breaks abruptly due to the fact that tangent vectors vanish at the boundary. Second, we have highlighted the rather subtle point of choosing a finite-dimensional family to project on that does not contain the initial condition of the original problem. As detailed in Chapter 4, this problem is rather subtle and depends on whether or not the projected evolution changed materially.

Some of the drawbacks of our method are that, as discussed in detail in Chapter 3, there are substantive differences between the finite- and infinite-dimensional settings. As a result, it might not be possible to study some of the consequences of extending the geometry to infinite dimensions. For example, issues related to the duality between the exponential and mixture flat manifolds in infinite dimensions cannot be studied in finite dimensions. In addition, problems regarding different possible splits of the tangent space needed to establish the orthogonal relationship between the tangent and cotangent spaces are better explored in infinite dimensions.

Another potential drawback is due to the computational demands needed in a high-dimensional simplex. In order to perform some of the computations required, such as computing the inverse of the Fisher matrix, or the left the eigenvectors of the infinitesimal generator, one must rely on numerical algorithms. These algorithms might suffer from numerical inaccuracies and/or instability as the dimension increases.

The projection method is a powerful tool for approximating optimal filters of stochastic processes in the general setting of stochastic filtering where having estimates that are optimal, causal, and online is a requirement. Even more, the optimality of the projection method in the infinite-dimensional case has been well established in the literature [9] relative to approximate filtering techniques such as Extended Kalman filter, and even Particle filter methods with the same number of parameters as the projection filter.

However, this optimality only holds if the finite-dimensional family is chosen such that the sufficient statistic is among the left eigenfunctions of the infinitesimal generator of the process (which depends on the test function). This limits the optimal use of the projection filter to situations where knowledge of the nature of the generator is available, and to a particular class of test functions.

Analyzing the eigenfunctions of the operator is a difficult problem on its own. In stochastic filtering, the problem is even harder since we are interested in the nature of the operator of the conditional process, not the operator associated with the signal. In addition, as we highlighted earlier, one needs to show that the operator exists in the tangent bundle structure of the infinite-dimensional exponential manifold, which is an open research problem.

It is important to note that even though the motivation for this dimensionality reduction

method is to solve stochastic filtering problems, the projection method, as noted by the authors in [23], cannot be applied directly to the nonlinear problem in the infinite-dimensional case without considering separately the different vector fields of the Kushner-Stratonovich equation, which ignores the total nature of the conditional distribution process. A rather interesting new line of research is to recast the SDE equation as a 2-jet whose projection has yet to be extended to infinite-dimensional statistical manifolds. For more details, see [9].

## 5.1 Future Directions

The following are a few inquiries that arose during our work in this thesis and may be worthwhile to investigate in further detail:

- Studying the conditional semi-group structure of the solution of stochastic filtering problems is important to develop an adequate theoretical basis for when it is justified to use the projection filter idea.
- The methods we developed in this thesis could be used to examine the convergence behavior of the approximation if the lower dimensional family is appropriately chosen. This convergence problem is still an open research problem.
- Use standard arguments to analyze the quality of the projection of a solution to a nonlinear stochastic filtering problem under different observational noise conditions. It has been noted that properties of the observational noise might impact the nature of the projection [23].
- Use standard methods we developed in Chapter 4 to study the projection of the jet formulation of a stochastic differential equation in the simplex to understand the properties of this jet projection method in comparison to other projection filters.

# References

- [1] Ethan Akin. *The Geometry of Population Genetics*, volume 31 of *Lecture Notes in Biomathematics*. Springer-Verlag, Berlin, 1946.
- [2] Shun-ichi Amari. Theory of information spaces - a geometrical foundation of the analysis of communication systems. *RAAG Memoirs*, 4, 01 1968.
- [3] Shun-ichi Amari. *Differential geometry in statistical inference*, volume 10. Institute of Mathematical Statistics, Hayward, Calif, 1987.
- [4] Shun-ichi Amari.  $\alpha$ -Divergence Is Unique, Belonging to Both  $f$ -Divergence and Bregman Divergence Classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.
- [5] Shun-ichi Amari. Information geometry and its applications. *Applied Mathematical Sciences*, 2016.
- [6] Erling Bernhard Andersen. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331):1248–1255, 1970.
- [7] John Armstrong and Damiano Brigo. Nonlinear filtering via stochastic PDE projection on mixture manifolds in L2 direct metric. *Mathematics of control, signals, and systems*, 28(1):1, 2016.
- [8] John Armstrong and Damiano Brigo. Stochastic filtering via L2 projection on mixture manifolds with computer algorithms and numerical examples. *Mathematics of Control, Signals and Systems*, 2016.
- [9] John Armstrong, Damiano Brigo, and Bernard Hanzon. Optimal projection filters with information geometry. *Information Geometry*, 7(Suppl 1):525–540, 2024.

- [10] Colin Atkinson and Ann F. S. Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 43(3):345–365, 1981.
- [11] Nihat. Ay, Jürgen. Jost, Hông Vân. Lê, and Lorenz. Schwachhöfer. *Information Geometry*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics, 64. Springer International Publishing, Cham, first edition, 2017.
- [12] Alan Bain and Dan Crisan. *Fundamentals of Stochastic Filtering*. Stochastic Modelling and Applied Probability. Springer New York, 2008.
- [13] Ole E. Barndorff-Nielsen. *Information and exponential families : in statistical theory*. Wiley series in probability and mathematical statistics. Wiley, Chichester, 1978.
- [14] Ole E. Barndorff-Nielsen. Parametric Statistical Models and Likelihood. *Lecture Notes in Statistics*, 1988.
- [15] Michèle Basseville. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.
- [16] Vaclav E. Benes. Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics An International Journal of Probability and Stochastic Processes*, 5:65–92, 1981.
- [17] Lev M. Brégman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. computational mathematics and mathematical physics*, 7(3):200–217, 1967.
- [18] Damiano Brigo. On the nice behaviour of the gaussian projection filter with small observation noise. *Systems & control letters*, 26(5):363–370, 1995.
- [19] Damiano Brigo. *Filtering by Projection on the Manifold of Exponential Densities*. Phd-thesis - research and graduation internal, Vrije Universiteit Amsterdam, 1996. Naam instelling promotie: Vrije Universiteit Naam instelling onderzoek: Vrije Universiteit.
- [20] Damiano Brigo. New results on the gaussian projection filter with small observation noise. *Systems & control letters*, 28(5):273–279, 1996.
- [21] Damiano Brigo and Bernard Hanzon. On some filtering problems arising in mathematical finance. *Insurance, mathematics & economics*, 22(1):53–64, 1998.

- [22] Damiano Brigo, Bernard Hanzon, and François Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 5(3):495–534, 1999.
- [23] Damiano Brigo and Giovanni Pistone. Projection based dimensionality reduction for measure valued evolution equations in statistical manifolds, 2016.
- [24] Roger W. Brockett. Remarks on finite dimensional nonlinear estimation. In *Analyse des systèmes*, number 75-76 in Astérisque, pages 47–55. Société mathématique de France, 1980.
- [25] Roger W. Brockett. The early days of geometric nonlinear control. *Automatica*, 50(9):2203–2224, 2014.
- [26] Richard S. Bucy. Optimum finite time filters for a special nonstationary class of inputs. 1959.
- [27] Alberto Cena and Giovanni Pistone. Exponential statistical manifold. *Annals of the Institute of Statistical Mathematics*, 59(1):27–56, 2007.
- [28] Nikolai N. Cencov. *Statistical decision rules and optimal inference (in Russian)*, volume 053 of *Translations of mathematical monographs*. American Mathematical Society,, Providence, R.I, 1982.
- [29] Mireille Chaleyat-Maurel and Dominique Michel. Des resultats de non existence de filtre de dimension finie. *Stochastics*, 13(1-2):83–102, 1984.
- [30] Goffredo Chirco and Giovanni Pistone. Dually affine information geometry modeled on a banach space. 2024.
- [31] David R. Cox and Hilton D. Miller. *The Theory of Stochastic Processes*. Wiley, 2017.
- [32] Dan Crisan. The stochastic filtering problem: a brief historical account. *Journal of Applied Probability*, 51(A):13–22, 2014.
- [33] Frank Critchley and Paul Marriott. Computational information geometry in statistics: theory and practice. *Entropy*, 16(5):2454–2471, 2014.
- [34] Imre Csiszár. Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

- [35] Georges Darmois. Sur les lois de probabilité à estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85, 1935.
- [36] A. P. Dawid. Further Comments on Some Comments on a Paper by Bradley Efron. *The Annals of Statistics*, 5(6):1249, 1977.
- [37] Bradley Efron. *Exponential Families in Theory and Practice*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2022.
- [38] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, New York; Toronto, 1986.
- [39] Ronald A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 14th ed., revised and enlarged. edition, 1970.
- [40] Kenji Fukumizu. Exponential manifold by reproducing kernel hilbert spaces. In *Algebraic and Geometric Methods in Statistics*, pages 291–306. Cambridge University Press, 2009.
- [41] Crispin W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004.
- [42] Ronald K. Gettoor. On the construction of kernels. *Séminaire de probabilités de Strasbourg*, 9:443–463, 1975.
- [43] Paolo Gibilisco and Giovanni Pistone. Connections on non-parametric statistical manifolds by orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 01(02):325–347, 1998.
- [44] Matheus R. Grasselli. Dual connections in nonparametric classical information geometry. *Annals of the Institute for Statistical Mathematics*, 62, 04 2001.
- [45] Matheus R. Grasselli. Dual connections in nonparametric classical information geometry. *Annals of the Institute of Statistical Mathematics*, 62(5):873–896, 2010.
- [46] Robert M. Gray. *Entropy and Information Theory*. Springer New York, New York, NY, 1990.

- [47] Bernard Hanzon. A differential-geometric approach to approximate nonlinear filtering. In *Geometrization of Statistical Theory*, C.T.J. Dodson, Editor, pages 219–223, University of Lancaster, 1987. ULMD Publications.
- [48] Michiel Hazewinkel, Steven I. Marcus, and Hector J. Sussmann. Nonexistence of finite-dimensional filters for conditional statistics of the cubic sensor problem. *Systems & Control Letters*, 3(6):331–340, 1983.
- [49] Alan T. James. The Variance Information Manifold and the Functions on It. In Paruchuri R. Krishnaiah, editor, *Multivariate Analysis–III*, pages 157–169. Academic Press, 1973.
- [50] Jürgen Jost and Xianqing Li-Jost. *Calculus of variations*. Cambridge studies in advanced mathematics ; 64. Cambridge University Press, Cambridge, UK ;, 1998.
- [51] Rudolph E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.
- [52] Rudolph E. Kalman and Richard S. Bucy. New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering*, 83(1):95–108, 03 1961.
- [53] Robert E. Kass. *The Riemannian Structure of Model Spaces: A Geometrical Approach to Inference*. University of Chicago, 1983.
- [54] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of differential geometry*, volume 15 of *Interscience tracts in pure and applied mathematics*. Interscience Publishers, New York, 1963.
- [55] Bernard O. Koopman. On Distributions Admitting a Sufficient Statistic. *Transactions of the American Mathematical Society*, 39(3):399–409, 1936.
- [56] Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [57] Harold J Kushner. On the dynamical equations of conditional probability density functions, with applications to optimal stochastic control theory. *Journal of Mathematical Analysis and Applications*, 8(2):332–344, 1963.
- [58] Harold J. Kushner. On the differential equations satisfied by conditional probability densities of markov processes, with applications. *Journal of the Society for Industrial and Applied Mathematics Series A Control*, 2(1):106–119, 1964.

- [59] Harold J. Kushner. Dynamical equations for optimal nonlinear filtering. *Journal of Differential Equations*, 3(2):179–190, 1967.
- [60] Anna Kutschireiter, Simone C. Surace, and Jean-Pascal Pfister. The Hitchhiker’s guide to nonlinear filtering. *Journal of Mathematical Psychology*, 94:102307, 2020.
- [61] Serge Lang. Introduction to Differentiable Manifolds. *Universitext*, 2002.
- [62] Steffen Lauritzen. *Differential geometry in statistical inference*, volume 10 of *Lecture notes-monograph series*. Institute of Mathematical Statistics, Hayward, Calif, 1987.
- [63] Paul Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.
- [64] Sanjoy K. Mitter. Filtering and stochastic control: a historical perspective. *IEEE Control Systems Magazine*, 16(3):67–76, 1996.
- [65] Michael K. Murray and John W. Rice. *Differential Geometry and Statistics*. Chapman & Hall, London; New York, 1993.
- [66] Nigel J. Newton. A class of non-parametric statistical manifolds modelled on Sobolev space. *Information Geometry*, 2(2):283–312, 2019.
- [67] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edition, January 2014.
- [68] Giovanni Pistone. Examples of the application of nonparametric information geometry to statistical physics. *Entropy*, 15(10):4042–4065, September 2013.
- [69] Giovanni Pistone. A lecture about the use of orlicz spaces in information geometry, 2021.
- [70] Giovanni Pistone and Maria P. Rogantin. The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 5(4):721–760, 1999.
- [71] Giovanni Pistone and Carlo Sempì. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of statistics*, 23(5):1543–1561, 1995.
- [72] Edwin J. G. Pitman. Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(4):567–579, 1936.

- [73] David Pollard. *Probability tools, tricks, and miracles*. 2023. <http://www.stat.yale.edu/~pollard/Books/Pttm/Orlicz.pdf> (visited 2024-07-23).
- [74] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Bulletin of Calcutta Mathematical Society*, volume 37, pages 81–91, 1945.
- [75] Ralph T. Rockafellar. *Convex analysis*. Princeton landmarks in mathematics and physics. Princeton University Press, Princeton, N.J, 1997-1970.
- [76] Claude E. Shannon. *A mathematical theory of communication*. American Telephone and Telegraph Company, New York, 1948.
- [77] Albert N. Shiryaev. Addendum: On stochastic equations in the theory of conditional markov processes. *Theory of Probability & Its Applications*, 12(2):342–342, 1967.
- [78] Lene T. Skovgaard. A Riemannian Geometry of the Multivariate Normal Model. *Scandinavian Journal of Statistics*, 11(4):211–223, 1984.
- [79] Ruslan L. Stratonovich. Conditional Markov Processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.
- [80] Daniel W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*. Grundlehren der mathematischen Wissenschaften ; 233. Springer Verlag, Berlin ;, 1979.
- [81] David V. Widder. *The Laplace transform*. Princeton mathematical series ; 6. Princeton University Press, Princeton, 1941.
- [82] Norbert Wiener. *Cybernetics: or Control and communication in the animal and the machine*. Technology Press, Cambridge, Mass, 1948.
- [83] Norbert Wiener. Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications. In *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*, 1949.
- [84] David Williams. *Probability with martingales*. Cambridge mathematical textbooks. Cambridge University Press, Cambridge, 1991.
- [85] David Williams and L.C.G. Rogers. *Diffusions, Markov Processes, and Martingales*, volume 2 of *Cambridge mathematical library*. Springer Verlag, Berlin; New York, 1979.