Arbitrage-free regularization, geometric learning, and non-Euclidean filtering in finance

Anastasis Kratsios

A Thesis
in the Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy (Mathematics) at
Concordia University
Montréal, Québec, Canada

July 2018

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared
By: Anastasis Kratsios
Entitled: Arbitrage-free regularization, geometric learning, and non-Euclidean filtering in finance

and submitted in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair
Dr. Greg Leblanc

______________________________ External Examiner
Dr. Tyrone Duncan

______________________________ External to Program
Dr. Arash Mohammadi

______________________________ Examiner
Dr. Frederic Godin

______________________________ Examiner
Dr. Arusharka Sen

______________________________ Thesis Supervisor
Dr. Cody Hyndman

______________________________ Thesis Supervisor
Dr. Alina Stancu

Approved by ______________________________
Dr. Arusharka Sen, Graduate Program Director

August 27, 2018 ______________________________
Dr. André Roy, Dean, Faculty of Arts and Science

ABSTRACT

Arbitrage-free regularization, geometric learning, and non-Euclidean filtering in finance

Anastasis Kratsios Ph.D.
Concordia University, 2018

This thesis brings together elements of differential geometry, machine learning, and pathwise stochastic analysis to answer problems in mathematical finance. The overarching theme is the development of new stochastic machine learning algorithms which incorporate arbitrage-free and geometric features into their estimation procedures in order to give more accurate forecasts and preserve the geometric and financial structure in the data.

This thesis is divided into three parts. The first part introduces the non-Euclidean upgrading (NEU) meta-algorithm which builds the universal reconfiguration and universal approximation properties into any objective learning algorithm. These properties state that a procedure can reproduce any dataset exactly and approximate any function to arbitrary precision, respectively. This is done through an unsupervised learning procedure which identifies a geometry optimizing the relationship between a dataset and the objective learning algorithm used to explain it. The effectiveness of this procedure is supported both theoretically and numerically. The numerical implementations find that NEU-ordinary least squares outperforms leading regularized regression algorithms and that NEU-PCA explains more variance with one NEU-principal component than PCA does with four classical principal components.

The second part of the thesis introduces a computationally efficient characterization of intrinsic conditional expectation for Cartan-Hadamard manifolds. This alternative characterization provides an explicit way of computing non-Euclidean conditional expectation by using geometric transformations of specific Euclidean conditional expectations. This reduces many non-convex intrinsic estimation problems to transformations of well-studied Euclidean conditional expectations. As a consequence, computationally tractable non-Euclidean filtering equations are derived and used to successfully forecast efficient portfolios by exploiting their geometry.

The third and final part of this thesis introduces a flexible modeling framework and a stochastic learning methodology for incorporating arbitrage-free features into many asset price models. The procedure works by minimally deforming the structure of a model until the objective measure acts as a martingale measure for that model. Reformulations of classical no-arbitrage results such as NFLVR, the minimal martingale measure, and the arbitrage-free Nelson-Siegel correction of the Nelson-Siegel model are all derived as solutions to specific arbitrage-free regularization problems. The flexibility and generality of this framework allows classical no-arbitrage pricing theory to be extended to models that admit arbitrage opportunities but are deformable into arbitrage-free models. Numerical implications are investigated in each of the three parts making up this thesis.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my co-advisors Prof. Cody Hyndman and Prof. Alina Stancu for their guidance, support, and direction through this journey. Their guidance has helped direct me in my studies, as a mathematician, and generally has helped me grow as a person.

Secondly, I would like to thank my parents and familly for their support throughout all the years and the unending love, help, and courage they have given me.

I would also like to thank my friends, colleagues, peers, and the faculty and staff members of the department of Mathematics and Statistics, each of which has made this experience incredibly pleasant, fulfilling, and truly allowed me to feel at home in the department.

Finally, I would like to express my thanks to Behnoosh for all her love and support throughout this entire express. None of this would have been even remotely possible without any you. Thank you.

Contents

List of Figures

List of Tables

1. Introduction

The application of machine learning to mathematical finance is a new and active research area. Innovative applications of deep learning to optimal hedging problems in [16, 34], price formation in [92], sparse estimation of diffusion process parameters in [25], forecasting of electricity prices in [20] amongst others, have been recently explored.

Many of these machine learning algorithms are both static and Euclidean making them unsuited to the dynamic nature of financial markets and unable to incorporate many of the geometric features present in financial data. An objective of this thesis is to develop new learning algorithms which are suited to the dynamic nature of financial data. To this end, arbitrage-free regularization is a non-anticipative learning procedure beginning with an empirical factor model for the price of a risky asset and progressively deforming it until the real-world measure becomes a local martingale.

Measure changes will be formulated as particular types of model deformations, allowing for the reformulation of many classical results from arbitrage-pricing theory can be reformulated as the existence, uniqueness, and solution to specific arbitrage-free regularization problems. Arbitrage-free regularization extends past measure change induced deformations and is used to extend arbitrage-pricing theoretic results to models which are not arbitrage-free but are deformable into arbitrage-free models.

In [38, 39, 21] it is shown that a wide range of factor models for the term-structure of interest rates admit arbitrage. Arbitrage-free regularization can be interpreted as optimally correcting these models while retaining the maximum amount of structure of the initial model after the deformation is complete. The arbitrage-free correction of the Nelson-Siegel term structure model of [21] is a model specific case of this general procedure.

The second direction of this thesis focuses on the development of new learning algorithms which have universal approximation properties built into them and are able to incorporate non-Euclidean features into their estimates. Non-Euclidean geometry occurs naturally in certain problems in finance. For example, in [10], short-rate models consistent with finite-dimensional smooth manifolds were characterized. In [59], highly accurate stochastic volatility model estimation methods are derived using heat kernel expansions of the Riemannian metric associated with a stochastic volatility model. In [39] arbitrage-free factor models for interest rates have been characterized using geometric methods.

Inspired by these results, the second direction of this thesis is concerned with the introduction and development of a meta-algorithm which learns and incorporates an optimal geometry into any learning algorithm. It is shown that this geometry improves the in-sample and out-of-sample forecasts of any learning algorithm. In numerical experiments, the performance of basic algorithms improved with the incorporation of this optimal geometry are seen to outperform their sophisticated counterparts while still retaining their simplicity. The second class of results in this direction develops a rigorous theory of non-Euclidean conditional expectation capable of naturally incorporating non-Euclidean features and is successfully used to obtain computationally tractable dynamics for the non-Euclidean conditional expectation. This differs from the current literature on the subject which considers Euclidean conditional expectations of functionals of a non-Euclidean signal and observations process. Implementation of these techniques successfully predicts efficient portfolio weights more accurately than the competitive methods such as basis-function regression methods and penalized regression methods, such as the LASSO of [99].

The thesis is organized as follows. Chapter 2 introduces the NEU meta-algorithm for incorporating non-Euclidean features and the universal reconfiguration property into any objective learning algorithm. Chapter 3 introduces non-Euclidean conditional expectation, proves its existence and alternative characterizations and uses this alternative characterization to solve for computable non-Euclidean filtering equations. Chapter 4 introduces the theory of arbitrage-free regularization, relates many classical arbitrage-pricing theory results to particular formulations of arbitrage-free regularization problems and introduces a meta-algorithm for incorporating arbitrage-free information into a wide range of estimation procedures. Chapter 5 summarizes the contributions made in the thesis and outlines future directions of research.

2. The NEU Meta-Algorithm for Geometric Learning with Applications in Finance

We introduce a meta-algorithm, called non-Euclidean upgrading (NEU), which learns algorithm-specific geometries to improve the training and validation set performance of a wide class of learning algorithms. Our approach is based on iteratively performing local reconfigurations of the space in which the data lie. These reconfigurations build universal approximation and universal reconfiguration properties into the new algorithm being learned. This allows any set of features to be learned by the new algorithm to arbitrary precision. The training and validation set performance of NEU is investigated through implementations predicting the relationship between select stock prices as well as finding low-dimensional representations of the German Bond yield curve.

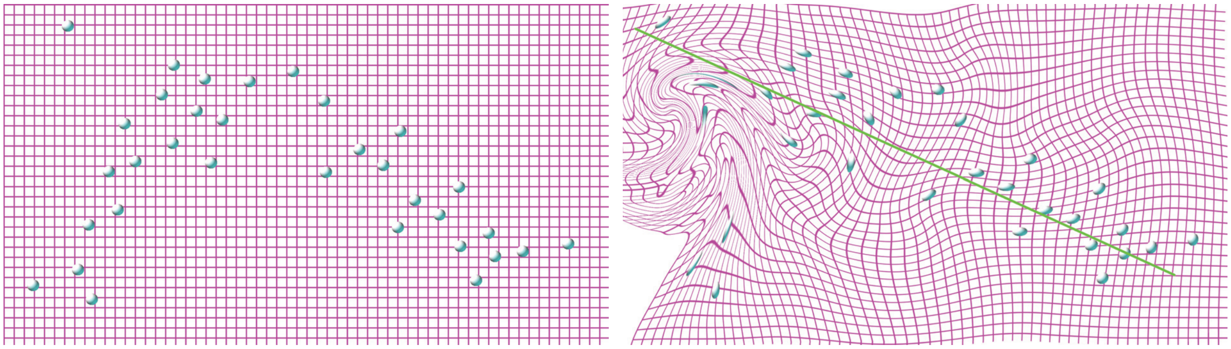
2.1 Introduction

Many statistical and learning algorithms such as ordinary linear regression and PCA admit linear algebraic formulations making them quick to execute. Their inability to capture non-linear features has motivated several non-linear generalizations. Non-linear generalizations of linear models require alternative, computationally costly, estimation procedures. Generalized additive models (GAM) and artificial neural networks (ANN) are examples of non-linear generalizations of linear regression that come with a significant increase in computational cost (see [55, Chapters 9 and 11] for a discussion of these methods).

Patterns in the data are typically interpreted as a function relating explanatory inputs to the observations which they explain. Alternatively, a pattern can be interpreted as the positioning of points in space. Since a function's graph is a specific set of points in space, interpreting a pattern as a configuration of points in space is more general than interpreting it as a function. The non-Euclidean Upgrading (NEU) methodology introduced in this chapter can learn any configuration of data. As a consequence, two versions of the universal approximation property (see [23] for details) of ANNs is also recovered.

Non-Euclidean Upgrading (NEU) is a meta-algorithm. Meta-algorithms are algorithms whose inputs and outputs are other algorithms. For example, the Boosting meta-algorithm of [86] efficiently combines learning algorithms to build a more accurate new learning algorithm. Bagging, as introduced in [14], is another meta-algorithm which generates bootstrapped samples from a given dataset, performs the input algorithm on those bootstrapped samples, and aggregates each of the predictions into a lower-variance estimate. NEU is also a meta-algorithm which inputs a learning algorithm and a dataset, and outputs a new algorithm with the universal approximation property built into it. Applying NEU to simple linear algorithms produces algorithms which are interpretable, have a low computational burden, and can predict any pattern to arbitrary precision once trained.

NEU works by first segmenting the input data into training and validation components, then performing local perturbations on the space on which the data is defined, executing the learning algorithm on the perturbed training and validation data sets, and evaluating if the validation set performance has increased. The procedure continues iteratively, stopping once the validation set performance begins to drop.



(a) Non-Linear Configuration of Euclidean Data.

(b) Linear Configuration of Non-Euclidean Data.

Figure 2.1: Visualization of Reconfiguration of the Data.

Figure 2.1 illustrates how perturbing $\mathbb{R}^2$ reconfigures the given dataset and allows for a linear regression to explain a non-linear relationship. After linear regression is performed, the transformations to $\mathbb{R}^2$ are inverted and the linear predictor becomes non-linear. This illustration is analogous to the non-Euclidean regression proposed in [41] with the central difference being that our methodology learns the geometry of the problem whereas the algorithm in [41] relies on a prespecified geometry.

Applying NEU to principal component analysis (PCA) generates an analogue of the principal geodesic analysis of [42] where the geometry is learned from the data. Applying NEU to the unscented Kalman filtering algorithm of [57] or to the geometric GARCH framework of [52] produces analogues of those algorithms but without a prespecified geometry. There are many other potential applications of NEU in statistics and machine learning.

We consider two examples from finance. The first example considers the use of principal component analysis (PCA) on German bond data. Using NEU on PCA shows that one NEU-principal component performs better than 4 standard principal components.

The second example from finance considers the relationship between Apple stock price and the stock prices of companies related to Apple. Using NEU on linear regression provides better out-of-sample predictions than the LASSO, Ridge regression, and non-linear extensions of Elastic-Net (ENET) procedures. While we consider only two examples from finance to illustrate NEU, the generality and flexibility should allow for similar performance gains in other areas of financial statistics and machine learning.

The remainder of this chapter is organized as follows. Section 2.2 introduces the mathematical framework for non-Euclidean upgrading, the main results regarding the technique's flexibility, and predictive performance enhancement are proven. Section 2.3 investigates the empirical performance of non-Euclidean upgrading on the two examples from finance. The relationship between Apple stock price and the stock price of related companies can be better explained training sets and validation sets using non-Euclidean upgraded regression. Parallels are drawn to the non-Euclidean generalizations of regression and principal geodesic analysis developed in [41] and [42], respectively. We adjoin an appendix with two sections, the first lists the regularity assumptions made and the second contains certain technical proofs.

2.2 Non-Euclidean Upgrading

This section introduces and develops the NEU meta-algorithm. Reconfigurations are first introduced and a universal approximation property is proven. The NEU meta-algorithm is then introduced and its performance gain property is proven.

For the remainder of this chapter, a dataset will be comprised of training and validation sets. The training set will be denoted by $X^I = \left\{X_1^I, \dots, X_{N_I}^I\right\}$ and the validation set will be denoted by $X^O = \left\{X_1^O, \dots, X_{N_O}^O\right\}$, where N_I and N_O are non-negative integers and $N_I \geq 1$.

Reconfigurations perturbing the dataset are smooth maps from $\mathbb{R}^D$ back into itself, smooth autodiffeomorphisms, which satisfy certain local properties. These are defined as follows.

Definition 2.2.1 (Reconfiguration Map) *Let Θ be an open subset of $\mathbb{R}^m$. A reconfiguration on $\mathbb{R}^D$ is a map*

$$\begin{aligned}\mathfrak{k} :&\mathbb{R}^D \times \Theta \to \mathbb{R}^D,\\ &(x,\theta) \mapsto \mathfrak{k}(x|\theta)\end{aligned}$$

satisfying the following properties:

(i) ***Invertiblility:*** *For every $\theta \in \Theta$, the map $f_\theta(x) \triangleq \mathfrak{k}(x|\theta)$ is a bijection,*

(ii) ***Smoothness:*** *For every $\theta \in \Theta$, the maps $f_\theta(x)$, and f_θ^{-1} are continuously differentiable,*

(iii) ***Smooth Parametrization:*** *For every x in $\mathbb{R}^D$, the map $\theta \mapsto \mathfrak{r}(x|\theta)$ is continuously differentiable,*

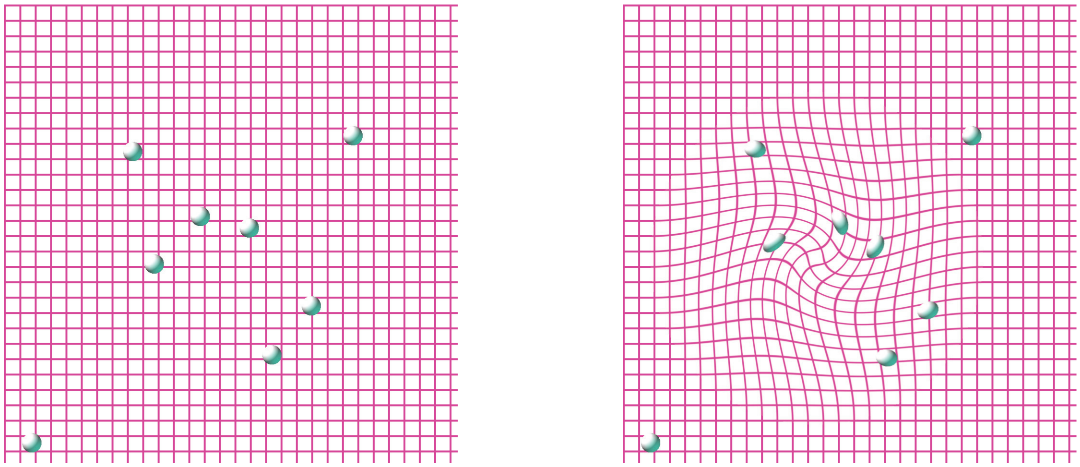
(iv) ***Local Transience:*** *For every x, y, z in $\mathbb{R}^D$ with $d(x, y) < d(x, z)$, there exists $\theta \in \Theta$ such that*

$$\begin{aligned}\mathfrak{r}\,(x|\theta) =& y\\ \mathfrak{r}\,(z|\theta) =& z,\end{aligned}$$

where $d(\cdot, \cdot)$ is the Euclidean distance on $\mathbb{R}^D$.

(v) ***Identity:*** *The subset $\Theta_0 \triangleq \left\{\theta \in \Theta : \mathfrak{r}\,(x|\theta) = x,\ \forall x \in \mathbb{R}^D\right\}$ of Θ is non-empty.*

The central example of a reconfiguration map, is a rapidly decaying rotation concentrated on a disc. These rotations slow exponentially as the boundary of the disc is approached. Beyond the disc's boundary the reconfiguration map becomes the identity transformation. Rapidly decaying rotations are illustrated by Figure 2.2.



(a) Data in Euclidean Space.

(b) A Rapidly Decaying Rotation.

Figure 2.2: Visualization of Rapidly Decaying Rotations.

Definition 2.2.2 (Rapidly Decaying-Rotations) *Let $\mathfrak{so}(D)$ denote the set of $D \times D$ skew-symmetric matrices and set $\Theta \triangleq \mathbb{R}^D \times (0, \infty) \times \mathfrak{so}(D)$. A rapidly decaying rotation is the map $\mathfrak{r}$ defined by*

$$\begin{aligned}\mathfrak{r} : \mathbb{R}^D \times \Theta \to \mathbb{R}^D&\\ \mathfrak{r}\,(x|(c, \sigma, X)) \mapsto exp\,(\psi(\|x - c\|; \sigma)X)\,(x - c) + c&,\end{aligned} \tag{2.1}$$

where ψ is the Gaussian bump-function supported on the unit sphere of radius σ centered at the point $c \in \mathbb{R}^D$, defined by

$$\psi(x; \sigma) \triangleq \begin{cases}\exp\left(\frac{-\sigma}{\sigma - \|x\|^2}\right) : & \|x\| < \sigma\\ 0 : & else,\end{cases} \tag{2.2}$$

and exp is the matrix exponential map.

Proposition 2.2.3. Rapidly decaying rotations are reconfiguration maps. Moreover, the inverse of $\mathfrak{x}(\cdot|(c,\sigma,X))$ is

$$\Phi(x|c,\sigma,X) = exp\left(-X\psi(\|y\|\,;\sigma)\right)(x-c)+c,$$

where y is the image of x under $\Phi\left(\cdot|c,\sigma,X\right)$.

Proof. The proof is deferred to the appendix □

Remark 2.2.4 (Geometric Interpretation)**.** The rapidly decaying rotations are interpolations between a rotation and the identity map interior to the disc of radius σ, centered at c. However, the interpolation does not take place in $\mathbb{R}^D$, but instead happens within the lie algebra $\mathfrak{so}(D)$ lying tangential to the space of all generalized rotation matrices $SO(D)$. This ensures that the map is invertible for all possible parameter choices.

Definition 2.2.5 (Planar Micro-Bumps) *A planar micro-bump on* $\mathbb{R}^2$*, is the map* $\mathfrak{x}$ *defined by*

$$\begin{gathered}\mathfrak{x}:\mathbb{R}^2\times\Theta\to\mathbb{R}^2\\ \mathfrak{x}\left((x_1,x_2)|(c,\sigma,X)\right)\mapsto x+\psi(\|x-c\|;\sigma)X,\end{gathered}\tag{2.3}$$

where $\Theta=\mathbb{R}^2\times[0,\infty)\times\mathbb{R}$*.*

Proposition 2.2.6. Planar micro-bumps are reconfigurations maps on $\mathbb{R}^2$.

Proof. The proof is deferred to the appendix □

Data points are deemed poorly placed if moving them increases the validation set performance of a learning algorithm. Iteratively applying reconfiguration maps allows poorly placed data-points to be moved to locations which increase an algorithm's validation set performance. The local transience property of reconfiguration maps, Definition 2.2.1 (iv), makes it possible to only move poorly placed data-points while leaving the others fixed. The procedure is summarized as follows.

Definition 2.2.7 (Reconfiguration) *Let* $\mathscr{M}$ *be a smooth sub-manifold of* $\mathbb{R}^D$*, which is diffeomorphic*[1] *to* $\mathbb{R}^D$*,* Φ *be a diffeomorphism*[2] *from* $\mathscr{M}$ *onto* $\mathbb{R}^D$*, let* $\mathfrak{x}$ *be a reconfiguration map on* $\mathbb{R}^D$*, and let* $\theta_0,\dots,\theta_N$ *be in* Θ *with* $\theta_0\in\Theta_0$*. Here* Θ_0 *is as in definition 2.2.1(v). A reconfiguration* $\mathbb{X}$*, is a map from* $\mathscr{M}$ *to* $\mathscr{M}$ *defined by*

$$\mathbb{X}\left(x|\theta_1,\dots,\theta_N;\Phi\right)\triangleq\Phi\left(X^{(N)}(x)\right)$$

where

$$\begin{aligned}X^{(i)}(x)&\triangleq\mathfrak{x}\left(X^{(i-1)}(x)|\theta_i\right);\quad i=1,\dots,k\\ X^{(0)}(x)&\triangleq\mathfrak{x}\left(x|\theta_0\right).\end{aligned}$$

[1] The Whitney embedding theorem implies that any smooth manifold is a smooth subset of a Euclidean space. In this chapter, a map will be quantified as being smooth if it is once continuously differentiable.

[2] A diffeomorphism is a bijection which is smooth and has a smooth inverse.

Reconfiguring a dataset on $\mathbb{R}^D$ maps it into new coordinates for the input D-variables. These coordinates may not be directly interpretable, therefore after performing the learning algorithm and obtaining an estimate in the new coordinate system the reconfiguration must be inverted. This inverse procedure is called deconfiguration.

Definition 2.2.8 (Deconfiguration) *Let $\mathbb{X}$ be a reconfiguration of $\mathscr{M}$. The deconfiguration of $\mathbb{X}$ is the map denoted by $\mathbb{X}^{-1}$ defined as*

$$\begin{aligned}\mathbb{X}^{-1}\left(x|\theta_1,\dots,\theta_N;\Phi\right) \triangleq& \Phi^{-1}\left(X^{(N)}(x)\right)\\ X^{(i)}(x) \triangleq& \mathfrak{r}^{-1}\left(X^{(i-1)}(x)|\theta_{N-1}\right); \quad i=1,\dots,N\\ X^{(0)}(x) \triangleq& \mathfrak{r}^{-1}\left(x|\theta_0\right).\end{aligned}$$

The universal approximation property of neural networks states that certain neural networks can approximate any function to arbitrary precision (see [23]). The first analogous property for reconfiguration states that any dataset can be transformed into any other dataset of equal size.

Theorem 2.2.9 (Universal Reconfiguration Property)**.** Assume that $D > 1$ and $\mathfrak{r}$ be a reconfiguration map on $\mathbb{R}^D$, and Φ be a diffeomorphism from $\mathscr{M}$ onto $\mathbb{R}^D$. Let $X \triangleq \{X_i\}_{i=1}^N$ and $\tilde{X} \triangleq \left\{\tilde{X}_i\right\}_{i=1}^N$ be subsets of $\mathscr{M}$. There exists a positive integer K, and $\theta_1,\dots,\theta_K$ in Θ for which

$$\mathbb{X}\left(X_i|\theta_1,\dots,\theta_K;\Phi\right) = \tilde{X}_i,$$

for every i in $\{1,\dots,N\}$.

Proof. The proof is deferred to the appendix. □

The universal reconfiguration property implies the following analogues to the universal approximation property of neural networks of [72]. The first captures general functions on a more restricted domain and the second captures a smaller class of functions on a larger domain.

Corollary 2.2.10 (Universal Approximation Property)**.** Let D_1, D_2 be positive integers, K be a subset of $\mathbb{R}^{D_1}$ and f, g be Borel-functions from K to $\mathbb{R}^{D_2}$. If K is diffeomorphic to $\mathbb{R}^{D_1}$, then for every countable subset Q of $\mathscr{M}$, probability measure $\mathbb{P}$ supported on Q, and every $n \in \mathbb{N}$, there exists $\theta_1^n,\dots,\theta_{N_n}^n \in \Theta$ such that for every $\epsilon > 0$ there exists a Borel-subset K_ϵ of Q satisfying

1. $\sup_{x\in K_\epsilon}\left\|f(x) - p\circ\mathbb{X}\left((x,g(x))\,|\theta_1^n,\dots,\theta_{N_n}^n\right)\right\| < \frac{1}{n}$,
2. $\mathbb{P}(Q - K_\epsilon) < \epsilon$.

Here p is the second canonical projection[3] of $\mathbb{R}^{D_1+D_2}$ onto $\mathbb{R}^{D_2}$. In the limiting case where $\epsilon = 0$, the convergence of $p\circ\mathbb{X}\left((x,g(x))\,|\theta_1^n,\dots,\theta_{N_n}^n\right)$ to $f(x)$ on Q is point-wise.

[3] The second canonical projection of the product space $X \times Y$ takes a pair (x, y) to y, see [73] for details.

Proof. The proof will be deferred to the appendix. □

Non-Euclidean upgrading uses reconfigurations to improve a class of learning algorithms which we call objective learning algorithms. These are discussed in the next section.

The learning algorithms we consider in this chapter optimize both the training set and validation set loss functions. Regularized regression, PCA, k-means, neural networks, Bayesian classifiers, support vector machines, and stochastic filters are all examples of objective learning algorithms.

Objective learning algorithms associate to every pair of training and validation sets of a given size, a pair of training set and validation set loss-functions as well as a pattern function linking the parameters being optimized to the prediction they can make. This formalization requires the definition of the set of all possible learning algorithms for a fixed set of hyper-parameters Γ and parameter to prediction function $\phi : \mathbb{R}^d \times \mathbb{R}^D \rightarrow \mathbb{R}^{D\times k}$. Here D is the dimension of the space in which the data-points lie, d is the dimension of the explanatory parameters, and k is the number of D-dimensional points out-putted by the algorithm.

For example, for a 1 factor PCA, $k = 1$ $d = D$ and for a two factor PCA $k = 2$ and $d = D$. In the case of linear regression, the regression weights are scalars therefore $d = 1$. If there is no intercept then $k = 1$ and if there is an intercept $k = 2$, in this formulation D is the number of columns of the design matrix.

Let N_I be a positive integer and N_O be a non-negative integer. Define $\Lambda_{\Gamma,\phi}^{N_I,N_O}$ to be the set of all pairs of maps $\left(\mathscr{L}_I^{N_I}, \mathscr{L}_O^{N_O}\right)$ such that

(i) The map $\mathscr{L}_I^{N_I} : \mathbb{R}^d \times C^\infty(\mathbb{R}^d \times \mathbb{R}^D, \mathbb{R}^{Dk}) \times \Gamma \times \mathbb{R}^{DN_I} \rightarrow [-\infty, \infty]$,

(ii) The map $\mathscr{L}_O^{N_O} : \mathbb{R}^d \times C^\infty(\mathbb{R}^d \times \mathbb{R}^D, \mathbb{R}^{Dk}) \times \Gamma \times \mathbb{R}^{DN_O} \rightarrow [-\infty, \infty]$,

(iii) Regularity condition 2.5.1 holds.

The function $\phi(\beta|\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^{D\times k}$ represents the estimated pattern, parameterized by β. The parameter β lies in the space $\mathbb{R}^d$ and is to be chosen by optimizing training set and validation set loss functions. $\mathscr{L}_I^{N_I}$ is the training set loss function on a dataset of size N_I and $\mathscr{L}_O^{N_O}$ is the out-of-sample loss function on a dataset of size N_O. The space of all learning algorithms for a specific pattern function ϕ is $\Lambda_{\Gamma,\phi} \triangleq \bigcup_{(N_I,N_O)\in\mathbb{N}^2} \Lambda_{\Gamma,\phi}^{N_I,N_O}$.

Definition 2.2.11 (Objective Learning Algorithm) *An objective learning algorithm is a map*

$$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi) : \bigcup_{(N_I,N_O)\in\mathbb{N}^2} \mathbb{R}^{D\times N_I} \times \mathbb{R}^{D\times N_O} \rightarrow \Lambda_{\Gamma,\phi},$$

$$\left(X^I, X^O\right) \mapsto \Lambda_{\Gamma,\phi}^{\frac{Dim(X^I)}{D},\frac{Dim(X^O)}{D}},$$

where the pair of an training set and a validation set (X^I, X^O) *are viewed as elements of* $\mathbb{R}^{D\times N^I}$ $\mathbb{R}^{D\times N^O}$*, and where* $Dim(\cdot)$ *is the non-negative integer-valued function mapping a point in Euclidean space to* $(\cdot)$*.*

Remark 2.2.12. Given a dataset consisting of N data-points, the regression analysis loss function is

$$\sum_{i=1}^{N}(\beta^i X^i - Y^i), \tag{2.4}$$

where $\{X^i\}_{i=1}^N$ are the data-points and $\{Y^i\}_{i=1}^N$ are the responses. Incorporating an additional data-point X^{101} and an additional response Y^{N+1} into the regression analysis changes the loss function of Equation (2.4) to

$$\sum_{i=1}^{N+1}(\beta^i X^i - Y^i). \tag{2.5}$$

Both Equations (2.4) and (2.5) are a 1-dimensional regression problem but technically are defined by different loss functions. Definition 2.2.11 overcomes the oddity of having a learning algorithm differ depending on the size of the dataset, by defining an objective learning algorithm as a map associating the size of a dataset to the corresponding loss function; which is what we do in inadvertently.

Principal component analysis and regression analysis are objective learning algorithms. This is illustrated by the following two examples.

Example 2.2.13 (Regression as an Objective Learning Algorithm)**.** Let $a < b$ be real numbers and $\{f_i(x)\}_{i=1}^d$ be a continuously differentiable linearly independent set of functions in $L^2([a, b])$. Non-linear regression is an objective learning algorithm which is represented by

(i) $\phi(\beta|x) = \sum_{i=1}^d \beta_i f_i(x^i)$,

(ii) $\mathscr{L}_I^N\left(\beta, \phi(\beta|\cdot) \mid X_1^I, \dots, X_N^I\right) \triangleq \sum_{i=1}^N \left(y_i - \phi(\beta|X_i^I)\right)^2$,

(iii) $\mathscr{L}_O^N\left(\beta, \phi(\beta|\cdot) \mid X_1^O, \dots, X_N^O\right) \triangleq \sum_{i=1}^N \left(y_i - \phi(\beta|X_i^O)\right)^2$,

(iv) $\Gamma = \{0\}$,

where x^i is the i^{th} component of the D-dimensional vector x and where y_i is the i^{th} observed data-point. Typically, the out-of-sample dataset is always taken to be empty unless a regularization or sparsity constraint is imposed.

By adding a penalty term, such as the ℓ^1 norm, to the training set and validation set loss functions and expanding the hyperparameter set Γ accordingly, most regularized regression problems, such the LASSO of [99], are seen to be objective learning algorithms.

Example 2.2.14 (PCA as an Objective Learning Algorithm)**.** Calculating the first principal component of a dataset's empirical covariance matrix Q is an objective learning algorithm. Here $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ are represented by

(i) $\phi(\beta|x) = x\beta^T$,

(ii) $\mathscr{L}_I^N\left(\beta, \phi(\beta|X^I) \mid X_1^I, \ldots, X_N^I\right) \triangleq -\left\{\frac{\beta^T \tilde{X}_I^T \tilde{X}_I \beta}{\beta^T \beta}\right\}$,

(iii) $\mathscr{L}_O^N\left(\beta, \phi(\beta|X^O) \mid X_1^O, \ldots, X_N^O\right) \triangleq -\left\{\frac{\beta^T \tilde{X}_O^T \tilde{X}_O \beta}{\beta^T \beta}\right\}$,

(iv) $\Gamma = \{0\}$,

where $\tilde{X}_I$ and $\tilde{X}_O$ are the training and validation sets X^I and X^O, viewed as matrices but with their column-wise means removed. Typically, the out-of-sample dataset is always taken to be empty. The higher principal components, as well as sparse principal components, can also be represented analogously as an objective learning algorithm.

The optimal evaluation of a learning algorithm, is a map taking a learning algorithm and a dataset to an optimized pattern. The optimal evaluation is only well-defined on datasets which admit a unique optimizer. This set of regular datasets, called the regular domain of definition of the learning algorithm, is defined as follows.

Definition 2.2.15 (Regular Domain of Definition) *Let $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ be a learning algorithm. The regular domain of definition of $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$, denoted by Dom$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$, is the set of all pairs of data points (X^I, X^O) in*

$$\bigcup_{(N_I, N_O) \in \mathbb{N}^2} \mathbb{R}^{D \cdot N_I} \times \mathbb{R}^{D \cdot N_O}$$

satisfying the regularity condition 2.5.2.

The map associating a dataset and an objective learning algorithm to the pattern best describing it is now defined.

Definition 2.2.16 (Optimal Evaluation) *Given an objective learning algorithm, $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ its optimal evaluation is the output of the function taking as input a pair training and validation sets in Dom$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ and returning the optimal parameter $\beta(\hat{\gamma})$ defined by*

$$\begin{aligned}
\hat{\gamma} &\in \underset{\gamma \in \Gamma}{\operatorname{arginf}} \mathscr{L}_O^{N_O}\left(\beta(\gamma), \phi(\beta(\gamma)|X^O); \gamma \mid (X_1^O, \ldots, X_{N_O}^O)\right) \\
\beta(\gamma) &= \underset{\beta \in \mathbb{R}^D}{\operatorname{arginf}} \mathscr{L}_I^{N_I}\left(\beta, \phi(\beta|X^I); \gamma \mid (X_1^I, \ldots, X_{N_I}^I)\right).
\end{aligned}$$

Remark 2.2.17. The optimal evaluation takes an objective learning algorithm and a dataset and returns the optimizer minimizing the loss function defined by the dataset. For example, in a LASSO regression the optimal evaluation returns the parameters of the line of best fit relating the explanatory variables to the responses, with the tuning parameter is optimized according to the validation set.

The requirement that the dataset be in the regular domain of definition of the learning algorithm means that the optimal evaluation is a well-defined function. For example the points $\{(1,1),(-1,1),(1,-1),(-1,-1)\}$ do not have a single line of best fit describing their relationship therefore the optimal evaluation of the regression problem is not defined on that dataset.

As in [55] the performance of a learning algorithm is defined as the negative of its loss function evaluated at the optimal value. The definition of performance of training and validation set performance of an objective learning algorithm is defined in an analogous manner.

Definition 2.2.18 (Performance) *Let $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ be a learning algorithm. The training set performance of $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ is the function, denoted by $\mathscr{P}^I(\mathscr{L}_I, \mathscr{L}_O)$, taking a dataset (X^I, X^O) in Dom$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ to the extended real number*

$$\mathscr{P}^I(\mathscr{L}_I, \mathscr{L}_O)\left(\tilde{X}^I, \tilde{X}^O\right) \triangleq -\mathscr{L}_I^{N_I}\left(\beta(\hat{\gamma}), \phi(\beta(\hat{\gamma})); \hat{\gamma} \mid (X_1^I, \dots, X_{N_I}^I)\right).$$

The validation set performance of $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ is the function, denoted by $\mathscr{P}^O(\mathscr{L}_I, \mathscr{L}_O)$, taking a dataset (X^I, X^O) in Dom$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ to the extended real number

$$\mathscr{P}^O(\mathscr{L}_I, \mathscr{L}_O)\left(\tilde{X}^I, \tilde{X}^O\right) \triangleq -\mathscr{L}_O^{N_O}\left(\beta(\hat{\gamma}), \phi(\beta(\hat{\gamma})); \hat{\gamma} \mid (X_1^O, \dots, X_{N_O}^O)\right).$$

Remark 2.2.19. The performance is the negative of the loss function evaluated at its optimal evaluation. It provides a measure of how well an objective learning algorithm can explain a given dataset.

A dataset in *Dom*$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ is said to maximize the in (resp. out-of) sample performance of $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ if there is no other dataset in *Dom*$(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ having the same number of training and validation data points and a higher validation set performance.

The main result can now be stated. If the data is in the regular domain of definition of a learning algorithm, and is not already in an optimal position, then there is a reconfiguration which increases the performance of that algorithm. An example of optimally positioned data for linear regression is data that is perfectly explained by a line both on the training and validation sets. In this extreme case, it is natural to expect that no improvement can be made to linear regression.

Theorem 2.2.20 (Performance Gain)**.** Let $D > 1$ and $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ be an objective learning algorithm. For every pair of integers N_I, N_O and every (X_I, X_O) in $Dom(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$, there exists $\theta_1^{N_I,N_O}, \ldots, \theta_K^{N_I,N_O}$ in Θ such that

$$\mathscr{P}^O\left(\mathscr{L}_I, \mathscr{L}_O\right)\left(\tilde{X}^I, \tilde{X}^O\right) \geq \mathscr{P}^O\left(\mathscr{L}_I, \mathscr{L}_O\right)\left(X^I, X^O\right), \tag{2.6}$$

$$\mathscr{P}^I\left(\mathscr{L}_I, \mathscr{L}_O\right)\left(\tilde{X}^I, \tilde{X}^O\right) \geq \mathscr{P}^I\left(\mathscr{L}_I, \mathscr{L}_O\right)\left(X^I, X^O\right), \tag{2.7}$$

where the reconfigured datasets $\tilde{X}^I$ and $\tilde{X}^O$ are defined as

$$\tilde{X}_i^I \triangleq \mathbb{X}\left(X_i^I | \theta_1^{N_I,N_O}, \ldots, \theta_K^{N_I,N_O}\right),$$
$$\tilde{X}_i^O \triangleq \mathbb{X}\left(X_i^O | \theta_1^{N_I,N_O}, \ldots, \theta_K^{N_I,N_O}\right).$$

The inequality in equation (2.7) (resp. equation (2.6)) is strict if (X_I, X_O) does not maximize $\mathscr{P}^O\left(\mathscr{L}_I, \mathscr{L}_O\right)$ (resp. $\mathscr{P}^I\left(\mathscr{L}_I, \mathscr{L}_O\right)$).

Proof. Without loss of generality assume that (X^I, X^O) does not maximize $\mathscr{P}^O\left(\mathscr{L}_I, \mathscr{L}_O\right)$, with the proof of the statement for $\mathscr{P}^I\left(\mathscr{L}_I, \mathscr{L}_O\right)$ being identical. Therefore there is $(\tilde{X}^I, \tilde{X}^O)$ in $Dom(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ which has a higher value of $\mathscr{P}^O\left(\mathscr{L}_I, \mathscr{L}_O\right)$ and has the same number of training and validation data-points.

Therefore, by the universal reconfiguration property of Theorem 2.2.9, there exists $\theta_1^{N_I,N_O}, \ldots, \theta_K^{N_I,N_O}$ such that $\tilde{X}_i = \mathbb{X}\left(X_i | \theta_1^{N_I,N_O}, \ldots, \theta_K^{N_I,N_O}\right)$. □

Theorem 2.2.20 guarantees that there exists a reconfiguration of the data which improves an algorithm's training set and validation set performance. The NEU meta-algorithm is a procedure which learns the reconfiguration of the space ensuring that the training and validation sets are positioned in a way which reduces the training set and validation set loss functions. This is formalized by the meta-algorithms illustrated by Figure 2.3 and made explicit in meta-algorithm 2.2.21.

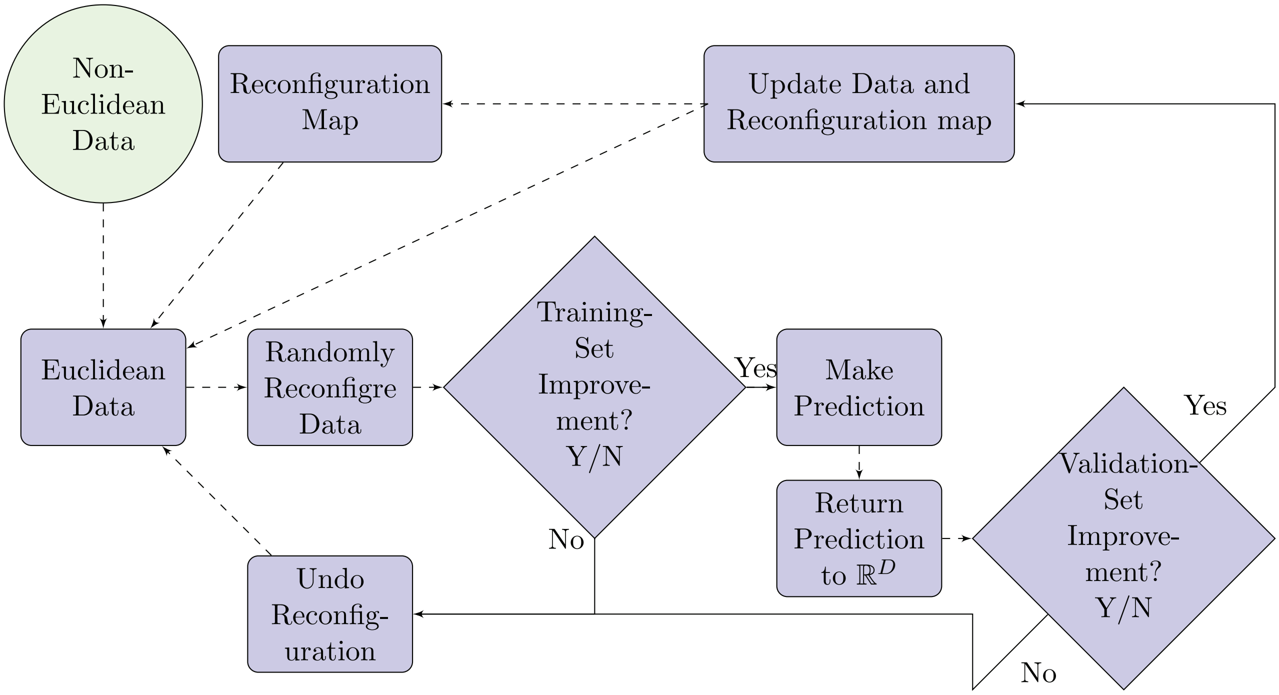

Non-Euclidean Data
Reconfiguration Map
Update Data and Reconfiguration map
Euclidean Data
Randomly Reconfigre Data
Training-Set Improvement? Y/N
Yes
Make Prediction
Yes
Return Prediction to $\mathbb{R}^D$
Validation-Set Improvement? Y/N
No
Undo Reconfig-uration
No

Figure 2.3: Work-flow of Reconfiguration Learning Phase of Non-Euclidean Upgrading

Meta-Algorithm 2.2.21 (Non-Euclidean Upgrading)**.** The inputs of the non-Euclidean upgrading algorithm are a diffeomorphism $\Phi : \mathscr{M} \to \mathbb{R}^D$, an objective learning algorithm $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$, a pair of training-set and validation-set data-points $\left(\{X_i^I\}_{i=1}^{N_I}, \{X_i^I\}_{i=1}^{N_O}\right)$ in $\mathscr{M}$ satisfying regularity condition 2.5.3, $\mathfrak{X}$ a reconfiguration map, $\theta_0 \in \Theta_0$, $\epsilon \in (0,1]$, and a positive integer N. Non-Euclidean upgrading takes these inputs and returns the following algorithms as its output

1. **Learning Reconfiguration:** Define a reconfiguration $\mathbb{X}$ through the following procedure,

   (a) Define the data-points $X_i^{(0)} \triangleq \Phi(p_i)$,

   (b) $\theta^{(0)} \triangleq \theta_0$,

   (c) For integers n between $0 < n \leq N$:

      (i) Define the tentative optimal evaluation $\beta^{\uparrow}(\hat{\gamma})$ to be

$$\hat{\gamma} \in \underset{\gamma \in \Gamma}{\operatorname{arginf}} \mathscr{L}_O^{N_O}\left(\beta^{\uparrow}(\gamma), \phi(\beta^{\uparrow}(\gamma)); \gamma \mid \left(\mathbb{X}\left(X_1^O|\theta\right), \dots, \mathbb{X}\left(X_{N_O}^O|\theta\right)\right)\right)$$

$$\left(\beta^{\uparrow}(\gamma), \theta(\gamma)\right) = \underset{\beta \in \mathbb{R}^D, \theta \in \Theta}{\operatorname{arginf}} \mathscr{L}_I^{N_I}\left(\beta, \phi(\beta); \gamma \mid \left(\mathbb{X}\left(X_1^I|\theta\right), \dots, \mathbb{X}\left(X_{N_I}^I|\theta\right)\right)\right),$$

      (ii) Define the tentative performance measurement, $\mathscr{P}^n\left(\mathscr{L}_I, \mathscr{L}_O\right)$ to be

$$\mathscr{P}^n\left(\mathscr{L}_I, \mathscr{L}_O\right) \triangleq \mathscr{P}^O\left(\mathscr{L}_I, \mathscr{L}_O\right)\left(\mathbb{X}\left(X_1^O|\theta\right), \dots, \mathbb{X}\left(X_{N_O}^O|\theta\right)\right),$$

```
            (iii) if $\mathscr{P}^n(\mathscr{L}_I, \mathscr{L}_O) > \mathscr{P}^{n-1}(\mathscr{L}_I, \mathscr{L}_O)$ then
                    define $\theta_1 \triangleq \theta(\hat{\gamma})$.
                  else
                    define $\theta_n \triangleq \theta_0$.
            (iv) Define the updated data $X_i^{(n)} \triangleq \mathfrak{x}\left(X_i^{(n-1)}|\theta_n\right)$,
        (d) Stop when $\frac{\mathscr{P}(\mathscr{L}_I, \mathscr{L}_O)}{\mathscr{P}^n(\mathscr{L}_I, \mathscr{L}_O)} < \epsilon$ or when $n = N$,
        (e) Define $X_i \triangleq X_i^{(n)}$,
        (f) Define the reconfiguration $\mathbb{X} \triangleq \mathbb{X}(\cdot|\theta_1, \dots, \theta_N; \Phi)$,
    2. Perform Algorithm: Perform $(\mathscr{L}_I, \mathscr{L}_O, \Gamma, \phi)$ on the data $(\mathbb{X}(X_i))_{i=1}^k$ and obtain
       the optimal evaluation $\hat{X}$,
    3. Deconfigure Prediction: Returns the values:
        (a) Prediction: $\mathbb{X}^{-1} \circ \phi(\hat{\beta}|\mathbb{X}(x))$,
        (b) Performance Gain: $\frac{\mathscr{P}^N(\mathscr{L}_I, \mathscr{L}_O)}{\mathscr{P}^0(\mathscr{L}_I, \mathscr{L}_O)}$,
        (c) Parameter Estimates: $\hat{\beta}$.
```

Geometric and algebraic interpretations of NEU are discussed now, as well as connection to other geometric algorithms.

Remark 2.2.22 (Geometric Interpretation)**.** The reconfiguration $\mathbb{X}$ is a diffeomorphism of $\mathbb{R}^D$ back into itself. The pullback of the Euclidean metric d_E along $\mathbb{X}$, denoted by $\mathbb{X}^\star(d_E)$, makes

$$(\mathbb{R}^D, \mathbb{X}^\star(d_E)) = (\mathbb{X}(\mathbb{R}^D), \mathbb{X}^\star(d_E)),$$

into a Riemannian manifold. The minimal distance curves in $(\mathbb{R}^D, \mathbb{X}^\star(d))$ are mapped to straight lines, through $\mathbb{X}$. Therefore the non-Euclidean algorithms in [42, 41, 52, 57] are all interpretable as parametric analogues to the NEU of PCA, regression, or Kalman filtering, but where the geometry is prespecified and not learned in an unsupervised manner.

Remark 2.2.23 (Algebraic Interpretation)**.** A smooth automorphism of $\mathbb{R}^D$ is a smooth bijection from $\mathbb{R}^D$ back onto itself, whose inverse is itself smooth. NEU is therefore a computational method for learning an autodiffeomorphisms which optimizes the validation-set and training set performance of a learning algorithm given a dataset.

In the next section, the numerical performance of NEU is investigated. The NEU algorithm is to improve regression analysis and principal component analysis and the resulting NEU-OLS and NEU-PCA algorithms are applied to financial time-series data.

2.3 Numerical Implementation of NEU-OLS and NEU-PCA

We begin by investigating the empirical performance of non-Euclidean upgrading. The first two implementations focus on real datasets and the second uses simulated data. The first two use the rapidly decreasing rotations to reconfigure the data whereas the last example uses micro-bumps since the data lies in $\mathbb{R}^2$.

The performance of the NEU meta-algorithm will be investigated both in the regression and dimensionality reduction settings on financial datasets beginning with a regression analysis study.

Example 2.3.1 (Regression Analysis: Apple Stock Tracker)**.** Predicting the relationship between the price of a set of assets is central to many trading strategies. For example, strategies that rely on illiquid assets may create a portfolio comprised entirely of liquid assets, which tracks the illiquid asset's movements. Since that is a particular application of tracking portfolios, in this example, the technique is demonstrated using liquid stocks. The target stock price will be denoted by S_t and the prices of the assets making the tracking portfolio will be denoted by $S_t^1, \dots, S_t^N$.

In this example, S_t will be the price of apple stock, and $S_t^1, \dots, S^N$ will be the stock prices for IBM, Google, Cisco Systems Inc., Microsoft Corporation, Acacia Communications Inc., NXP Semiconductors NV, Qualcomm, Analog Devices Inc., Glu Mobile Inc., Jabil Inc., Micron, and STMicroelectronics NV. These portfolio is chosen as being comprised of the stock of major companies in the same same industry as well as major companies making up apple's supply chain (see [1] for a discussion on apple's supply chain and [96] for a discussion of the 10 tech companies with the largest market capitalization).

A tracking portfolio consisting of these assets is built by minimizing the ordinary least-squares loss function on the training dataset

$$\sum_{i=1}^{N}\left(\left[\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}}\right]+\sum_{j=1}^{d}\beta_t^j\left[\frac{S_{t_i}^j-S_{t_{i-1}}^j}{S_{t_{i-1}}^j}\right]\right)^2,$$

where N is the number of data points and d is the number of assets used to track the Apple stock price. For illustrative and comparative purposes, the LASSO of [99], the Ridge (or Tykhonov regularization) regression of [100], the Elastic-Net regularization (ENET) of [104], and the NEU-OLS are compared.

The ENET selects the optimal regression weights by minimizing the loss function ENET Opt. Power denotes the solution to

$$\sum_{i=1}^{N}\left(\left[\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}}\right]+\sum_{j=1}^{d}\beta_t^j\left[\frac{S_{t_i}^j-S_{t_{i-1}}^j}{S_{t_{i-1}}^j}\right]\right)^2+\lambda\left[(1-\alpha)\sum_{j=1}^{d}\left|\beta^j\right|+\alpha\sum_{j=1}^{d}(\beta^j)^2\right],$$

with α, λ selected by sequential-validation. The LASSO is the special case where α is fixed to 0 and Ridge regression is the special case where $\alpha = 1$. The penalty

$$\lambda \left[(1-\alpha) \sum_{j=1}^{d} \left|\beta^j\right| + \alpha \sum_{j=1}^{d} (\beta^j)^2 \right]$$

reduces the number of explanatory parameters in a model by forcing the regression weights towards 0, thereby forcing the most significant parameters to only be fit. The meta-parameter λ controls the strength of this sparsity penalty, $\alpha \in [0,1]$ controls the aggressiveness of the variable-selection process, with $\alpha = 0$ giving a more aggressive choice and $\alpha = 1$ towards a non-aggressive penalty. ENET, LASSO, and Ridge regression are interpreted in [103] as robust regression problems where the regression problem is optimized against varying types of shocks in the data, or alternatively these can be interpreted as in [99, 105] as modifications of the regression problem that are able to detect and converge to the true set of explanatory variables, under linear and Gaussian noise assumptions.

In this example, 2 years of adjusted stock prices are used to compute the weights, ending on July 25^{th} 2018. The modeling assumption that the data does not follow a constant pattern throughout time is made and the data is broken up into rolling windows. Regression weights are dynamically updated on each window as is standard in practice (for example see [37, 11, 98]). In order to extract meaningful weights $\beta_t^1, \dots, \beta_t^N$, the time-series must be shown to be co-integrated. The Dickey-Fuller, unit root test is performed on the returns of the adjusted stock price time-series and the null-hypothesis that there exists a unit root is rejected with a p-value of less than .01 and Dickey-Fuller statistic -2.8453, therefore the β_t can meaningfully be computed from the adjusted stock price's returns using regression methods (see [81] for more details on co-integrated time-series).

	Mean	95 L	95 U	99 L	99 U
OLS	4.185	4.038	4.385	4.017	4.448
Ridge	-0.831	-0.916	-0.715	-0.928	-0.678
LASSO	0.581	0.568	0.599	0.566	0.604
ENET	0.526	0.519	0.535	0.518	0.538
NEU-OLS	0.204	0.202	0.208	0.202	0.209

Table 2.1: Mean Aggregate Training Errors.

Each window is sequentially divided into a training, a validation, and a test set. Each of the training sets consists of 200 observations, the validation sets consist of 2 weeks, and the test sets consists of the last week of each moving window. The proportions invested in each asset, denoted are the regression weights on that window, and are recalibrate on each window using each of the stocks' returns. The mean training, validation, and test errors aggregated across each windows are reported in the Tables 2.1, 2.3, and 2.2, respectively. The optimal parameters for the Ridge, LASSO, ELASTIC-NET, and NEU-OLS are re-calibrated on every window using sequential validation. The optimization of the parameters defining the reconfiguration of the data on were performed by alternating between stochastic gradient descent and randomized searches of the parameter space.

	Mean	95%L	95%U	99%L	99%U
OLS	4.217	4.214	4.222	4.214	4.224
Ridge	-0.853	-0.946	-0.726	-0.959	-0.686
LASSO	0.582	0.573	0.594	0.572	0.598
ENET	0.525	0.518	0.534	0.517	0.537
NEU-OLS	0.204	0.203	0.206	0.203	0.206

Table 2.2: Mean Aggregate Testing Errors.

	Mean	95%L	95%U	99%L	99%U
OLS	4.202	4.058	4.397	4.038	4.458
Ridge	-0.845	-0.928	-0.734	-0.939	-0.699
LASSO	0.581	0.571	0.594	0.569	0.598
ENET	0.525	0.521	0.530	0.520	0.531
NEU-OLS	0.204	0.203	0.206	0.202	0.206

Table 2.3: Mean Aggregate Validation Errors.

As expected the OLS performs worst and the ENET performs best amongst the benchmark regression methods. All the methods, except the Ridge regression are conservative and under-estimate the price of apple stock. The NEU-OLS has the lowest error in the training, validation, and test sets across every window. Moreover, it has the tightest confidence intervals. Therefore the NEU-OLS performs achieves a lower bias as well as a lower variance.

Algorithm	OLS	NEU-OLS	Ridge	LASSO	ENET
Run Time (sec)	0.01	104.02	0.02	0.02	0.07
$\frac{\text{Run Time}}{\text{Run Time OLS}}$	1	12,980.03	2.74	2.57	9.11

Table 2.4: Runtime Comparison.

The NEU-OLS does have its own drawbacks, namely computational time. Once the reconfiguration of the data is learned the OLS algorithm can be run directly on the reconfigured dataset making NEU-OLS and OLS just as fast. However on the first run, when the reconfiguration is being learned the NEU-OLS is significantly slower than the other methods compared within this chapter.

Table 2.4 reports the run-times of performing the OLS, NEU-OLS, Ridge regression, LASSO, and ENET algorithms on the dataset considered in this example using an Intel(R) Core(TM) i5-6200U CPU at 2.30GHz, with 7844MB available RAM machine running 18.04 LTS version of the Ubuntu Linux distribution.

We conclude that after learning the NEU-OLS has the lowest prediction error amongst the regression methods considered in this example and its execution speed is just as fast as OLS after the reconfiguration has been learned. However, on the first run when the reconfiguration is being learned NEU-OLS is notably slower than the other methods. Therefore, NEU-OLS may be the best of these options when speed is not a large factor, but it may not be ideal for setting when the runtime of an algorithm is a determining factor, such as for live high-frequency trading.

Example 2.3.2 (Dimensionality Reduction: German-Bond Yield Curve)**.** Principal component analysis (PCA) is a non-parametric technique which converts correlated data $\{x_1, \dots, x_N\}$ into a set of uncorrelated vectors $v_{(1)}, \dots, v_{(K)}$, each explaining progressively less of the data's variance than the last one. The vectors $\left\{v_{(k)}\right\}_{k=1}^{K}$, called principal components, are obtained through the recursion relation:

$$\begin{aligned}
\hat{\mathbf{Q}}_k &\triangleq \mathbf{Q} - \sum_{s=1}^{k-1} \mathbf{Q}\mathbf{x}_{(s)} v_{(s)}^{\mathrm{T}} \\
v_{(k)} &\triangleq \operatorname*{arg\,max}_{\|v\|=1} \left\{ \|\hat{\mathbf{Q}}_k v\|^2 \right\} \\
v_{(0)} &\triangleq 0.
\end{aligned} \tag{2.8}$$

where $\mathbf{Q}$ is the empirical data matrix with column-wise means removed.

PCA is commonly used in finance, where high dimensional data is typical. A classical use is for pricing zero-coupon bonds. Denote by $B(t,T)$ the price of a zero-coupon bond with maturity T at time t. The price $B(t,T)$ can be modeled using the yield curve $y(t,T)$, which is defined as the rate at which the price of the bond is equal to the discounted cash flows. That is,

$$y(t,T) \triangleq \ln\left(\frac{B(t,T)}{T-t}\right).$$

The first three principal components of the yield curve are known to explain its level, slope, and curvature respectively (see [30] for more details). The validation-set loss function which we will use is

$$\min_{\beta_1,\dots,\beta_k \in \mathbb{R}^{\tilde{K}}} \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{\tilde{K}} \beta_k v_{(k)} \right)^2, \tag{2.9}$$

where Y_i is the vector of Bond yields observed on the i^{th} day in the validation set (resp. training set) and $\tilde{K} \leq K$ is the number of principal components used to give a low dimensional approximation of the yield curve. As discussed in [30], the first three principal components $v_{(1)}, v_{(2)}, v_{(3)}$ of most yield curves tend to explain about 95% of the data's variance.

As a benchmark a two common alternatives to PCA, Kernel PCA (kPCA) and sparse PCA (sPCA) will be also be considered. Kernel PCA, performs first maps the data into another space, called the feature space, wherein the data can be more naturally partitioned by hyperplanes and the performs PCA in the feature space. The transformation into the feature space is typically made indirect by only describing the feature space's inner product, which is possible due to the reproducing kernel Hilbert space structure of the feature space. A choice of inner product between two vectors v_1, v_2 in the feature space is

$$\begin{aligned} &t(v_1)Kv_2 \\ K \triangleq &\left(e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}} \right)_{i,j=1}^{N} . \end{aligned}$$

Unlike NEU-PCA, the non-linear transformation used in kPCA is not learned from the data but chosen before the algorithm is executed. Since kPCA does not make computations directly in the feature space but works indirectly to it by exploiting its inner product, kPCA does not allow for reconstruction of the data. However this is not the case with NEU-PCA, since it is entirely constructive.

Analogously to the LASSO, Ridge regression, and ENET regularization problems, sPCA penalizes the Equation (2.8) to in order to obtain sparser principal components. The implementation considered in this chapter will use the sPCA formulation of [36]. Sparse PCA has the advantage over PCA of being more interpretable, lower-dimensional, and being more robust due to its low dimensionality (see [106, 36] for more details on sPCA).

For this illustration PCA, kPCA, sPCA, NEU-PCA, NEU-kPCA, and NEU-sPCA will all be performed on bond yield data. The daily bond data considered in this example consists of stripped German government bond prices between January 4^{th} 2010 and December 30^{th} 2014. The considered bond maturities are between 6 months and 30 years. The training-set consists of the first 1000 days of data, the validation set of the next 200 days, and the test set consists of the remainder. The reconfigurations defining the NEU methods with be learned using NEU-PCA. The NEU-kPCA and NEU-sPCA methods will be use the reconfigurations learned from NEU-PCA.

The NEU-PCA algorithm is implemented by optimized the training and validation objective functions by alternating between random searches and performing bulk iterations of the Nelder-Mead heuristic search method (see [78] for details Nelder-Mead optimization). This heuristic scheme provided faster convergence results than direct use of stochastic gradient descent as in Example 2.2.13 due to the data's high dimensionality. After learning the reconfigurations defining the NEU-PCA algorithm, the same reconfigurations were used to define NEU-kPCA and NEU-sPCA. This is interpreted as a form of transfer learning between analogous models.

N.Fact.	PCA	NEU-PCA	kPCA	NEU-kPCA	sPCA	NEU-sPCA
1	0.7749	0.7868	0.0906	0.0894	0.9756	0.9774
2	0.8833	0.8936	0.9171	0.9175	0.9942	0.9949
3	0.9417	0.9506	0.9948	0.9955	0.9992	0.9996
4	0.9654	0.9688	0.9981	0.9981	0.9999	0.9999

Table 2.5: Comparison of Variance Explained in Training Set.

Table 2.5 shows that NEU-PCA explains more of the training set variance than PCA does. However, kPCA and sPCA seem to explain more training set variance than NEU-PCA, but not as much as NEU-kPCA or NEU-sPCA. However, examining the test-set predictive performance of the four algorithms in Table 2.6, it is observed that the kPCA based algorithms are not able to accurately forecast the yield curve. Therefore, NEU-PCA is the most parsimonious option for prediction between the four methods and NEU-kPCA explains the most training set variance of the data.

The more modest gains of this method are due to the district training and validation loss functions. For example, removing the validation loss-function and thereby the early stopping criterion in the definition of NEU, it can be seen that one NEU-PCA can explain more than 99.99% of the training set variability of the data. However, this leads to poor out-of-sample predictions of the test set yield curves as well as uninterpretable NEU-PCAs.

N.Fact.	PCA	NEU-PCA	kPCA	NEU-kPCA	sPCA	NEU-sPCA
1	2,245.643	2,153.412	829.210	827.651	497.683	471.695
2	344.961	294.106	829.200	827.644	290.040	265.822
3	28.633	17.927	829.197	827.640	14.489	12.400
4	4.424	2.975	829.190	827.634	12.061	12.210

Table 2.6: Comparison of test set Predictions

In this implementation, the NEU-PCAs of the yield curve. Figure 2.4 shows that, upon rescaling, the first and fourth PCA and NEU-PCAs have identical interpretation, while the second and fourth NEU-PCAs look similar a flipped version of the second and fourth PCAs. The NEU-PCAs in Figure 2.4 are in the transformed, non-Euclidean space, whereas the PCAs in Figure 2.4 are in Euclidean space itself. It should not be surprising that the 1 and 4 factor sPCA outperforms the 1-factor NEU-sPCA since the reconfiguration used for the NEU-sPCA was trained using the PCA algorithm.

In this implementation, the NEU-PCAs provided the most robust out-of-sample predictions of the yield curve, explained more of the training set variance than PCAs did and retained the interoperability of each of the principal components. Moreover like PCA, the approach is constructive therefore can be used for reconstruction purposes, which is not

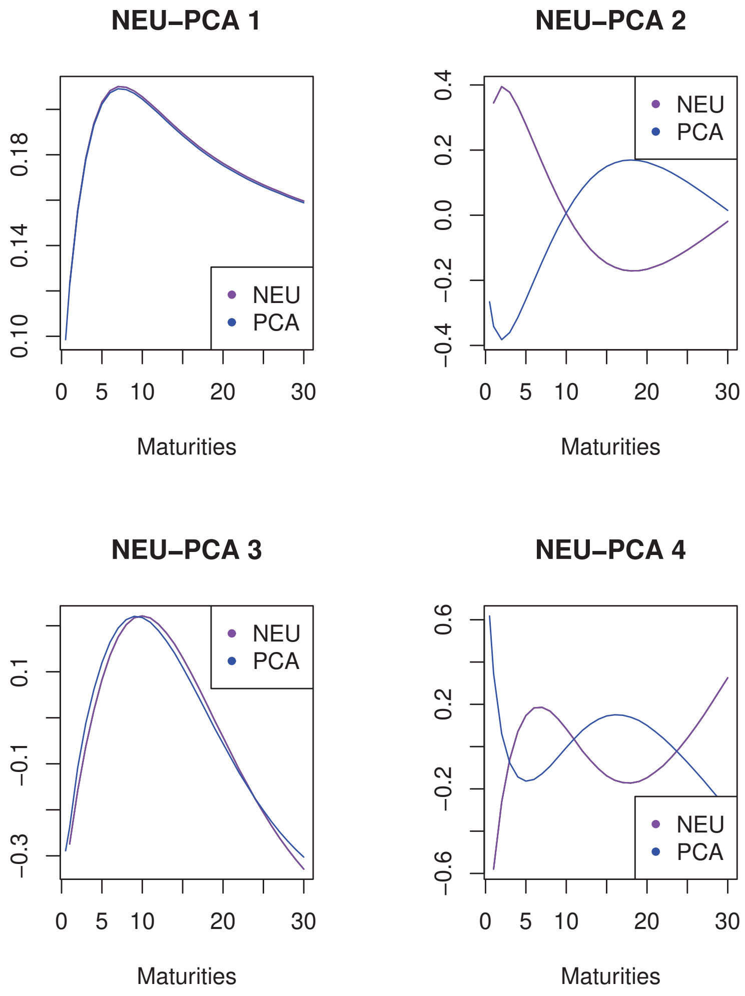


Figure 2.4: First four principal components of the German Bond Yield-curve.

the case for kPCA due to it indirectly working with the feature space (see [87, Section 4] for a brief discussion on the data-reconstruction shortcomings of kPCA).

Table 2.7 examines the runtime of each method. All six algorithms were run on a machine with the same specs as those of Example 2.2.13.

Algorithm	Run Time (sec)	$\frac{\texttt{Run Time}}{\texttt{Run Time PCA}}$
PCA	0.01	1
NEU-PCA	2.89	474.99
kPCA	0.08	12.50
NEU-kPCA	2.96	486.48
sPCA	0.81	132.40
NEU-sPCA	3.70	606.39

Table 2.7: Runtime Comparison.

The central shortcoming of the NEU meta-algorithm is underlined by Table 2.7. Its second row shows that the runtime of the NEU algorithms are about 1000 times slower than PCA and 100 times slower than kPCA. Therefore if speed is necessary it may be more desirable to turn to PCA or kPCA than their NEU counterparts. However, if time can be spared then the first three NEU-PCAs makes 3-factors NEU-PCA the best overall choice due to its interpretability, out-of-sample predictive power and it explaining a competitive level of the training set's variance.

Both numerical implementation show that the NEU algorithm makes simple algorithms competitive by embedding the universal approximation and universal reconfiguration properties into them. Future research could investigate the performance of the NEU meta-algorithm applied to other learning procedures such as clustering or classification tasks.

2.4 Conclusion

In this chapter reconfigurations were introduced and shown to have the universal reconfiguration property introduced in Theorem 2.2.9, which stated than any dataset could be transformed into any other dataset using a reconfiguration. Applying the universal reconfiguration property to the graph of a continuous function, it was shown that reconfigurations also have the universal approximation property (Corollary 2.2.10) of neural networks.

The NEU meta-algorithm was introduced. NEU builds the universal reconfiguration and universal approximation properties into any objective learning algorithm. The resulting algorithm is found in three steps. First the optimal reconfiguration, which best relates a given dataset to the loss function defining the learning algorithm is learned. The old algorithm is then preformed on the new reconfigured space and subsequently the prediction made by the learning algorithm is moved back to the original space by deconfiguration. Given any objective learning algorithm A, the algorithm NEU-A was shown to outperform A, in the sense that it exhibits a lower validation-set loss.

The performance increase was justified both theoretically and supported empirically. The empirical experiments found that the variance of German bond yields was better explained with one NEU principal component than with 4 ordinary principal components. Likewise, the investigations of Apple stock price found that the residuals of the NEU-OLS algorithm was smaller than those of OLS, Ridge, LASSO, and ENET regression. The effectiveness of NEU-OLS as a non-parametric estimator was explored in three simulation studies which showed that using the universal reconfiguration and universal approximation properties of reconfigurations. It was confirmed that NEU-OLS is not only competitive with other non-parametric regression methods but can better approximate functions with discontinuities, have non-locally determined behavior, or exhibit an oscillatory behavior.

A consequence of the construction of the NEU of an algorithm, is that once a correct geometry is learned for the algorithm given the data, the algorithm can be executed directly in the associated non-Euclidean space. This gave fast and simple algorithms such as linear regression higher validation set performance than their complicated and difficult to train Euclidean counterparts. These techniques can be applied outside of mathematical finance and we believe there are many applications in geonomics and mathematical imaging, where traditional machine learning algorithms are used.

The NEU meta-algorithm was shown to increase the explanatory and predictive power of an algorithm, both theoretically and through the implementations considered in this chapter. However, NEU does reduce the speed of the original algorithm on the first run, when the reconfigurations are being learned. Future research needs to be done to find a way to minimize this computational shortcoming.

2.5 Appendix

This appendix contains a list of all the technical regularity conditions used in this chapter.

Regularity Condition 2.5.1 (Regularity Condition) *For every choice of hyper-parameters $\gamma \in \Gamma$ and every data-sample $X \in \mathbb{R}^{DN_I}$ the map*

$$\begin{aligned} \mathscr{L}_I^{N_I}\left(\cdot, \phi(\cdot|X^I); \gamma \mid X^I\right) &: \mathbb{R}^d \to \mathbb{R} \\ &\beta \mapsto \mathscr{L}_I^{N_I}\left(\beta, \phi(\beta|X^I); \gamma \mid X^I\right), \end{aligned}$$

has a, possibly not unique, infimum.

Regularity Condition 2.5.2 *Let (X^I, X^O) be a pair of training and validation datasets.*

(i) For every hyper-parameter $\gamma \in \Gamma$, there exists a unique optimal parameter $\beta(\gamma)$ in $\mathbb{R}^d$ such that for all other parameters $\tilde{\beta}$ in $\mathbb{R}^d$

$$\mathscr{L}_I^{N_I}\left(\beta(\gamma), \phi(\beta(\gamma)|X^I); \gamma \mid X^I\right) < \mathscr{L}_I^{N_I}\left(\tilde{\beta}, \phi(\tilde{\beta}|X^I); \gamma \mid X^I\right),$$

(ii) There exists a unique hyper-parameter $\hat{\gamma}$ in Γ such that for every other hyper-parameter γ in Γ

$$\mathscr{L}_O^{N_O}\left(\beta(\hat{\gamma}),\phi(\beta(\hat{\gamma})|X^O);\hat{\gamma}\mid X^O\right)<\mathscr{L}_O^{N_O}\left(\beta(\gamma),\phi(\beta(\gamma)|X^O);\gamma\mid X^O\right).$$

Regularity Condition 2.5.3 *The pair $\left(\left\{\Phi\left(X_j^I\right)\right\}_{j=1}^{N_I},\left\{\Phi\left(X_i^O\right)\right\}_{i=1}^{N_O}\right)$ is in $Dom(\mathscr{L}_I,\mathscr{L}_O,\Gamma,\phi)$.*

Regularity Condition 2.5.4 *$D>1$ and there exists $\left(\tilde{X}_I,\tilde{X}_O\right)\in Dom(\mathscr{L}_I,\mathscr{L}_O,\Gamma,\phi)$ such that $\mathscr{P}\left(\mathscr{L}_I,\mathscr{L}_O\right)\left(\tilde{X}_I,\tilde{X}_O\right)<\mathscr{P}\left(\mathscr{L}_I,\mathscr{L}_O\right)\left(X^I,X^O\right)$.*

Regularity Condition 2.5.5 *There exists a regular compact set V containing the compact set $[K\times f(K)]\cup[K\times g(K)]$such that for every $\theta\in\Theta$, the reconfiguration map $\mathfrak{r}\left(\cdot|\theta\right)$ on $int(V)$. For every $\theta\in\Theta$, $\mathfrak{r}\left(\cdot|\theta\right)$ the partial derivatives of $\mathfrak{r}$ are uniformly bounded by 1.*

Proof of Proposition 2.2.3. Since additive conjugation by c is its own inverse, we may assume that $c=0$. For any $X\in\mathfrak{so}(D)$, $exp(X)$ is a rotation matrix and is therefore an isometry from $\mathbb{R}^D$ onto itself. Therefore,

$$\|exp(f(\|x\|)X)x\|=\|x\|, \tag{2.10}$$

from which it follows that

$$f(\|x\|)=f(\|exp(f(\|x\|)X)x\|). \tag{2.11}$$

Moreover, since exp is a group homomorphism

$$exp(X+Y)=exp(X)exp(Y). \tag{2.12}$$

Combining Equations (2.11) and (2.12) we obtain

$$\begin{aligned} I=&\,(exp(f(\|x\|)X)x-f(\|x\|)X)x)\\ =&\,(exp(f(\|x\|)X)x)\,(exp(f(\|x\|)(-X))x)\\ =&\,(exp(f(\|x\|)X)x)\,(exp([f(\|exp(f(\|x\|)X)x\|)]\,(-X))x)\,. \end{aligned} \tag{2.13}$$

Therefore Definition 2.2.1 (i) holds.

The maps *exp* and ϕ are infinitely differentiable, see [68] and [48] respectively. Moreover, since $\|\cdot\|$ and $\cdot\pm c$, therefore $\Psi(\cdot|\theta)$ and $\Phi(\cdot|\theta)$ are infinitely differentiable. Therefore Definition 2.2.1 (ii) and (iii) hold.

Let c be the midpoint between x and y, moreover let $\epsilon\triangleq\|x-y\|$ and $\sigma\triangleq\frac{\|x-z\|+\epsilon}{2}$. Therefore, for any choice of X in $\mathfrak{so}(D)$, $\mathfrak{r}(z|(c,\sigma,X))=z$. Since *exp* maps $\mathfrak{so}(D)$ onto the set $SO(D)$ which is the collection of all maps from the D-sphere onto itself and x,y lie on the same sphere centered at c of radius ϵ, then there exists X in $\mathfrak{so}(D)$ for such that $\mathfrak{r}\left(x|(c,\sigma,X)\right)$ is a rotation taking x to y. Therefore Definition 2.2.1 (iv) holds.

For a triple $(c,\sigma,0)$, the application of the map $\mathfrak{r}\left(x|(c,\sigma,X)\right)$ becomes multiplication by the identity matrix in this case, hence (v) holds. □

Proof of Proposition 2.2.6. Definition 2.2.1 (i,ii,iii,v) hold analogously to the proof of Proposition 2.2.3. To see Definition 2.2.1 (iv) note that if c is the midpoint between x and y and σ is taken to be $d(x,y)$ then $\mathfrak{k}$ is the identity function outside the ball centered at c of radius σ, therefore $\mathfrak{k}(z|c,\sigma,X) = z$ for any $X \in \mathbb{R}$. Taking $X = (x-y)/\psi(\|x-y\|;\sigma)$ establishes Definition 2.2.1 (iv). □

The proof of Theorem 2.2.9 relies on the construction of a particular curve described by the next Lemma.

Lemma 2.5.6. Let $D > 1$ and let $x_1,\dots,x_n,x,z$ be distinct points in $\mathbb{R}^D$ and let $\Delta \in (0,\infty]$. Then there exists positive integers n, K and a curve γ from x to z satisfying

(i) γ is rectifiable with length l,

(ii) $\gamma(0) = x$ and $\gamma(1) = z$,

(iii) $0 < \frac{1}{n} < \min_{\substack{t\in[0,1]\\ i,j=1,\dots,N}} \{\|\gamma(t) - x_i\|, \Delta\}$,

(iv) $\gamma([0,1]) = \gamma([0,1]) \cap \left[\bigcup_{k=1}^K Ball\left(\gamma\left(\frac{k}{K}\right); \frac{1}{K}\right)\right]$,

(v) $\emptyset = \bigcup_{i=2}^n \{x_i\} \cap \left[\bigcup_{k=1}^K Ball\left(\gamma\left(\frac{k}{K}\right); \frac{1}{K}\right)\right]$.

Proof of Lemma 2.5.6. The existence of such a curve is equivalent to looking for a smooth curve inside the open set $\mathbb{R}^D - \bigcup_{i=1}^N \{X_i\} \cup \left[\bigcup_{i=1}^N \{\tilde{X}_i\}\right]$, which is in-turn equivalent to the open subset $\mathbb{R}^D - \bigcup_{i=1}^N \{X_i\} \cup \left[\bigcup_{i=1}^N \{\tilde{X}_i\}\right]$ of $\mathbb{R}^D$ being simply connected. More generally, let X_N be $\mathbb{R}^D$ with the N-distinct points $\{x_1,\dots,x_N\}$ deleted.

Simply connectedness of X_N will be proven by strong induction on N. If $N = 1$ then X_N is simply connected since X_1 is homeomorphic to $\mathbb{R}^D - \{0\}$ which is a deformation retract of the $(D-1)$-sphere S^{D-1} (see [56, Excerise 0.2]). Since homeomorphisms and deformation retractions induce chain homotopies in their associated chain complexes (see [56, Chapter 2.1]) and since the homology functors H_n are invariant under chain homotopies (see [56, Proposition 2.1.2]), then there are group isomorphism

$$H_n(X_1) \cong H_n(\mathbb{R}^D - \{0\}) \cong H_n(S^{D-1}) \cong \begin{cases} \mathbb{Z} & \text{if } n = D, 0 \\ 0 & \text{else.} \end{cases}$$

where the last isomorphism is computed in [95, Theorem 4.6.6][4]. Applying [95, Lemma 4.4.7] implies that X_1 is path-connected. Suppose that X_N is path connected for some

[4] Actually, it is computed for the reduced homology. However, [95, Lemma 4.3.1] permits the translation into singular homology.

$N \geq 1$. Since the interiors of the sets[5] $A \triangleq \mathbb{R}^D - \cup_{i=2}^N \{x_i\}, B \triangleq \mathbb{R}^D - \{x_1\}$ cover $\mathbb{R}^D$ and their intersection is X_N. The Mayer-Vietoris sequence (see [95, page 190]) implies that there is a long-exact sequence in singular homology

$$0 \cong H_0(\mathbb{R}^D) \leftarrow H_0(\mathbb{R}^D) \leftarrow H_0(A) \oplus H_0(B) \leftarrow H_0(A \cap B) \leftarrow H_1(\mathbb{R}^D). \tag{2.14}$$

By [95, Lemma 4.4.1], $\mathbb{R}^D$ is contractible, therefore $H_0(\mathbb{R}^D) \cong \mathbb{Z}$ and $H_n(\mathbb{R}^D) \cong 0$ if $n > 0$. Applying the strong induction hypothesis that $H_0(A) \cong \mathbb{R} \cong H_0(B)$, it follows that

$$0 \cong \mathbb{Z} \leftarrow \mathbb{Z} \oplus \mathbb{Z} \leftarrow H_0(A \cap B) \leftarrow 0$$

is an exact sequence of groups. The Splitting Lemma implies that

$$\mathbb{Z} \cong \mathbb{Z} \oplus H_0(A \cap B);$$

therefore $H_0(A \cap B) = \mathbb{Z}$, hence $A \cap B$ is path connected by [95, Lemma 4.4.7]. Since $A \cap B = X_d$, it follows that X_d is path connected. Picking $x_1, \dots, x_N$ to be the data-points $\bigcup_{i=1}^N \{X_i\} \cup \bigcup_{i=1}^N \{\tilde{X}_i\}$ it follows that there exists a path $\tilde{\gamma}$ which is interior to $\mathbb{R}^D - \bigcup_{i=1}^N \{X_i\} \cup \bigcup_{i=1}^N \{\tilde{X}_i\}$ connecting X_1 to $\tilde{X}_1$. This completes the induction step.

Since $\tilde{\gamma}([0,1])$ is compact there exists a finite open cover $\{V_j\}_{j=1}^J$ of $\tilde{\gamma}([0,1])$. Therefore, the open sets $\left\{W_j \triangleq V_j \cap \left[\mathbb{R}^D - \bigcup_{i=1}^N \{X_i\} \cup \bigcup_{i=1}^N \{\tilde{X}_i\}\right]\right\}_{j=1}^J$ is a finite collection of sets diffeomorphic to $\mathbb{R}^D$, via some diffeomorphism $\{\phi_j\}_{j=1}^J$. Therefore each $\phi_j \circ \gamma|_{W_j}$ defines a continuous path from into $\mathbb{R}^D$. The Whitney Approximation Theorem (see [71, Theorem 6.12]) implies that for each j in $\{1, \dots, J\}$ there exists a smooth curve $\tilde{\tilde{\gamma}}_j$ which is δ-close to $\phi_j \circ \gamma$; where

$$\delta \triangleq \min_{i=2,\dots,N} \frac{\left\{\|X_i - \tilde{\gamma}(t)\|, \|\tilde{X}_i - \tilde{\gamma}(t)\|\right\}}{2}.$$

Therefore the composition $\phi_j^{-1} \circ \tilde{\tilde{\gamma}}_j$ a piecewise-smooth curve, denoted by γ, which joins X_1 to $\tilde{X}_1$ and is contained entirely within $\mathbb{R}^D - \left[\bigcup_{i=1}^N \{X_i\} \cup \bigcup_{i=1}^N \{\tilde{X}_i\}\right]$. Since every piecewise-smooth curve is rectifiable, by definition of arc-length γ has finite arc-length, which we denote by l. This establishes (i) and (ii).

Define $\epsilon > 0$ as

$$\epsilon \triangleq \min_{i=2,\dots,N} \frac{\left\{\|X_i - \gamma(t)\|, \|\tilde{X}_i - \gamma(t)\|\right\}}{2}.$$

By the Archimedean property of $\mathbb{R}$, there exists a positive integer n for which $0 < \frac{1}{n} <$

[5] The interior of an open set is itself.

$min\{\epsilon, l\}$. Observe that the definitions of ϵ and γ imply that

$$\begin{aligned}\gamma([0,1]) =& U \cap \gamma([0,1]),\\ \emptyset =& \bigcup_{i=2}^{N} \{X_i, \tilde{X}_i\} \cap U,\\ U \triangleq & \bigcup_{k=1}^{n} Ball\left(\gamma\left(\frac{k}{n}\right); \frac{1}{n}\right);\end{aligned}$$

therefore $(iii) - (v)$ hold. □

Proof of Theorem 2.2.9. $\mathscr{M}$ is diffeomorphic to $\mathbb{R}^D$ it may be assumed without loss of generality that $\mathscr{M} = \mathbb{R}^D$. The case that $X = \tilde{X}$ must hold with $K = 1$ by choosing θ_1 to be any element of Θ_0, which is possible since Θ_0 is non-empty. Assume without loss of generality that the collections X and $\tilde{X}$ are formed of distinct elements.

We proceed by induction. Suppose that $N = 1$. Let $\epsilon \triangleq 2\|X_1 - \tilde{X}_1\|$ and let Z be any point in $\mathbb{R}^D$ for which $\|Z - X_1\| > \epsilon$. Then the local-transience property of $\mathfrak{X}$ implies that there exists $\theta_1 \in \Theta$ such that

$$\mathfrak{X}(X_1|\theta_1) = \tilde{X}_1; \mathfrak{X}(Z|\theta) = Z.$$

Suppose now that the claim holds for $N \geq 1$. In the notation of Lemma 2.5.6, let $x_1, \dots, x_n = X_2, \dots, X_N, \tilde{X}_2, \dots, \tilde{X}_N$, $x = X_1$, and $z = \tilde{X}_1$. Since there exists a rectifiable curve γ connecting x to z and γ is uniformly bounded away from each x_i by a distance of at least $\frac{1}{n}$, where $\frac{1}{n}$ is set to be less than the locality Δ, then there exists a set of open balls $\left\{Ball\left(\gamma\left(\frac{k}{K}\right); \frac{1}{K}\right)\right\}_{k=1}^{K}$ covering γ which are separated from the points $x_1, \dots, x_n$, by a distance of at least $\frac{1}{n}$. The local transience property of $\mathfrak{X}$ implies that there exist $\theta_1^1, \dots, \theta_K^1$ in Θ satisfying

$$\begin{aligned}\mathfrak{X}\left(\gamma\left(\frac{k-1}{K}\right)|\theta_k^1\right) =& \gamma\left(\frac{k}{K}\right)\\ \mathfrak{X}\left(x_i|\theta_k^1\right) =& x_i,\end{aligned}$$

for every i in $\{1, \dots, n\}$ and every k in $\{1, \dots, K\}$.

Repeating this construction and process for every data-point X_i we find a list of parameters

$$\theta_1^1, \dots, \theta_{K_1}^1, \theta_1^2, \dots, \theta_{K_2}^2, \dots, \theta_{K_N}^N,$$

such that

$$\mathbb{X}\left(X_j|\theta_1^i, \dots, \theta_{K_i}^i\right) = \begin{cases}\tilde{X}_i & \text{if} i = j\\ X_j & \text{else.}\end{cases}$$

Definition 2.2.1[(iv)] implies that any Z not in $\cup_{k=1}^{K} Ball(\gamma\left(\frac{k}{K}\right); \frac{1}{n})$ must remain fixed by $\mathbb{X}$. □

Proof of Corollary 2.2.10. Let K be an open subset of $\mathbb{R}^{D_1}$, diffeomorphic to $\mathbb{R}^{D_1}$ and let Q be a countable subset of K. Let $\{x_i\}_{i\in\mathbb{N}}$ be an enumeration of Q and define the sequences of points $\{X_i\}_{i\in\mathbb{N}}$ and $\{\tilde{X}_i\}_{i\in\mathbb{N}}$ in $\mathbb{R}^{D_1+D_2}$ by

$$\begin{aligned} X_i &\triangleq (x_i, g(x_i)) \\ \tilde{X}_i &\triangleq (x_i, f(x_i)). \end{aligned}$$

Since $\mathbb{R}^{D_1+D_2}$ is of dimension at least 2 and K is diffeomorphic to $\mathbb{R}^{D_1}$, then for every $n \in \mathbb{N}$ Theorem 2.2.9 applies to the sets of points $\{X_i\}_{i=1}^n$ and $\{\tilde{X}_i\}_{i=1}^n$ in $\mathbb{R}^{D_1+D_2}$. Therefore for every $n \in \mathbb{N}$ there exists $\theta_1^n, \dots, \theta_{N_n}^n \in \Theta$ such that

$$\mathbb{X}\left(X_i|\theta_1^n, \dots, \theta_{N_n}^n\right) = \tilde{X}_i; i = 1, \dots, n. \tag{2.15}$$

Define the sequence of functions $\{f_n\}_{n\in\mathbb{N}}$ from K to $\mathbb{R}^{D_2}$ by

$$f_n(x) \triangleq p \circ \mathbb{X}\left((x, g(x))|\theta_1^n, \dots, \theta_{N_n}^n\right), \tag{2.16}$$

where p is the second canonical projection on $\mathbb{R}^{D_1} \times \mathbb{R}^{D_2}$ onto $\mathbb{R}^{D_2}$. From equation (2.15), it follows that the sequence $\{f_n\}_{n\in\mathbb{N}}$ converge point-wise to f on Q. This establishes the $\epsilon = 0$ case of (i).

Now assume that $\epsilon > 0$. Since $\mathbb{P}$ is a probability measure, Q is of finite $\mathbb{P}$-measure. Therefore $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of Borel-measurable functions over a set of finite $\mathbb{P}$-measure converging point-wise to the Borel-measurable function f, where f takes values in the separable metric space $(\mathbb{R}^{D_2}, d_E)$. Here d_E is the Euclidean metric on $\mathbb{R}^{D_2}$. Hence, Egorov's Theorem (see [83] for details) gives the existence of the set K_ϵ as well as the uniform convergence of the sequence $\{f_n\}_{n\in\mathbb{N}}$ to f on K_ϵ. This establishes (i).

□

3. Non-Euclidean Conditional Expectation and Efficient Portfolio Filtering

A non-Euclidean generalization of conditional expectation is introduced, proven to exist, and characterized as the minimizer of expected intrinsic squared-distance from a manifold-valued target. The computational tractable formulation expresses the non-convex optimization problem as transformations of Euclidean conditional expectation. This gives computationally tractable filtering equations for the dynamics of the intrinsic conditional expectation of a manifold-valued signal and is used to obtain accurate numerical forecasts of efficient portfolios by incorporating their geometric structure into the estimates.

3.1 Introduction

Non-Euclidean geometry occurs naturally in problems in finance. Short-rate models consistent with finite-dimensional Heath-Jarrow-Morton (HJM) models are characterized using Lie group methods, in [10]. In [59], highly accurate stochastic volatility model estimation methods are derived using Riemannian heat-kernel expansions. In [39], the equivalent local martingale measures (ELMMs) of finite-dimensional term-structure models for zero-coupon bonds are characterized using the smooth manifold structure associated which factor models for the forward-rate curve. In [15], information-geometric techniques for yield-curve modeling which consider finite-dimensional manifolds of probability densities are developed. In [53, 52], Riemannian geometric approaches to stochastic volatility models and covariance matrix prediction are employed to successfully predicts stock prices. In [70], it is shown that considering a relevant geometric structures on a mathematical finance problem leads to more accurate out-of-sample forecasts. The superior forecasting power of non-Euclidean methods is interpreted as encoding information present in mathematical finance problems which is otherwise overlooked by the classical Euclidean methods. Each of these methodologies approach distinct problems in mathematical finance using differential geometry.

Conditional expectation and stochastic filtering are some of the most fundamental tools used in applied probability and finance. Geometric formulations of conditional expectation, such as those used in [94, 85] are solutions to non-convex optimization problems. The non-convexity of the problem makes computation of these formulations of non-Euclidean conditional expectations difficult or intractable.

Non-Euclidean filtering formulations such as those of [31], [79], or [64] assume that the signal and/or noise processes are non-Euclidean and estimate functionals of the noisy signal using the classical Euclidean conditional expectation. In [85] dynamics for the intrinsic conditional expectation of a manifold-valued signal was found, using the Le Jan-Watanabe connection. This connection reduced the intrinsic non-Euclidean filtering problem to a Euclidean filtering problem. However, the authors of [85] remark that implementing their results may be intractable due to the added complexity introduced by the Le Jan-Watanabe connection.

This chapter presents an alternative computationally tractable characterization of intrinsic conditional expectation, called dynamic conditional expectation, and uses it to produce a computable solution to a non-Euclidean filtering problem similar to that of [85]. The implementation is similar to [94] for a non-Euclidean particle filter. However, in [94] the convergence of the algorithm to the non-Euclidean conditional expectation is left unjustified. The dynamic conditional expectation expresses the intrinsic conditional expectation as a limit of certain transformations of Euclidean conditional expectations associated to the non-Euclidean signal process. Analogue to [85], this reduces the computation of the non-Euclidean problem the computation of a Euclidean problem with the central difference being that the required transformations are available in closed form. The infinitesimal linearization transformations considered here are similar to those empirically postulated in the engineering, computer-vision, and control literature in [42, 41, 53, 57, 3, 94].

Portfolios maximizing returns given a fixed risk-appetite have a natural non-Euclidean structure. The empirical performance of our filtering algorithm is evaluated on the space of efficient portfolios called the Markowitz space. These are compared against other intrinsic filtering algorithms from the engineering and computer vision literature.

3.2 Preliminaries and Notation

For the duration of this chapter, $(\Omega, \mathscr{F}, \mathscr{F}_\cdot, \mathbb{P})$ will be a complete stochastic base on which independent Brownian motions, denoted by W_t and B_t are defined. Furthermore, $\mathscr{G}_\cdot$ will denote a sub-filtration of $\mathscr{F}_\cdot$. The vector-valued conditional expectation will be denoted by $\mathbb{E}_{\mathbb{P}}[X_t|\mathscr{G}]$.

The measure m will denote the Lebesgue measure, $L^p_m(\mathscr{F};\mathbb{R}^D)$ will denote the Bochner-Lebesgue spaces for $\mathscr{F}$-measurable $\mathbb{R}^D$-valued functions with respect to the D-tuples of Lebesgue measure m. If $D = 1$, the Bochner-Lebesgue spaces will be abbreviated by $L^p_m(\mathscr{F})$. For a Riemannian manifold $(\mathscr{M}, g)$, the intrinsic measure is denoted by μ_g and the induced distance function is denoted by d_g.

The disjoint union, or coproduct, of topological spaces will be denoted by $\coprod$. The set of càdlàg paths from $\mathbb{R}$ into the metric space induced by $(\mathscr{M}, g)$, is defined by $D(\mathbb{R}; \mathscr{M}, d_g)$.

The next section motivates the geometries studied in this chapter by introducing and discussing the geometry of efficient portfolios.

3.3 The Geometry of Efficient Portfolios

A fundamental problems in mathematical finance is choosing an optimal portfolio. Typically, in modern portfolio theory, a portfolio is comprised of D predetermined risky assets and a riskless asset. Efficient portfolios are portfolios having the greatest return but not exceeding a fixed level of risk. Classically, the return level is measured by the portfolio's expected (log)-returns. The portfolio's risk is quantified as the portfolio's variance. The optimization problem defining efficient portfolios may be defined in a number of ways, the one considered in this chapter is the following Sharpe-type ratio

$$\hat{w}(\gamma, \mu, \Sigma) \triangleq \underset{\substack{w \in \mathbb{R}^D \\ \bar{1}^\star w = 1}}{\operatorname{argmin}} \left(-\gamma \mu^\star w + \frac{w^\star \Sigma w}{2} \right). \tag{3.1}$$

Here w is the vector of portfolio weights expressed as the proportion of wealth invested in each risky asset, $\mu \in \mathbb{R}$ is the vector of the expected log-returns of the risky assets, Σ is the covariance matrix of those log-returns, γ is a parameter balancing the objectives of maximizing the portfolio return versus minimizing the portfolio variance, $\bar{1}$ is the vector with all its components equal to 1, and * indicates matrix transpose operation. If Σ is not degenerate, the unique optimal solution to Equation (3.1) is

$$\hat{w}(\gamma, \mu, \Sigma) = \frac{\Sigma^{-1}\bar{1}}{\bar{1}^\star \Sigma \bar{1}} + \gamma \left(\Sigma^{-1}\mu - \frac{\bar{1}^\star \Sigma^{-1} \mu}{\bar{1}^\star \Sigma^{-1} \bar{1}} \Sigma^{-1} \bar{1} \right). \tag{3.2}$$

The particular case where t is set to 0 is the minimum variance portfolio of [74]. The minimum-variance portfolio $\hat{w}(0, \mu, \Sigma)$ may also be derived by minimizing the portfolio variance subject to the budget constraint $\bar{1}^\star w = 1$. By adding a risk-free asset to the portfolio, one can derive similar expressions for the market portfolio and the capital market line (for more details on this approach to portfolio theory see [7]).

Unlike the returns vector μ, a portfolio's covariance matrix is not meaningfully represented in Euclidean space. That is, a covariance matrix does not scale linearly and the difference of covariance matrices need not be a covariance matrix. Therefore, forecasting a future covariance matrix, even through a simple technique such as linear regression directly to the components of Σ, can lead to meaningless forecasts. Using the intrinsic geometry of the set of positive-definite matrices, denoted by $\mathscr{P}_D^+$, avoids these issues.

The space $\mathscr{P}_D^+$, has a well studied and rich geometry lying at the junction of Cartan-Hadamard geometry and Lie theory. Empirical exploitation of this geometry has found many applications in mathematical imaging (see [76]), computer vision (see [80]), and signal processing (see [6]). Moreover, connections between this geometry and information theory have been explored in [93], linking it to the Cramer-Rao lower bound.

The set $\mathscr{P}_D^+$ is smooth and comes equipped with a natural infinitesimal notion of distance called Riemannian metric. Denoted by g, the Riemannian metric on $\mathscr{P}_D^+$ quantifies the difference in making infinitesimal movements in Euclidean space along $\mathscr{P}_D^+$ to making infinitesimal movements with respect to the geometry of $\mathscr{P}_D^+$. The description of Riemannian manifolds as subsets of Euclidean space is made rigorous by Nash in the embedding theorem in [77]. Distance between two points on $\mathscr{P}_D^+$ is quantified by the length of the shortest path connecting the two points, called a geodesic. On $\mathscr{P}_D^+$, any two points can always be joined by geodesic. The distance function taking two points to the length of the unique most efficient curve joining them can be expressed as

$$d_g^2\left(\Sigma_1, \Sigma_2\right) \triangleq \left\|\log\left(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}}\right)\right\|_F^2 = \sum_{i=1}^{d} \lambda_i^2\left(\log\left(\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}\right)\right). \tag{3.3}$$

The function d_g makes $\mathscr{P}_D^+$ into a complete metric space, where the distance between two points corresponds exactly to the length of the unique distance minimizing geodesic connecting them. Where, $\|\cdot\|_F$ is the Frobenius norm, which first treats a matrix as a vector and subsequently computes its Euclidean norm, $\Sigma^{\frac{1}{2}}$ is the matrix square-root operator, log is the matrix logarithm, and $\lambda_i(\Sigma)$ denotes i^{th} eigenvalue of Σ. Both the log and $\Sigma^{\frac{1}{2}}$ operators are well-defined on $\mathscr{P}_D^+$.

The disparity between the distance measurements is explained by the intrinsic curvature of $\mathscr{P}_D^+$. Sectional curvature is a formalism for describing curvature intrinsically to a space, such as $\mathscr{P}_D^+$. It is measured by sliding a plane tangentially to geodesic paths and measuring the twisting and turning undergone by that tangential plane. A detailed measurement of $\mathscr{P}_D^+$ shows that its sectional curvature is everywhere non-positive. This means that locally the space $\mathscr{P}_D^+$ is locally curved somewhat between a pseudo-sphere and Euclidean space. Alternatively this can be described by stating that $\mathscr{P}_D^+$ nowhere bulges out like a circle but is instead puckered in or flat.

A smooth subspace of Euclidean space having everywhere non-positive curvature when equipped with a Riemannian metric, and for which every pair of points can be joined by a unique distance minimizing geodesic is called a Cartan-Hadamard manifold. These spaces posses many well-behaved properties, as studied in [5], but for this discussion the most relevant property of Cartan-Hadamard manifolds to this chapter is the existence of a smooth map $\mathrm{Log}^g\left(\right)$ from $\mathscr{P}_D^+ \times \mathscr{P}_D^+$ onto $\mathbb{R}^d$. Here $\mathbb{R}^d$ is the Euclidean space of equal dimension to $\mathscr{P}_D^+$. For every fixed input, this map is infinitely differentiable, has an infinitely differentiable inverse and therefore puts $\mathscr{P}_D^+$ in smooth correspondence with $\mathbb{R}^d$. The map $\mathrm{Log}^g\left(\right)$ is called the Riemannian Logarithm. It is related to the distance between two covariance matrices through

$$\begin{aligned} d_g\left(\Sigma_1, \Sigma_2\right) =& \|\mathrm{Log}^g_{\Sigma_1}\left(\Sigma_2\right)\|_2, \\ Log_{\Sigma_1}(\Sigma_2) \triangleq& \Sigma_1^{\frac{1}{2}} log\left(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}\right)\Sigma_1^{\frac{1}{2}}. \end{aligned} \tag{3.4}$$

The Riemannian Exponential map, denoted by $\mathrm{Exp}^g\left(\right)$, is the inverse of $\mathrm{Log}^g\left(\right)$. The Riemannian Exponential map takes a covariance matrix Σ_1 and a tangential velocity vector v to Σ_1, and maps it to the covariance matrix Σ_2, found by traveling along $\mathscr{P}_D^+$ at the most efficient path beginning at Σ_1 with initial velocity v and stopping the movement after one time unit. Geodesics on $\mathscr{P}_D^+$ are obtained by scaling the initial velocity vector in the $\operatorname{Exp}^g()$ map, which is expressed as

$$\operatorname{Exp}^g_{\Sigma_1}(v) \triangleq \Sigma_1^{\frac{1}{2}} exp\left(\Sigma_1^{-\frac{1}{2}} Sym(v) \Sigma_1^{-\frac{1}{2}}\right) \Sigma_1^{\frac{1}{2}} \tag{3.5}$$

$$Sym(v) \triangleq \begin{pmatrix} v_1 & v_2 & \dots & v_D \\ v_2 & v_{D+1} & \dots & v_{2D-1} \\ \vdots & & \ddots & \vdots \\ v_D & & \dots & v_{\frac{D(D+1)}{2}}. \end{pmatrix},$$

where exp is the matrix exponential.

Returning to portfolio theory, any efficient portfolio in the sense of Equation (3.2), is entirely characterized by the log-returns, the non-degenerate covariance structure between the risky assets, and the risk-aversion level. The space parameterizing all the efficient portfolios, which will be called the Markowitz space after [74], has a natural geometric structure.

Definition 3.3.1 (Markowitz Space) *Let g_E^1, g_E^D, and g be the Euclidean Riemannian metrics on $\mathbb{R}$, $\mathbb{R}^D$, and the Riemannian metric on $\mathscr{P}_D^+$. The Riemannian manifold*

$$\left(\mathscr{M}_D^{Mrk}, g_D^{Mrk}\right) \triangleq \left(\mathbb{R} \times \mathbb{R}^d \times \mathscr{P}_D^+, g_E^1 \oplus g_E^2 \oplus g\right)$$

is called the (D-dimensional) Markowitz space.

Proposition 3.3.2 (Select Properties of the Markowitz Space)**.** The Markowitz space is connected, of non-positive curvature, and its associated metric space is complete. The distance function is

$$d^2_{Mrk}\left((\gamma_1, \mu_1, \Sigma_1), (\gamma_2, \mu_2, \Sigma_2)\right) \triangleq \|\gamma_2 - \gamma_1\|_2^2 + \|\mu_2 - \mu_1\|_2^2 + \left(\sum_{i=1}^{d} \lambda_i^2 \left(\log\left(\Sigma_2^{-\frac{1}{2}} \Sigma_1 \Sigma_2^{-\frac{1}{2}}\right)\right)\right)^2. \tag{3.6}$$

The Riemannian $\operatorname{Log}^g()$ and $\operatorname{Exp}^g()$ maps on $\mathscr{M}^{Mrk}$ are of the form

$$\begin{aligned} \operatorname{Exp}^g_{(\gamma_1, \mu_1, \Sigma_1)}\left((v_1, v_2, v_3)\right) &\triangleq \left(\gamma_1 + v_1, \mu_1 + v_2, \Sigma_1^{\frac{1}{2}} exp\left(\Sigma_1^{-\frac{1}{2}} Sym(v_3) \Sigma_1^{-\frac{1}{2}}\right) \Sigma_1^{\frac{1}{2}}\right) \\ \operatorname{Log}^g_{(\gamma_1, \mu_1, \Sigma_1)}\left((\gamma_2, \mu_2, \Sigma_2)\right) &\triangleq \left(\gamma_2 - \gamma_1, \mu_2 - \mu_1, \Sigma_1^{\frac{1}{2}} log\left(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}\right) \Sigma_1^{\frac{1}{2}}\right). \end{aligned} \tag{3.7}$$

Note that the Riemannian exponential and logarithm maps are defined everywhere and put $\mathscr{M}^{Mrk}$ in a smooth 1 to 1 correspondence with $\mathbb{R}^{1+D+\frac{D(D+1)}{2}}$.

Proof. The proof is deferred to the appendix. □

The Markowitz space serves as the prototypical example of the geometric spaces considered in the rest of this chapter, these are Riemannian manifolds, of non-positive curvature, for which every two points can be joined by a unique distance minimizing geodesic.

In the remainder of this chapter, all Riemannian manifolds will be Cartan-Hadamard manifolds. Cartan-Hadamard manifolds appear in many places in mathematical finance, for example in [60] the natural geometry associated with stochastic volatility models with two driving factors are Cartan-Hadamard manifolds.

On Cartan-Hadamard manifolds, such as the Markowitz space, there is no rigorously defined notion of conditional expectation. Therefore rigorous estimation intrinsic to these spaces' geometries is still a generally unsolved problem. We motivate this problem by discussing a few formulations of intrinsic conditional expectation and related empirical techniques present in the mathematical imaging literature.

The least-squares formulation of conditional expectation is

$$\mathbb{E}_{\mathbb{P}}[X_t|\mathscr{G}] \triangleq \operatorname{argmin}_{Z\in L^2_{\mathbb{P}}(\mathscr{G};\mathbb{R}^d)} \mathbb{E}_{\mathbb{P}}\left[\|X_t - Z\|_2^2\right].$$

Replacing the expected Euclidean distance by the expected intrinsic distance gives the typical formulation of a non-Euclidean conditional expectation. This formulation will be referred to as intrinsic conditional expectation.

Alternatively, estimates in a Riemannian manifold are made by locally linearizing the data using the Riemannian log map, performing the estimate in Euclidean space, and returning the data back onto the manifold. This type of methodology has been used extensively in the computer vision and mathematical imaging literature by [42, 53, 57, 3], and [94]. In [94], the authors empirically support estimating the intrinsic conditional expectation a following procedure which first linearizes the observation using the Riemannian Log transform, subsequently computes the conditional expectation in Euclidean space, and lastly returns the prediction onto the Riemannian manifold using the Riemannian Exp map.

This chapter provides a rigorous framework for the two methods described above, proves the existence of their optimum, and shows that both formulations agree. The rigorous formulation of the non-Euclidean filtering algorithm of [94] is used to derive non-Euclidean filtering equations. The non-Euclidean filtering problem is implemented and used to accurately forecast efficient portfolios by exploiting the geometry of the Markowitz space.

Empirical evidence for the importance of considering non-Euclidean geometry will be examined in the next section before developing a general theory of non-Euclidean conditional expectation in Section 3.4.

3.4 Non-Euclidean Conditional Expectations and Intrinsic Forecasting

Let $\mathbb{E}_{\mathbb{P}}[X_t|\mathscr{G}_t]$ denote the vector-valued conditional expectation in $\mathbb{R}^d$. Let $0 < \Delta < t$ and consider

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[X_t|\mathscr{G}_t] &= \lim_{\Delta\downarrow 0} \mathbb{E}_{\mathbb{P}}[X_t|\mathscr{G}_t] \\ &= \left(\lim_{\Delta\downarrow 0} \mathbb{E}_{\mathbb{P}}[X_{t-\Delta}|\mathscr{G}_{t-\Delta}] + \mathbb{E}_{\mathbb{P}}[(X_t - \mathbb{E}_{\mathbb{P}}[X_{t-\Delta}|\mathscr{G}_{t-\Delta}])|\mathscr{G}_t]\right) \\ &= \lim_{\Delta\downarrow 0} \mathrm{Exp}^g_{\mathbb{E}_{\mathbb{P}}[X_{t-\Delta}|\mathscr{G}_{t-\Delta}]}\left(\mathbb{E}_{\mathbb{P}}\left[\mathrm{Log}^g_{\mathbb{E}_{\mathbb{P}}[X_{t-\Delta}|\mathscr{G}_{t-\Delta}]}(X_t)\middle|\mathscr{G}_t\right]\right).\end{aligned} \tag{3.8}$$

The equality is obtained by taking the limit of a constant sequence and the second line it achieved using the $\mathscr{G}_t$-measurability of $\mathbb{E}_{\mathbb{P}}[X_{t-\Delta}|\mathscr{G}_{t-\Delta}]$ and the linearity of conditional expectation. The last line is obtained by using the fact that the Riemannian Exponential and Logarithm maps in Euclidean space respectively correspond to addition and subtraction.

Equation (3.8) expresses the conditional expectation at time t as moving from the conditional expectation at an arbitrarily close past time along a straight line with initial velocity given determined by the position of X_t and the last computed conditional expectation. The past time-period is made arbitrarily small by taking the limit of Δ to 0.

The last line is obtained by noting that in Euclidean space, the Riemannian exponential and logarithm maps correspond to vector addition and subtraction. Equation (3.8) may be generalized and taken to be the definition of conditional expectation in the general Cartan-Hadamard manifold setting.

In general, this definition will rely on a particular non-anticipative pathwise extension of a process. The definition of this pathwise extension is similar to the horizontal path extensions introduced in [32]. The extension $X_t^{\mathfrak{e}:t}$ of a process X_t holds the initial realized value X_0 constant back to $-\infty$ and the time t value constant all the way to ∞. Formally, $X_t^{\mathfrak{e}:t}$ is defined pathwise by

$$X_s^{\mathfrak{e}:t}(\omega) \triangleq \begin{cases} X_t(\omega) & t \leq s \\ X_s(\omega) & 0 \leq s \leq t \\ X_0(\omega) & s \leq 0 \end{cases}.$$

Figure 3.1: Extension of the process X_t.

The next assumption will be made to ensure that the initial conditional probability laws exist on $\mathscr{M}$.

Assumption 3.4.1 *Suppose that X_0 is $\mathscr{G}_0$-measurable and is absolutely continuous with respect to the intrinsic measure μ_g on $(\mathscr{M}, g)$. Denote its density by f_0, and assume that there exists at-least one point x_0 in $\mathscr{M}$ such that the integral $\int_{y \in \mathscr{M}} d_g^2(x_0, y) f_0(y) \mu_g(dy)$ is finite.*

Definition 3.4.2 (Dynamic Conditional Expectation) *Let X_t be an $(\mathscr{M}, g)$-valued càdlàg process and $\mathscr{G}_t$ be a sub-filtration of $\mathscr{F}_t$. The intrinsic conditional expectation of X_t given $\mathscr{G}_t$, denoted by X_t^g is defined to be the solution to the recursive system*

$$X_t^g \triangleq \begin{cases} \lim\limits_{n \mapsto \infty} \operatorname{Exp}^g_{X^g_{t-\frac{1}{n}}} \left(\mathbb{E}_{\mathbb{P}} \left[\operatorname{Log}^g_{X^g_{t-\frac{1}{n}}} (X_t^{\mathfrak{c}:t}) \middle| \mathscr{G}_t \right]^o \right) & if t > 0 \\ \operatorname*{argmin}\limits_{x \in \mathscr{M}} \int_{y \in \mathscr{M}} d_g^2(x, y) f_0(y) \mu_g(dy) & if\ t \leq 0, \end{cases} \tag{3.9}$$

where Y^o is the $\mathscr{G}_t$-optional projection.

The geometric intuition behind Equation (3.9) is that the current dynamic conditional expectation at time t is computed by first predicting the infinitesimal velocity describing the current state on $(\mathscr{M}, g)$ from the previous estimate at time $t - \frac{1}{n}$, and then moving across the infinitesimal geodesic along $(\mathscr{M}, g)$ in that direction. The computational implication of Equation (3.9) is that all the classical tools for computing the classical Euclidean conditional expectation may be used to compute the dynamic conditional expectation, once the Riemannian Exp and Riemannian Log maps are computed.

Lemma 3.4.3 (Existence of Initial Condition)**.** Under Assumption 3.4.1, X_0^g exists and is $\mathbb{P}$-a.s unique.

Proof. Under Assumption 3.4.1, [5, Exercise 5.11] guarantees the existence of X_0. □

Dynamic conditional expectation is an atypical formulation of non-Euclidean conditional expectation. Typically, non-Euclidean conditional expectation is defined as the $\mathscr{M}$-valued random element minimizing the expected intrinsic distance to X_t.

Following [69], by first isometrically embedding $(\mathscr{M}, g)$ into a large Euclidean space $\mathbb{R}^D$, the space $L^p_{\mathbb{P}}(\mathscr{F}; \mathscr{M})$ is subsequently defined as the subset of the Bochner-Lebesgue space $L^p_{\mathbb{P}}\left(\mathscr{F}; \mathbb{R}^D\right)$ consisting of the equivalence classes of measurable maps which are $\mathbb{P}$-a.s. supported on $\mathscr{M}$, and for which there exists some $\hat{X} \in \mathscr{M}$ for which

$$\left(\int_{\omega \in \mathscr{M}} d_g^p\left(X(\omega), \hat{X}\right) \mathbb{P}(d\omega) \right)^{\frac{1}{p}} < \infty. \tag{3.10}$$

The set $L^p_{\mathbb{P}}(\mathscr{F}; \mathscr{M})$ is a Banach manifold (see [82] for more general results).

Definition 3.4.4 (Intrinsic Conditional Expectation) *The intrinsic conditional expectation with respect to the σ-subalgebra $\mathscr{G}_t$ of $\mathscr{F}$ of an $\mathscr{M}$-valued stochastic process X_t is defined as the optimal Bayesian action*

$$\mathbb{E}^{g,p}_{\mathbb{P}}[X_t | \mathscr{G}_t] \triangleq \operatorname*{arginf}_{Z_t \in L^p_{\mathbb{P}}(\mathscr{G}_t; \mathscr{M})} \mathbb{E}_{\mathbb{P}}\left[d_g^p(Z_t, X_t)\right].$$

When $p = 2$, we will simply write $\mathbb{E}^g_{\mathbb{P}}[X_t|\mathscr{G}_t]$.

Intuition about intrinsic conditional expectation is gained by turning to the Markowitz space.

Example 3.4.5. Let $\gamma \geq 0$ be fixed and constant. Let $X_t \triangleq (\gamma, \mu_t, \Sigma_t)$ be a process taking values in the Markowitz space, Equation (3.6). Then the intrinsic conditional expectation of X_t given $\mathscr{G}_t$ is

$$\begin{aligned}\mathbb{E}^g_{\mathbb{P}}[(\gamma, \mu_t, \Sigma_t)|\mathscr{G}_t] = \underset{(\tilde{\mu}_t, \tilde{\Sigma}_t)\in \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})}{\operatorname{argmin}} & \mathbb{E}_{\mathbb{P}}\left[\|\mu_t - \tilde{\mu}_t\|_2^2\right] \\ & + \mathbb{E}_{\mathbb{P}}\left[\left(\sum_{i=1}^{d} \lambda_i^2 \left(\log\left(\sqrt{\tilde{\Sigma}_t}^{-1} \Sigma_t \sqrt{\tilde{\Sigma}_t}^{-1}\right)\right)\right)^2\right].\end{aligned} \tag{3.11}$$

The conditional expectation intrinsic to the Markowitz space seeks portfolio weights which give the most likely log-returns given the information in $\mathscr{G}_t$, while penalizing for the variance taken on by following that path.

In the case where Σ_t is independent of μ_t and Σ_t is $\mathscr{G}_t$-measurable, Equation (3.11) simplifies. Since μ_t does not depend on Σ_t and the latter is in $\mathbb{L}^2_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$, Σ_t may be substituted into the second term, which sets it to zero. Therefore, in this simplified scenario the least-squares property of Euclidean conditional expectation (see [65, Page 80]) that

$$\mathbb{E}^g_{\mathbb{P}}[(\gamma, \mu_t, \Sigma_t)|\mathscr{G}_t] = \underset{(\tilde{\mu}_t, \tilde{\Sigma}_t)\in \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})}{\operatorname{argmin}} \mathbb{E}_{\mathbb{P}}\left[\|\mu_t - \tilde{\mu}_t\|_2^2\right] = \mathbb{E}_{\mathbb{P}}[\mu_t|\mathscr{G}_t]. \tag{3.12}$$

There is a natural topology defined on $L^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$ which is characterized as being the weakest topology on which sequences of cádl'ag process process $\left\{X_{t-\frac{1}{n}}\right\}_{n\in\mathbb{N}}$ in $L^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$ converge to X_t in $L^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$ (see 3.7.2 for a rigorous discussion). For any two elements X and Y of $L^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$ with this topology, we will write

$$X \equiv Y,$$

if X and Y are indistinguishable in this topology. Intuitively, this means that they cannot be further separated in the topology. For example in $\mathbb{R}^D$ two points are indistinguishable if and only if they are equal, the same is true for example in metric spaces. Whereas in the space of measurable functions from $\mathbb{R}$ to itself which are square integrable equipped with its usual topology, two functions are inst indistinguishable if and only if they are equal on almost all points (see [67] for details on topological indistinguishability.)

Under mild assumptions, the dynamic conditional expectation and intrinsic conditional expectation agree on Cartan-Hadamard spaces as shown in the following theorem.

Theorem 3.4.6 (Unified Conditional Expectations)**.** Let X_t be an $\mathscr{M}$-valued process with càdlàg paths which is in $L^2(\mathscr{G}_t;\mathscr{M})$ for m-a.e. $t \geq 0$ and is such that Assumptions 3.4.1

and 3.7.7 hold. For $1 \leq p < \infty$, the intrinsic conditional expectation $\mathbb{E}_{\mathbb{P}}^{g}[X_t|\mathscr{G}_t]$ exists. Moreover, if $p = 2$, then

$$\mathbb{E}_{\mathbb{P}}^{g}\left[X_t^{\mathfrak{e}:t}\middle|\mathscr{G}_t\right]^{\mathfrak{e}:t} \equiv X_t^{g\,\mathfrak{e}:t}, \tag{3.13}$$

where the left-hand side of Equation (3.13) is the intrinsic conditional expectation and its right-hand side is the dynamic conditional expectation.

Theorem 3.4.6 justifies the particle filtering Algorithm of [94]. Before proving Theorem 3.4.6 and developing the required theory, a few implications and examples will be explored.

Theorem 3.4.6 has computational implications in terms of forecasting the optimal intrinsic conditional expectation using the dynamic conditional expectation. These implications are in the computable solution to the certain filtering problems.

Instead of discussing the dynamics of a coupled pair of $\mathscr{M}$-valued signal process X_t and observation processes Y_t intrinsically to $(\mathscr{M}, g)$, Theorem 3.4.6 justifies locally linearizing X_t and Y_t, then subsequently describing their Euclidean dynamics before finally returning them onto $\mathscr{M}$. More, specifically assume that

$$\begin{aligned}
\tilde{X}_t^i =& \int_0^t f^i(\tilde{X}_u^i)du + \int_0^t \beta^i(u, \tilde{X}_u^i)dB_u^i,\\
\tilde{Y}_t^i =& \int_0^u c^i(\tilde{X}_u^i, \tilde{Y}_u^i)du + \int_0^t \alpha^i(u, \tilde{Y}_u^i)dW_u^i,\\
\tilde{X}_t^i \triangleq& \langle \operatorname{Log}_{X_{t-\frac{1}{n}}^g}^g(X_t), e_i\rangle_{\mathbb{R}^d}\\
\tilde{Y}_t^i \triangleq& \langle \operatorname{Log}_{X_{t-\frac{1}{n}}^g}^g(Y_t), e_i\rangle_{\mathbb{R}^d}\\
B^i \perp\!\!\!\perp W^j, B^i \perp\!\!\!\perp& B^j, W^i \perp\!\!\!\perp W^j; i \neq j
\end{aligned} \tag{3.14}$$

where B_t and W_t are independent Brownian motions and $\tilde{X}_t^i, \tilde{Y}_t^i$ satisfy the usual existence and uniqueness conditions (see [22, Chapter 22.1] for example). This implies that X_t^i depends on only itself and that $\tilde{Y}_t^i$ depends only one X_t^i and itself. In particular, this implies that

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{X}_t^i\middle|\mathscr{G}_t\right] = \mathbb{E}_{\mathbb{P}}\left[\tilde{X}_t^i\middle|\mathscr{G}_t^i\right], \tag{3.15}$$

where $\mathscr{G}_t^i$ is the filtration generated only by $\tilde{Y}_t^i$. Using these dynamics, asymptotic local filtering equations for the dynamics of the dynamic conditional expectation $X_t^g \triangleq \mathbb{E}_{\mathbb{P}}^{g}\left[X_t^{\mathfrak{ert}}\middle|\mathscr{G}_t\right]$ in terms of $\eta_t^{\mathfrak{ert}}$ can be deduced and are summarized in the following Corollary of Theorem 3.4.6.

Corollary 3.4.7 (Asymptotic Non-Euclidean Filter)**.** Let $\mathscr{M} = \mathbb{R}^d$, denote the i^{th} coordinate of X_t^g by $X_t^{g,i}$, and suppose Assumptions 3.4.1 and 3.7.7 as well as the assumptions on X_t and Y_t made in [22, Chapter 22.1]. If X_t^g is $\mathbb{P} \otimes m$-a.e. unique, a version of the intrinsic conditional expectation $X_t^g \triangleq \mathbb{E}_{\mathbb{P}}^g\big[X_t^{\mathfrak{ert}}\big|\mathscr{G}_t\big]$ must satisfy the SDE

$$\begin{aligned} X_t^{g,i} = & \lim_{\Delta \mapsto 0^+} \langle \mathrm{Exp}^g_{X^g_{t-\Delta}} \left(\sum_{i=1}^d X_0^i \right), e_i \rangle_{\mathbb{R}^d} \\ & + \int_0^t \left[\sum_{i=1}^d \left\langle \frac{\partial}{\partial x_i} \mathrm{Exp}^g_{X^g_{t-\Delta}} \left(\sum_{i=1}^d X_t^i \right), e_i \right\rangle_{\mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\big[f^i(X_u)\big|\mathscr{G}_u^i\big] \right. \\ & \left. + \frac{1}{2} \sum_{i,j=1}^d \left\langle \frac{\partial^2}{\partial x_i x_j} \mathrm{Exp}^g_{X^g_{t-\Delta}} \left(\sum_{i=1}^d X_t^i \right), e_i \right\rangle_{\mathbb{R}^d} \Xi_u^{i,j} \right] du \\ & + \int_0^t \sum_{i=1}^d \left\langle \frac{\partial}{\partial x_i} \mathrm{Exp}^g_{X^g_{t-\Delta}} \left(\sum_{i=1}^d X_t^i \right), e_i \right\rangle_{\mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\big[f^i(X_u)\big|\mathscr{G}_u^i\big] dV_u, \end{aligned} \tag{3.16}$$

where the limit is taken with respect to the metric topology on $\mathbb{L}^2_{\mathbb{P}}\left(\mathscr{G}_t; \mathscr{M}\right)$ and the processes $\Xi_t^{i,j}$ are defined by

$$\Xi_t^{i,j} \triangleq \left(\mathbb{E}_{\mathbb{P}}\left[\tilde{X}_u^i c^i \middle| \mathscr{G}_u^i \right] - \mathbb{E}_{\mathbb{P}}\left[\tilde{X}_u^i \middle| \mathscr{G}_u^i \right] \mathbb{E}_{\mathbb{P}}\left[c^i(\tilde{X}_u^i) \middle| \mathscr{G}_u^i \right] \right) \left(\mathbb{E}_{\mathbb{P}}\big[X_u^j c^j \big| \mathscr{G}_u^j\big] - \mathbb{E}_{\mathbb{P}}\big[X_u^j \big| \mathscr{G}_u^j\big] \mathbb{E}_{\mathbb{P}}\big[c^i(X_u^j) \big| \mathscr{G}_u^j\big] \right).$$

Proof. The proof is deferred to the appendix. □

Corollary 3.4.7 gives a way to use classical Euclidean filtering methods to obtain arbitrarily precise approximations to an SDE for the non-Euclidean conditional expectation. It is two-fold recursive as it requires the previous non-Euclidean conditional expectation $X^g_{t-\Delta}$ to compute the next update. In practice, X_t^g will be taken to be the previous asymptotic estimate.

The next section investigates the numerical performance of our non-Euclidean filtering methodology.

3.5 Numerical Performance

To evaluate the empirical performance of the filtering equations of Corollary 3.4.7, 1000 successive closing prices ending on June 30^{th} 2018, for the Apple and Google stock are considered. The unobserved signal process X_t is the covariance matrix between the closing prices at time t and the observation process Y_t, is the empirical covariance matrix generated on 7-day moving windows.

The signal and observation processes X_t and Y_t are assumed to be coupled by Equation (3.14). The functions f^i and c^i are modeled as being deterministic linear functions and β^i, α^i are modeled as being constants.

$$\begin{aligned}
\tilde{X}_t^i =& \int_0^t A^{i,i}\tilde{X}_u^i du + \int_0^t C^{i,i} dB_u^i \\
\tilde{Y}_t^i =& \int_0^t H^{i,i}\tilde{X}_u^i du + \int_0^t K^{i,i} dW_u^i, \\
\tilde{Y}_t^i \triangleq& \langle \operatorname{Log}^g_{X^g_{t-\frac{1}{n}}}(Y_t), e_i\rangle_{\mathbb{R}^d} \\
\tilde{X}_t^i \triangleq& \langle \operatorname{Log}^g_{X^g_{t-\frac{1}{n}}}(X_t), e_i\rangle_{\mathbb{R}^d} \\
B^i \perp\!\!\!\perp W^j, B^i \perp\!\!\!\perp& B^j,\, W^i \perp\!\!\!\perp W^j; i \neq j
\end{aligned} \tag{3.17}$$

where A, B, C, H, and K are invertible diagonal matrices non-zero determinant.

Analogous dynamics are for the benchmark methods, ensuring that the Kalman filter is the solution to the stochastic filtering problem. The values of A, B, C, H, and K are estimated using maximum likelihood estimation.

Both the classical (KF) and proposed methods (N-KF) are also benchmarked against the non-Euclidean Kalman filtering algorithm of [57] (N-KF-int). This algorithm proposes that the dynamics of X_t^i and Y_t^i be modeled in Euclidean space using the transformations

$$\begin{aligned}
\tilde{X}_t^i &\triangleq \left\langle \operatorname{Log}^g_{\bar{\Sigma}}(X_t), e_i \right\rangle_{\mathbb{R}^d}, \\
\tilde{Y}_t^i &\triangleq \left\langle \operatorname{Log}^g_{\bar{\Sigma}}(Y_t), e_i \right\rangle_{\mathbb{R}^d}, \\
\bar{\Sigma} &\triangleq \underset{\Sigma \in \mathscr{P}_D^+}{\operatorname{argmin}} \frac{1}{15}\sum_{j=1}^{15} d_g^2(\Sigma, Y_{t_j}),
\end{aligned}$$

where $\bar{\Sigma}$ is the intrinsic Riemannian Barycenter (see [8] for details properties of the intrinsic mean), and the Riemannian Log and Exp functions are derived from the geometry of $\mathscr{P}_D^+$ and not of $\mathscr{M}^{Mrk}$, 15 was chosen by sequential-validation. Unlike equations (3.14), the Riemannian log and exp maps are always performed about the same point $\bar{\Sigma}$ and do not update. This will be reflected in the estimates whose performance progressively degrades over time.

The Riemannian Barycenter $\bar{\Sigma}$, is computed both intrinsically and extrinsically using the first 15 empirical covariance matrices. The extrinsic Riemannian Barycenter on $\mathscr{P}_D^+$ is defined to be the minimizer of

$$\bar{\Sigma}^{ext} \triangleq \operatorname{Exp}^g_{Y_1}\left(\frac{1}{15}\sum_{j=1}^{15} \operatorname{Log}^g_{Y_1}(Y_j)\right).$$

The extrinsic formulation of the Kalman filtering algorithm of [57] (N-KF-ext), models the linearized signal and observation processes by

$$\begin{aligned}
\tilde{X}_t^i &\triangleq \left\langle \operatorname{Log}^g_{\bar{\Sigma}^{ext}}(X_t), e_i \right\rangle_{\mathbb{R}^d}, \\
\tilde{Y}_t^i &\triangleq \left\langle \operatorname{Log}^g_{\bar{\Sigma}^{ext}}(Y_t), e_i \right\rangle_{\mathbb{R}^d}.
\end{aligned}$$

The length of the moving window was calibrated in a way which maximized the performance of the standard Kalman-filter performed componentwise (EUC). The choice of 15 observed covariance matrices used to compute the intrinsic mean was made by sequential validation on the initial 25% of the data. The findings are reported in the following table.

Table 3.1: Efficient Portfolio One-Day Ahead Forecasts

	$\gamma = 0$		$\gamma = 0.5$		$\gamma = 1$	
	ℓ^2	ℓ^∞	ℓ^2	ℓ^∞	ℓ^2	ℓ^∞
EUC	0.780	0.695	0.811	0.723	0.834	0.746
N-KF	0.101	0.087	0.162	0.140	0.162	0.140
N-KF-int	0.550	0.493	0.671	0.608	0.671	0.608
N-KF-extr	0.865	0.828	0.969	0.931	0.969	0.931

Table 3.1 examines the one day ahead predictive power by evaluating the accuracy of the forecasted portfolio weights. N-KF is our proposed algorithm. N-KF-int is the algorithm of [57] based on the methods of [43], without the unscented transform. N-KF-int computes the Riemannian $\operatorname{Log}_\mu^g(\cdot)$ and $\operatorname{Log}_\mu^g(\cdot)$ maps where μ is the intrinsic mean to the first 15 observed covariance matrices and N-KF-ext is the same with the mean computed extrinsically (see [8] for a detailed study of intrinsic and extrinsic means on Riemannian manifolds). The one-day ahead predicted weights are evaluated both against the next day's optimal portfolio weights using both the ℓ^2 and ℓ^∞ norms for portfolios with the risk-aversion levels $\gamma = 0, 0.5, 1$.

According to each of the performance metrics, the forecasted efficient portfolios using the intrinsic conditional expectation introduced in this chapter performs best. An interpretation is that the Euclidean method disregards all the geometric structure, and that the competing non-Euclidean methods do not update their reference points for the $\operatorname{Exp}^g()$ and $\operatorname{Log}^g()$ transformations. The failure to update the reference point results in progressively degrading performance. This effect is not as noticeable when the data is static as in [43, 42], however the time-series nature of the data makes the need to update the reference point for the transformations numerically apparent.

Table 3.2: Comparison of Covariance Matrix Prediction

	Frobenius	Max Modulus	Infinity	Spectral	Intrinsic
EUC	0.001	0.001	0.001	0.001	2.069
N-KF	0.0001	0.0001	0.0001	0.0001	0.843
N-KF-int	0.003	0.003	0.003	0.003	2.797
N-KF-extr	0.0004	0.0003	0.0005	0.0004	14.395

Table 3.2 examines the covariance matrix forecasts of all four methods directly. The performance metrics considered are the Frobenius, Maximum Modulus, Infinite and Spec-

tral matrix norms of the difference between the forecasted covariance matrix and the realized future covariance matrix of the two stocks closing prices. An intrinsic performance metric, the distance on $\mathscr{P}_D^+$ between the forecasted covariance of the stock prices matrix and the realized future covariance matrix is also considered.

In Table 3.2, all the non-Euclidean methods out-perform the component-wise classical Euclidean forecasts of the one-day ahead predicted covariance matrix. The prediction of covariance matrices is less sensitive than that of the efficient portfolio weights, this is most likely due to $(\bar{1}^\star\Sigma^{-1}\bar{1})^{-1}$ term appearing in Equation (3.2) which is sensitive to small changes due to the observably small value of Σ.

Table 3.3: Bootstrapped Adjusted Confidence Intervals for Performance Metrics

	95.l	Mean	95 U
Frobenius	0.001	0.001	0.001
Max Modulus	0.0004	0.001	0.001
Infinity	0.001	0.001	0.001
Spectral	0.001	0.001	0.001
Intrinsic	1.793	2.069	2.359

(a) Euclidean Kalman Filter

	95.l	Mean	95 U
Frobenius	0.00005	0.0001	0.0001
Max Modulus	0.00004	0.0001	0.0001
Infinity	0.0001	0.0001	0.0001
Spectral	0.00005	0.0001	0.0001
Intrinsic	0.407	0.843	1.900

(b) Asymptotic Non-Euclidean Kalman Filter

	95.l	Mean	95 U
Frobenius	0.001	0.003	0.007
Max Modulus	0.001	0.003	0.008
Infinity	0.001	0.003	0.008
Spectral	0.001	0.003	0.007
Intrinsic	2.549	2.797	3.064

(c) Non-Updating Intrinsic Barycenter

	95.l	Mean	95 U
Frobenius	0.0003	0.0004	0.001
Max Modulus	0.0002	0.0003	0.0004
Infinity	0.0003	0.0005	0.001
Spectral	0.0003	0.0004	0.001
Intrinsic	11.721	14.395	17.128

(d) Non-Updating Extrinsic Barycenter

Tables 3.3 and 3.2 report 95% confidence intervals about the estimated mean of the one-day ahead mean error of each respective distance measure. The error distribution of the performance metrics is non-Gaussian according to the Shapiro-Wilks test performed for normality (see [90] for details). The bootstrap adjusted confidence (BAC) interval method of [29] is used instead to non-parametrically generate the 95%-confidence intervals. The BAC method is chosen since it does not assume that the underlying distribution is Gaussian, it corrects for bias, and it corrects for skewness in the data. The bootstrapping was performed by re-sampling $10,000$ times from the realized error distributions of the performance metrics.

Tables 3.1 and 3.4 show that the N-KF method is the most accurate and has the lowest variance amongst all the methods according to the Frobenius, Maximum modulus, infinity, and spectral matrix norms. However the variance of the intrinsic distance is not the lowest, but it's variance is. This is interpreted as a bias-variance trade-off.

Table 3.4: Performance Metrics of Portfolio Weights One-Day Ahead Predictions

		95 L	Mean	95 U
$\gamma = 0$	ℓ^2	0.610	0.780	0.962
	ℓ^∞	0.547	0.695	0.842
$\gamma = 0.5$	ℓ^2	0.649	0.811	0.995
	ℓ^∞	0.558	0.723	0.894
$\gamma = 1$	ℓ^2	0.669	0.834	1.004
	ℓ^∞	0.600	0.746	0.908

(a) Euclidean Kalman Filter

		95 L	Mean	95 U
$\gamma = 0$	ℓ^2	0.069	0.101	0.171
	ℓ^∞	0.056	0.087	0.146
$\gamma = 0.5$	ℓ^2	0.084	0.162	0.367
	ℓ^∞	0.067	0.140	0.292
$\gamma = 1$	ℓ^2	0.080	0.162	0.347
	ℓ^∞	0.072	0.140	0.335

(b) Non-Euclidean Kalman Filter

		95 L	Mean	95 U
$\gamma = 0$	ℓ^2	0.429	0.550	0.691
	ℓ^∞	0.388	0.493	0.604
$\gamma = 0.5$	ℓ^2	0.516	0.671	0.879
	ℓ^∞	0.480	0.608	0.807
$\gamma = 1$	ℓ^2	0.514	0.671	0.848
	ℓ^∞	0.467	0.608	0.792

(c) Non-Updating Intrinsic Barycenter

		95 L	Mean	95 U
$\gamma = 0$	ℓ^2	0.731	0.865	1.013
	ℓ^∞	0.692	0.828	0.966
$\gamma = 0.5$	ℓ^2	0.791	0.969	1.113
	ℓ^∞	0.755	0.931	1.078
$\gamma = 1$	ℓ^2	0.803	0.969	1.115
	ℓ^∞	0.761	0.931	1.084

(d) Non-Updating Extrinsic Barycenter

Tables 3.2 and 3.1 reflect that the forecasting performance for the efficient portfolio weights of the N-KF method is more accurate than the others. This is again seen in the lower bias and tighter 95% confidence interval reported in the Table 3.4.

The numerics reflect the importance of incorporating relevant geometry to mathematical finance problems and that the manner in which it is incorporated is a subtle matter.

The next section summarizes the contributions made in this chapter.

3.6 Summary

The need to incorporate relevant geometric information into probabilistic estimation procedures was supported by the efficient portfolio weight prediction at the start of the chapter. Non-Euclidean filtering was seen to outperform traditional Euclidean filtering methods with the estimates presenting a lower bias and generally having tighter confidence intervals.

The numerical procedure was justified by Theorem 3.4.6 which proved the equivalence and existence of common formulations of intrinsic conditional expectation to transformations of a specific Euclidean conditional expectation. These results were established using the variational-calculus theory of Γ-convergence introduced in [24] and subsequently developed by [13] by temporarily passing through the larger $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.; \mathscr{M})$-spaces. To our knowledge, these are novel proofs techniques within the field of mathematical finance and applied probability theory.

A central consequence of Theorem 3.4.6 is the potential to write down computable stochastic filtering equations for the dynamics of the intrinsic conditional expectation on $(\mathscr{M}, g)$ using classical Euclidean filtering equations. Our results differed from those of [79], [31], or [64] since dynamics for an intrinsic conditional expectation are forecasted and not dynamics of the Euclidean conditional expectation of a function of a non-Euclidean signal and/or observation process. Likewise, out results did not rely on the Le Jan-Watanabe connection as those of [85] and the only computational bottleneck may be to compute the Riemannian Logarithm and Riemannian Exponential maps. However, these are readily available in many well-studied geometries not discussed in this chapter, for example the hyperbolic geometry used to study the λ-SABR models in [60].

Many other naturally occurring spaces in mathematical finance have the required properties for the central theorems of this chapter to apply. For instance the geometry of two-factor stochastic volatility models developed in [60] do. The techniques developed here can find applications to that geometry and other relevant geometries in mathematical finance and could find many other areas of applied probability theory where standard machine learning methods have been used extensively.

3.7 Appendix

In this, section technical proofs or results from the main body of this chapter are given.

Proof of Proposition 3.3.2. In general, for any three Riemannian manifolds $(M, g^M), (N, g^N)$, $(\tilde{N}, g^{\tilde{N}})$, there is a natural bundle-isomorphism $T(M \times N \times \tilde{N}) \cong TM \times TN \times T\tilde{N}$ (see [63, Section 2.1] for a discussion on vector bundles). Under this identification, define the metric on $T(M \times N \times \tilde{N})$ as follows for each $(p, q, r) \in M \times N \times \tilde{N}$.

$$g^{M\times N\times\tilde{N}}_{(p,q,r)} : T_{(p,q,r)}(M \times N \times \tilde{N}) \times T_{(p,q,r)}(M \times N \times \tilde{N}) \to \mathbb{R},$$

$$((x_1, y_1, z_1), (x_2, y_2, z_3)) \mapsto g^M_p(x_1, x_2, z_3) + g^N_q(y_1, y_2, z_3) + g^{\tilde{N}}_r(y_1, y_2, z_3).$$

Let $\nabla^{M\times N\times\tilde{N}}$ be the Levi-Civita connection on the product of two Riemannian manifolds, then $\nabla^{M\times N} = \nabla^M + \nabla^N + \nabla^{\tilde{N}}$. Therefore for if $\gamma^M, \gamma^N, \gamma^{\tilde{N}}$ are geodesics on M, N, $\tilde{N}$ respectively then

$$\nabla^{M\times N\times\tilde{N}}\dot{(\gamma^M, \gamma^N, \gamma^{\tilde{N}})} = \nabla^M\dot{\gamma}^M + \nabla^N\dot{\gamma}^N + \nabla^{\tilde{N}}\dot{\gamma}^{\tilde{N}} = 0 + 0 + 0 = 0,$$

whence $M \times N \times \tilde{N}$-valued curve $t \mapsto \left(\gamma^M(t), \gamma^N(t)\right)$ is a geodesic on the product Riemannian manifold. Therefore geodesics, and hence the $\operatorname{Exp}^g()$ as well as the $\operatorname{Log}^g()$ maps can be expressed component-wise on the product Riemannian manifold. Particularizing

$M, N, \tilde{N}$ to $\mathbb{R}, \mathbb{R}^D$, and $\mathscr{P}_D^+$ implies that the Markowitz space is a well-defined Riemannian manifold. The formula for d^{Mrk} is just the formula for the product metric between metric spaces.

Using the natural isomorphism discusses above, the sectional curvature of the product Riemannian is the sum of the sectional curvatures. Since Euclidean space has 0-sectional curvature and $\mathscr{P}_D^+$ has non-positive sectional curvature (see[12]), then the Markowitz space has non-positive sectional curvature.

The general linear group $GL_D(\mathbb{R})$ has two connected components corresponding to the matrices with negative or positive determinant. Since $\mathscr{P}_D^+$ is a subset comprised of matrices with strictly positive eigenvalues, its elements all have a strictly positive determinant. Therefore $\mathscr{P}_D^+$ is simply connected. Since each of the component spaces of the Markowitz space is geodesically complete (see [12] for the statement concerning $\mathscr{P}_D^+$) the Hopf-Rinow Theorem implies that the associated metric space is complete. The non-positive curvature of the Markowitz space together with the Cartan-Hadamard Theorem imply that the Riemannian exponential map at every point of the Markowitz space is a diffeomorphism onto the $\mathbb{R}^{1+D+\frac{D(D+1)}{2}}$, where $\frac{D(D+1)}{2}$ is the dimension of the $\mathscr{P}_D^+$. The dimension is obtained by counting the entries on and above the main diagonal of a *symmetric* matrix. □

Proof of Corollary 3.4.7. Denote the conditional expectation $\mathbb{E}_{\mathbb{P}}\left[\tilde{X}_t^i \middle| \mathscr{G}_t\right]$ by X_t^i. The filtering Equations of [22, Remark 22.1.15] imply that each of the conditional mean of each locally linearized coordinate processes $\tilde{X}_t^i$ given the filtration $\mathscr{G}_t{}^i$ is

$$X_t^i = \mathbb{E}_{\mathbb{P}}\left[X_0^i \middle| \mathscr{G}_0^i\right] + \int_0^t \mathbb{E}_{\mathbb{P}}\left[f^i(X_u) \middle| \mathscr{G}_u^i\right] du + \int_0^u \left(\mathbb{E}_{\mathbb{P}}\left[\tilde{X}_u^i c^i \middle| \mathscr{G}_u^i\right] - \mathbb{E}_{\mathbb{P}}\left[\tilde{X}_u^i \middle| \mathscr{G}_u^i\right] \mathbb{E}_{\mathbb{P}}\left[c^i(\tilde{X}_u^i) \middle| \mathscr{G}_u^i\right]\right) dV_u \tag{3.18}$$

where the innovations processes V_t^i and the optional projections of c^i are defined by

$$\begin{aligned} V_t^i &\triangleq \int_0^t \alpha(u, Y_u)^{-1} dY_u - \int_0^t \alpha^i(s, \tilde{Y}_u^i)^{-1} \left(\hat{c}^i(\omega, u, \tilde{Y}_u^i)\right) du \\ \hat{c}^i(\omega, t, y) &\triangleq \mathbb{E}_{\mathbb{P}}\left[c^i(t, X_t, y) \middle| \mathscr{G}_t{}^i\right], \end{aligned}$$

(see [22, Chapter 22.10] for more details on the innovations process and [22, Chapter 7.6] for more details on optional projections).

Abbreviate $\mathbb{E}_{\mathbb{P}}^g[X_t|\mathscr{G}_t]$ by Xt^g Applying the Itô-Lemma to the (smooth) function

$$x \mapsto \left\langle \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^d x\right), e_i \right\rangle_{\mathbb{R}^d}$$

to the process $\sum_{i=1}^{d} Xte_i$ yields

$$
\begin{aligned}
\langle \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_t\right), e_i\rangle_{\mathbb{R}^d} =& \langle \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_0\right), e_i\rangle_{\mathbb{R}^d} \\
&+ \int_0^t \sum_{i=1}^{d} \left\langle \frac{\partial}{\partial x_i} \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_t\right), e_i \right\rangle_{\mathbb{R}^d} d\hat{X}^i_t \\
&+ \frac{1}{2}\int_0^t \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i x_j} \left\langle \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_t\right), e_i \right\rangle_{\mathbb{R}^d} d[\hat{X}^i, \hat{X}^j]_t \\
=& \langle \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_0\right), e_i\rangle_{\mathbb{R}^d} \\
&+ \int_0^t \left[\sum_{i=1}^{d} \left\langle \frac{\partial}{\partial x_i} \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_t\right), e_i \right\rangle_{\mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\left[f^i(X_u)\middle|\mathscr{G}^i_u\right] \right.\\
&+ \left. \frac{1}{2} \sum_{i,j=1}^{d} \left\langle \frac{\partial^2}{\partial x_i x_j} \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_t\right), e_i \right\rangle_{\mathbb{R}^d} \Xi^{i,j}_u \right] du \\
&+ \int_0^t \sum_{i=1}^{d} \left\langle \frac{\partial}{\partial x_i} \operatorname{Exp}^g_{X^g_{t-\Delta}}\left(\sum_{i=1}^{d} X^i_t\right), e_i \right\rangle_{\mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\left[f^i(X_u)\middle|\mathscr{G}^i_u\right] dV_u,
\end{aligned}
\tag{3.19}
$$

where the processes $\Xi^{i,j}_t$ is defined by

$$
\Xi^{i,j}_t \triangleq \left(\mathbb{E}_{\mathbb{P}}\left[\tilde{X}^i_u c^i \middle| \mathscr{G}^i_u\right] - \mathbb{E}_{\mathbb{P}}\left[\tilde{X}^i_u \middle| \mathscr{G}^i_u\right]\mathbb{E}_{\mathbb{P}}\left[c^i(\tilde{X}^i_u)\middle|\mathscr{G}^i_u\right]\right)\left(\mathbb{E}_{\mathbb{P}}\left[X^j_u c^j|\mathscr{G}^j_u\right] - \mathbb{E}_{\mathbb{P}}\left[X^j_u|\mathscr{G}^j_u\right]\mathbb{E}_{\mathbb{P}}\left[c^i(X^j_u)|\mathscr{G}^j_u\right]\right).
$$

The results follow by applying Theorem 3.4.6 and the Optional Projection [22, Theorem 7.6.2]. □

We return to the proof of Theorem 3.4.6. This will require moving to a slightly larger space where things become more manageable.

Definition 3.7.1 (The $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$ Spaces) *Let $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$ denote the subset of the disjoint union $\coprod_{t\in\mathbb{R}} L^p_{\mathbb{P}}(\mathscr{G}_{t\vee 0};\mathscr{M})$ consisting of all families $\{X_t\}_{t\in\mathbb{R}}$ satisfying*

$$
t \mapsto X_t(\omega) \in D(\mathbb{R};\mathscr{M}, d_g);\mathbb{P} - a.s.
$$

The natural topology on $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$ induced by these operations will be denoted by τ_0.

Refine the topology on $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$ into the coarsest topology on $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$ satisfying

(i) τ is no coarser than the topology on $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$,

(ii) $\{Z^n_t\}_{n\in\mathbb{N}}$ converges to an element of $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})$ if and only if it converges to Z_t with respect to τ and $\{Z^n_{t-\frac{1}{n}}\}_{n\in\mathbb{N}}$ converges to Z_t in τ.

The one-point compactification of $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ is denoted by $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$, the new point, denoted by ∞ is called the escape point. Elements of $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ are called eternal processes and are denoted by $Z_\cdot$.

Remark 3.7.2. Since $L^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$ is a topological subspace of $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ then it inherits a relative topology. The indistinguishability discussed in Theorem 3.4.6 is with respect to this relative topology.

Remark 3.7.3 (Escape Point)**.** The escape point ∞ is interpreted as describing the eternal processes which either fail the finiteness condition of Equation (3.10) or fail to take values in $\mathscr{M}$ at a given point in time $\mathbb{P}$-a.s.

Remark 3.7.4 (Points in $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ are Eternal and May Explode)**.** Every element of $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ is indexed by the time t which takes values in $\mathbb{R}$ and not only in $[0,\infty)$. The time $t=0$ is interpreted as when the observer first gained information of the process. In this way the part above time $t=0$ is a process which may explode arbitrary number of times and the part below is interpreted as a *pre-history* to an observer at time $t=0$. In this way, processes in $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ are thought of as eternal. Note that the eternal process $X^{\mathfrak{e}:T}_t$ is $\mathscr{G}_{t\wedge T}$-adapted.

Lemma 3.7.5 (Existence)**.** *The space* $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ *exists and is unique up to homeomorphism. Moreover,* $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ *is dense in* $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$.[1]

Proof. The uniqueness and the density of $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ are in $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ the properties of the one-point compactification.

Let τ_0 denote the topology on $\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$. Let $\mathscr{T}$ denote the set of topologies containing τ_0 and for which (ii) holds. $\mathscr{T}$ is non-empty since the discrete topology satisfies both (i) and (ii). Since the intersections of topologies is again a topology (see [67, page 55 Problem A.a]) then the topology on $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ exists and is $\cap_{\tau\in\mathscr{T}}\tau$. Existence follows from the existence of the one point-compactification of the topological space $\left(\tilde{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M}), \cap_{\tau\in\mathscr{T}}\tau\right)$. □

The Riemannian Log and Riemannian Exponential maps extend to a correspondence

[1] These spaces also exhibit universal properties that follow directly from those of the Alexandroff one-point compactification used to construct them, but they are besides the central focus of this chapter and so will not be discussed here.

between $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ and $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathbb{R}^d)$. To see this consider the maps

$$\begin{aligned}
&\mathrm{LOG}^g\,() : \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M}) \times \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M}) \to \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathbb{R}^d)\\
&\qquad\qquad \mathrm{LOG}^g_{Z_\cdot}(Y_\cdot) \mapsto \left\{\begin{cases} 0 & : Z_\cdot = Y_\cdot = \infty \\ \mathrm{Log}^g_{Z_t}(Y_t) & : Z_\cdot \text{ and } Y_\cdot \neq \infty \\ \infty & : \text{else} \end{cases}\right\}_{t\in\mathbb{R}},\\
&\mathrm{EXP}^g\,() : \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M}) \times \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathbb{R}^d) \to \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})\\
&\qquad\qquad \mathrm{EXP}^g_{Z_\cdot}(Y_\cdot) \mapsto \left\{\begin{cases} \mathrm{Exp}^g_{Z_t}(Y_t) & : Z_\cdot \neq \infty \text{ and } Y_\cdot \neq \infty \\ \infty & : \text{else} \end{cases}\right\}_{t\in\mathbb{R}}.
\end{aligned}$$

Both these maps collapse to component-wise post-composition by $\mathrm{Log}^g\,()$ (resp. $\mathrm{Exp}^g\,()$) if the eternal process $Z_\cdot$ never hits ∞.

The map $d_g(\cdot,\cdot)$ also induces a map from $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M}) \times \mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ into $[0,\infty]$. The induced map, denoted by $D_g(\cdot,\cdot)$ is defined by

$$Z_\cdot \mapsto \begin{cases} d_g(Z_t,X_t) & : \text{ if } X_\cdot \text{ and } Z_\cdot \neq \infty \\ \infty & : \text{else.} \end{cases}$$

All of these collapse to their usual definitions when the escape point is not encountered. They will play a key technical role for the remainder of this chapter.

Lemma 3.7.6. For every $1 \leq p < \infty$ and every sub-filtration $\mathscr{G}_\cdot$ of $\mathscr{F}_\cdot$, the functionals

$$F_n(Z_\cdot) \triangleq \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}(Z_t) - \mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}(X_t)\right\|_2^p\right]dt,$$

Γ-converges to the functional

$$F(Z_\cdot) \triangleq \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[D_g^p(Z_t,X_t)\right]dt$$

on $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$.

Proof. Let $Z_\cdot$ be an element of $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$, $\{Z^n_\cdot\}_{n\in\mathbb{N}}$ be a sequence converging to $Z_\cdot$ in $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_t;\mathscr{M})$ and $X_\cdot$ be an element of $\mathbb{L}^p_{\mathbb{P}}(\mathscr{F}_t;\mathscr{M})$. For every $t\in\mathbb{R}$, Reverse Fatou's Lemma implies that

$$\overline{\lim_{n\mapsto\infty}} \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(Z^n_t) - \mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(X_t)\right\|_2^p\right]dt \tag{3.20}$$

$$\leq \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\overline{\lim_{n\mapsto\infty}}\left\|\mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(Z^n_t) - \mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(X_t)\right\|_2^p\right]dt \tag{3.21}$$

The continuity of $\|\cdot\|_2^2$, $\mathrm{Log}^g\,()$, and the $\mathbb{P}$-a.s. continuity of the path $t \mapsto Z_t(\omega)$ and the choice of topology on $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$ implies that the limit on the RHS of Equation (3.21) exists and can be computed to be

$$\begin{aligned}
&\overline{\lim_{n\mapsto\infty}} \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(Z^n_t) - \mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(X_t)\right\|_2^p\right] dt\\
&\leq \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z_t}(Z_t) - \mathrm{LOG}^g_{Z_t}(X_t)\right\|_2^p\right] dt \qquad (3.22)\\
&= \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z_t}(X_t)\right\|_2^p\right] dt\\
&= \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[D^p_g(Z_t, X_t)\right] dt. \qquad (3.23)
\end{aligned}$$

Here the fact that $\mathrm{LOG}^g_x(x) = 0$ was used along with the relationship between the Riemannian Logarithm and the Riemannian metric, as exemplified in $\mathscr{P}^+_D$ by Equation (3.4). Analogously, by the ordinary Fatou's Lemma

$$\int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[D^p_g(Z_t, X_t)\right] dt \leq \underline{\lim}_{n\mapsto\infty} \int_{t\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(Z^n_t) - \mathrm{LOG}^g_{Z^n_{t-\frac{1}{n}}}(X_t)\right\|_2^p\right] dt. \qquad (3.24)$$

By the definition of Γ-convergence, F is the Γ-limit of the functionals F_n on $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$. □

Assumption 3.7.7 *Both $X^g_\cdot, X_\cdot \neq \infty$.*

The proof of Theorem 3.4.6 relies on a result of central interest in the theory of Γ-convergence. This results [75, Theorem 7.8], is also called the Fundamental Theorem of Γ-convergence in [13, Theorem 2.10] in the metric space formulation. It may be reformulated as stating that if a sequence of functionals F_n Γ-converges to a functional on a compact topological space[2] X, then it must satisfy

$$\min_{x\in X} \Gamma\text{-}\lim_{n\mapsto\infty} F_n(x) = \lim_{n\mapsto\infty} \inf_{x\in X} F_n(x). \qquad (3.25)$$

Proof of Theorem 3.4.6. Lemma 3.4.3 established the required Γ-convergence between the discussed functionals on the compact topological space $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot;\mathscr{M})$; this gives existence of the intrinsic conditional expectation $\mathbb{E}^{g,p}_{\mathbb{P}}[X_t|\mathscr{G}_t]$, for every $1 \leq p < \infty$.

For the remainder of this proof, p will be equal to 2. Equation (3.11) will be established by an uncountable strong induction, indexed by the totally ordered set $(\mathbb{R}, \leq)$. By the definitions of X^g_t and $\mathbb{E}^{g,p}_{\mathbb{P}}[X_t|\mathscr{G}_0]$if follows that

$$X^g_0 = \mathbb{E}^g_{\mathbb{P}}[X_0|\mathscr{G}_0] = Z_0.$$

[2] The assumption of compactness is a special case of the statement which only requires *equicoercivity*.

Since ${X^{g}}^{\mathfrak{e}:0}_{t} = X_0^g$ and $\mathbb{E}_{\mathbb{P}}^{g}[X_t^{\mathfrak{e}:0}|\mathscr{G}_t] = \mathbb{E}_{\mathbb{P}}^{g}[X_0|\mathscr{G}_0]$ for every $t \leq 0$, the base case of the (uncountable) strong induction hypothesis is established.

Suppose that for every $t \leq T$, $X_t^g = \mathbb{E}_{\mathbb{P}}^{g,p}[X_t|\mathscr{G}_t]^{\mathfrak{e}:t}$. It follows from the Γ-convergence of F_n to F, that

$$\begin{aligned}\min_{Z.\in\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathscr{M})}\int_{t=0}^{T}\mathbb{E}_{\mathbb{P}}\left[D_g^p(Z_t, X_t^{\mathfrak{e}:T})\right]dt =& \min_{Z.\in\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathscr{M})}\int_{t\in\mathbb{R}}\mathbb{E}_{\mathbb{P}}\left[D_g^p(Z_t, X_t^{\mathfrak{e}:T})\right]dt \\ =& \lim_{n\mapsto\infty}\inf_{Z.\in\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathscr{M})}\int_{t\in\mathbb{R}}\mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}(Z_t) - \mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}\left(X_t^{\mathfrak{e}:T}\right)\right\|_2^p\right]dt \\ =& \lim_{n\mapsto\infty}\inf_{Z.\in\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathscr{M})}\int_{t=0}^{T}\mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}(Z_t) - \mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}\left(X_t^{\mathfrak{e}:T}\right)\right\|_2^p\right]dt.\end{aligned} \tag{3.26}$$

Here the fact that $X_t^{\mathfrak{e}:T}$ is identical above T and below 0 was used. The non-negativity of the integrands on of both sides of Equation (3.26) and the monotonicity of integration implies that the LHS of Equation (3.26) must minimize $\mathbb{E}_{\mathbb{P}}\left[D_g^p(Z_t, X_t^{\mathfrak{e}:T})\right]$ for m-a.e. value of t between 0 and the current time T. Therefore by the definition of intrinsic conditional expectation, the left-hand side of Equation (3.26) is minimized by the eternal process

$$\mathbb{E}_{\mathbb{P}}^{g}\left[X_t^{\mathfrak{e}:T}\middle|\mathscr{G}_t\right]^{\mathfrak{e}:T}. \tag{3.27}$$

Likewise, the right-hand side of Equation (3.26) is minimized by the minimizers of

$$\mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}(Z_t) - \mathrm{LOG}^g_{Z_{t-\frac{1}{n}}}\left(X_t^{\mathfrak{e}:T}\right)\right\|_2^p\right].$$

Since $t - \frac{1}{n} < t$, the induction hypothesis may be, applied hence

$$Z_{t-\frac{1}{n}} = {X^{g}}^{\mathfrak{e}:t}_{t-\frac{1}{n}} = \mathbb{E}_{\mathbb{P}}^{g}\left[X^{\mathfrak{e}:t}_{t-\frac{1}{n}}\middle|\mathscr{G}_{t-\frac{1}{n}}\right]. \tag{3.28}$$

Equation (3.28) implies that $\mathrm{LOG}^g_{{X^{g}}^{\mathfrak{e}:t}_{t-\frac{1}{n}}}(X_t^{\mathfrak{e}:t})$ no longer enters into the optimization as a variable. The correspondence between $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathscr{M})$ and $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathbb{R}^d)$ defined by the map $\mathrm{LOG}^g_{X^{\mathfrak{e}:t}_{t-\frac{1}{n}}}()$ gives

$$\begin{aligned}&\inf_{Z.\in\mathbb{L}^2_{\mathbb{P}}(\mathscr{G}.;\mathscr{M})}\mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{X^{\mathfrak{e}:t}_{t-\frac{1}{n}}}(Z_t) - \mathrm{LOG}^g_{X^{\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X_t^{\mathfrak{e}:t}\right)\right\|_2^p\right] \\ =&\inf_{Z.\in\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}.;\mathbb{R}^d)}\mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{X^{\mathfrak{e}:t}_{t-\frac{1}{n}}}(Z_t) - \mathrm{LOG}^g_{X^{\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X_t^{\mathfrak{e}:t}\right)\right\|_2^p\right] \qquad (3.29) \\ =&\mathbb{E}_{\mathbb{P}}\left[\mathrm{LOG}^g_{X^{\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X_t^{\mathfrak{e}:t}\right)\middle|\mathscr{G}_t\right], \qquad (3.30)\end{aligned}$$

where the least-squares property of the L^2-formulation of conditional expectation (see [65, page 80]) was used. Since the Riemannian Logarithm is a diffeomorphism, the change of variables may be undone. Hence

$$\begin{aligned}
&\underset{Z_{\bullet}\in\mathbb{L}^2_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathscr{M})}{\operatorname{arginf}}\ \mathbb{E}_{\mathbb{P}}\left[\left\|\mathrm{LOG}^g_{X^{g\,\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(Z_t\right)-\mathrm{LOG}^g_{X^{g\,\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X^{\mathfrak{e}:t}_t\right)\right\|_2^p\right]\\
=&\mathrm{EXP}^g_{X^{g\,\mathfrak{e}:t}_{t+\frac{1}{n}}}\left(\underset{\tilde{Z}_{\bullet}\in\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_{\bullet};\mathbb{R}^d)}{\operatorname{arginf}}\ \mathbb{E}_{\mathbb{P}}\left[\left\|\tilde{Z}_t-\mathrm{LOG}^g_{X^{g\,\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X^{\mathfrak{e}:t}_t\right)\right\|_2^p\right]\right)\\
=&\mathrm{EXP}^g_{X^{g\,\mathfrak{e}:t}_{t+\frac{1}{n}}}\left(\underset{\tilde{Z}_t\in L^p_{\mathbb{P}}(\mathscr{G}_t;\mathbb{R}^d)}{\operatorname{arginf}}\ \mathbb{E}_{\mathbb{P}}\left[\left\|\tilde{Z}_t-\mathrm{LOG}^g_{X^{g\,\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X^{\mathfrak{e}:t}t_t\right)\right\|_2^p\right]\right)\\
=&\mathrm{EXP}^g_{X^{g\,\mathfrak{e}:t}_{t+\frac{1}{n}}}\left(\mathbb{E}_{\mathbb{P}}\left[\mathrm{LOG}^g_{X^{g\,\mathfrak{e}:t}_{t-\frac{1}{n}}}\left(X^{\mathfrak{e}:t}_t\right)\middle|\mathscr{G}_t\right]\right).
\end{aligned}\tag{3.31}$$

Recombining equations (3.26), (3.27), and (3.31) yields

$$\mathbb{E}^g_{\mathbb{P}}\left[X^{\mathfrak{e}:T}_T\middle|\mathscr{G}_t\right]^{\mathfrak{e}:T}=\lim_{n\mapsto\infty}\mathrm{EXP}^g_{X^{g\,\mathfrak{e}:t}_{T-\frac{1}{n}}}\left(\mathbb{E}_{\mathbb{P}}\left[\mathrm{LOG}^g_{X^{g\,\mathfrak{e}:T}_{T-\frac{1}{n}}}\left(X^{\mathfrak{e}:T}_T\right)\middle|\mathscr{G}_t\right]\right)^{\mathfrak{e}:T}=X^{g\,\mathfrak{e}:T}_T.\tag{3.32}$$

Assumption 3.7.7 implies that D_g, $\mathrm{LOG}^g\,()$, $\mathrm{EXP}^g\,()$ reduce to their usual counterparts. This completes the induction and establishes Theorem 3.4.6. □

The proof of Theorem 3.4.6 showed how passing through the larger space $\mathbb{L}^2_{\mathbb{P}}\left(\mathscr{G}.;\mathscr{M}\right)$ conclusions about the smaller $L^2_{\mathbb{P}}(\mathscr{G}_t,\mathscr{M})$ spaces could be made.

4. Arbitrage-Free Regularization

We introduce an unsupervised and non-anticipative machine learning algorithm which is able to detect and remove arbitrage from a wide variety of models. In this framework, fundamental results and techniques from risk-neutral pricing theory such as NFLVR, market completeness, and changes of measure are given an equivalent formulation and extended to models which are deformable into arbitrage-free models. We use this scheme to construct a meta-algorithm which ensures that a wide range of factor estimation schemes return arbitrage-free estimates and incorporate this additional information into their estimation procedure. We show that, using our meta-algorithm, we are able to produce more accurate estimates of forward-rate curves, specifically at the long-end. The spread between a model and its arbitrage-free regularization is then used to construct a mis-pricing detection or classification algorithm, which is in turn used to develop a pairs trading strategy. Our theory provides a sound theoretical foundation for a risk-neutral pricing theory capable of handling models which potentially admit arbitrage but which can be deformed into arbitrage-free models.

4.1 Introduction

This chapter introduces a novel machine learning framework founded on stochastic calculus and arbitrage-pricing theory which extracts financial features from relevant time-series data in order to learn an arbitrage-free model describing the data. The framework we introduce is built on the following modeling principles, analogous to those discussed in [54].

A modeling framework should produce *interpretable* models, in that the model learned should either directly depend on factors which can easily be understood, or be as close as possible to a model whose factors are interpretable. The models produced from the framework should be *describable*. That is they should rely only on a finite number of factors and their evolution must be described by a finite number of SDEs. The modeling approach should be *data-driven*. By this we mean that the learned model should be dynamic and continuously updating according to the statistical and financial features of any relevant incoming data. This self-updating property should happen with minimal input from the user and, therefore, produce a low modeling bias. All models produced using the procedure should not conflict with the efficient market hypothesis. More precisely, it will be required that any model produced by the framework should be *self-correcting*, removing the potential for any arbitrage opportunities in an unsupervised fashion. The modeling framework should be *natural* in that it applies to, relates, explains, and transfers meaning between a variety of asset classes with minor changes. This requirement not only ensures that the modeling framework is widely applicable, but that it captures core properties which are fundamental to most market's behavior irrespectively of the market asset's particularities.

A commonly employed approach for obtaining interpretable models is to use factor models to either model the asset price or an auxiliary process. It was shown in [39] that, for very large classes of factor models for the term-structure of a zero-coupon bond introduce arbitrage opportunities into the bond market. Since many empirically chosen factor models allow arbitrage these models must disregard certain subtle financial features of the data and therefore are not as data-driven as they may appear. Although employing empirically chosen factor models may result in an interpretable and reusable picture of the market, this approach does not meet all of our modeling principles.

In [28], it is argued that the absence of arbitrage and the related subtleties in the data are not of great consequence. The authors empirically demonstrated that the predictive power of certain arbitrage-free interest-rate models have comparable performance to their factor-model approximation at the short end of the curve. However, a detailed inspection of these results confirms that the performance of an empirically chosen factor model rapidly degrades for maturities at the long end of the forward-rate curve. Therefore, a model which lacks a low-dimensional representation, interpretability, or admits arbitrage opportunities is not practically tractable or theoretically viable. To meet these three modeling requirements, we introduce the arbitrage-free regularization algorithm. This algorithm extracts financial features from the realized time-series and uses them to remove arbitrage opportunities from the empirical factor model. This predictive advantage will be illustrated on the long-end of the forward-rate curve.

Our framework can be interpreted as producing a dynamically self-correcting model with a factor model at its core. More precisely, our approach will begin by taking an interpretable empirically chosen factor model ϕ and deforming it into a factor-like model $\Phi_t(\phi)$. The predictable function Φ_t can be interpreted as dynamically deforming the factor model ϕ in an optimal way to remove arbitrage opportunities. The deformation Φ_t will be optimal in the sense that it minimizes the objective function

$$D(\phi, \Phi_t(\phi); x_t) + AF(\Phi_t(\phi)), \tag{4.1}$$

where D is a distance measuring how far the deformed factor model $\Phi_t(\phi)$ is from the interpretable empirical factor model ϕ on the realized data's path up to time t, that path will be denoted by x_t. Here AF is a penalty detecting the presence of arbitrage opportunities in $\Phi_t(\phi)$. A deformation which is far from the data or the empirical factor model would either be poorly performing or uninterpretable. In contrast, a model which admits arbitrage fails to be theoretically sound and will be shown to have poorer predictive powers than the closest arbitrage-free model. The *Arbitrage-Free Regularization* problem of Equation (4.1) is defined as the model selection criteria for determining $\Phi_t(\phi)$.

As a consequence of our methodology, we are able to use the spread between the optimal model $\Phi_t(\phi)$ and the model ϕ to detect and classify market mispricings. This is then used to construct a pairs trading strategy which trades pairs of assets whose price frequently fluctuate between over and under-priced states. The pairs trading strategy developed as an application of the theory serves as an alternative to the usual pairs trading strategy of [19] which can be deployed on pairs assets exhibiting co-integration.

The Heath-Jarrow-Morton (HJM) framework of [58], provides a flexible modeling framework which describes the evolution of bond prices for every maturity through an infinite dimensional system of SDEs. We will refer to the extension of the HJM framework to other asset classes, as in [32, 18], the Generalized-HJM (GHJM) framework and reserve the acronym HJM for the bond setting. A central point of interest of the GHJM models, other than their flexibility, is that the existence of arbitrage opportunities for the GHJM model representing the price of a bond, or pair of a call option and stock have been characterized in [58, 66, 17] through specifications on the driving process drift. However, there still is no general characterization of the existence of arbitrage opportunities to the general GHJM models of [32, 18].

The infinite dimensionality of GHJM models make them computationally intractable. The computational intractability of the infinite dimensionality of this approach is resolved in [9], where finite-factor models could be found to be consistent with an infinite dimensional GHJM model. These consistent finite dimensional models are called finite dimensional realizations of an GHJM model (FDR-HJM), and the dimension reduction is achieved by turning to non-Euclidean methods. FRD-HJMs characterize the HJM model which can be represented as a finite-factor model. These additional benefits come with the cost that the characterization of the existence of arbitrage opportunities is lost in the case of call options. Under certain assumptions, a characterization of equivalent local martingale measures for FDR-HJMs representing the price of a zero-coupon bond is provided by [39]. Furthermore, a characterization of the existence of arbitrage opportunities for FDR-HJMs modeling zero-coupon bonds whose factor model is an exponential polynomial is given in [38].

Another drawback of the FDR-HJM models is that the introduction of factor models makes them parametric. The parametric nature of FDR-HJM models is a further drawback since it introduces model selection bias introduced by the user's choice of model as opposed to being learned non-parametrically. An alternative to the HJM and FDR-HJM approach is the non-parametric principal component analysis approach. The FDR-HJM method is computationally tractable, low-dimensional and generally interpretable, it's factor model component is static. The static factor model component of the FDR-HJM methodology is not suited to the dynamic nature of the financial landscape and is responsible for many of the drawbacks associated to FDR-HJM models.

Flow models provide an alternative extension to the FDR-HJMs extension of GHJM models. Flow models are designed to be a general modeling framework on which Arbitrage-Free Regularization is defined. Arbitrage-Free Regularization is an unsupervised learning method for learning a what we will call a flow model from an empirical factor model. The method predictably deforms the empirical factor model until the objective measure $\mathbb{P}$ becomes risk-neutral for the deformed model. Arbitrage-free regularized models are semi-parametric since they are non-parametric deformations of parametric models.

The comparison of modeling approaches is summarized in this table.

Approach	AF-Reg	GHJM	FDR-HJM	PCA
N. Factors	**Finite**	N/A	**Finite**	**Finite**
N. Diffusions	**Finite**	Infinite	**Finite**	N/A
Path Dependence	**Yes**	No	No	No
Non-Euclidean Features	**Yes**	No	**Yes**	No
Estimation	**Semi-Param.**	NA	Parametric	**Non-Parametric**
Characterization of No-Arbitrage	**Yes**	Partial	Partial	No.

Table 4.1: AF-Reg abbreviates arbitrage-free regularization of flow models.

Section 4.3 introduces the class of models, called flow models, on which arbitrage-free regularization is defined. This section focuses on providing example of flow models and characterizes flow models which do not admit arbitrage opportunities. The mathematics of regularization is subsequently developed in Section 4.4. Section 4.5 focuses of the theoretical implications of arbitrage-free regularization. Specifically, central results to risk-neutral pricing theory such as NFLVR, market completeness and the minimal martingale measure are shown to be have an equivalent reformulations in terms of specific arbitrage-free regularization problems. Corrections to models which admit arbitrage, such as the Arbitrage-Free Nelson-Siegel correction of the Nelson-Siegel model, are shown to be model specific solutions to particular arbitrage-free regularization problems. The arbitrage-free regularization framework provides a methodology, which is not model specific, for removing arbitrage from models. Section 4.5 extends the classical risk-neutral pricing theory beyond semi-martingales to models which admit arbitrage, but are deformable into models arbitrage-free flows models.

Computational aspects of arbitrage-free regularization are considered in Section 4.6, with numerical illustrations set within the fixed-income setting. It is observed that the arbitrage-free regularized models outperform their empirical factor counterparts. Their improved performance is interpreted as the incorporation of financial features into the learned model and this interpretation is validated through information theoretic methods. In particular, it is observed that this newly assimilated information yields more accurate price forecasts for bonds maturing in at-least 20 years. As an application, an algorithm exploiting the spread between the arbitrage-free regularization of a factor model to detect and classify types of mis-pricing in the bond market is introduced. With the addition of Hidden Markov Model (HMMs), a low-risk trading strategy simultaneously shorting overpriced bonds, and going long on underpriced bonds is introduced as situational alternative to classical pairs trading strategies.

The next section discusses the conventions and notation used in this chapter.

4.2 Preliminaries and Notation

For the remainder of this chapter, we assume that all processes are defined on a stochastic base $(\Omega, \mathscr{G}_t, \mathscr{G}, \mathbb{P})$, with $\mathbb{P}$-complete left-continuous filtration on which a Brownian motion

W_t exists. For the remainder of this chapter, $X_t(u)$ will denote the price of an asset depending on the parameter u.

We denote the complete left-continuous sub-filtration of $\mathscr{G}_t$, jointly generated by $X_t(u)$ and W_t by $\mathscr{F}_t$, and the complete left-continuous sub-filtration of $\mathscr{G}_t$ generated by $X_t(u)$ (resp. W_t) by $\mathscr{F}_t^X$ (resp. $\mathscr{F}_t^W$). We will denote by μ_t, a cádlág $\mathscr{F}_t^X$-predictable process taking values in the set of σ-finite Borel measures on $\mathcal{U}$.

Moreover $(\mathscr{M}, g_t)$ will denote a Riemannian manifold with time-dependent connection, $D([0,t];\mathbb{R}^D)$ (resp. $D([0,t];S_d^+)$) the space of paths with values in $\mathbb{R}^d$ (resp. S_d^+, the set of $d \times d$-positive-definite matrices), and $\mathcal{U}$ be a Borel subset of $\mathbb{R}^D$, for some $D \geq d$. We will denote the Lebesgue measure on $[0,\infty)$ by m and by $\mathscr{B}_t$ the Borel σ-algebra on $[0,\infty)$. A family of functionals $(F_t)_{t\in[0,\infty)}$ will be abbreviated by F_t and said to be non-anticipative if is non-anticipative (resp. predictable) with respect to the paths of $X_t(u)$ and of W_t. By $H_\mu^s(\mathcal{U})$ we mean the Sobolev space $W_{\mu_t}^{s,2}(\mathcal{U})$ on $\mathcal{U}$. By $L_\mu^2(H^s(\mathcal{U})_\mu; H^s(\mathcal{U})_\mu)$ we mean the square-integrable Bochner-Lebesgue space.

The next section introduces and presents examples of flow models. Flow models which are arbitrage-free are also characterized.

4.3 Arbitrage-Free Flows

Arbitrage-free flows are dynamically updating models motivated by empirical factor models. The definition is presented and will be followed by a series of examples emphasizing the use and necessity of each of a flow model's defining components.

Definition 4.3.1 (Flow Model) *Let $X_t \triangleq \{X_t(u)\}_{u\in\mathcal{U}}$ be a family of price processes such that there exists (F, ϕ, β_t, g_t) such that for $\mathbb{P} \otimes m \otimes \mu$-a.e. (ω, t, u) in $\Omega \times [0,\infty) \times \mathcal{U}$*

$$X_t(u) = F_t\left(\phi(t, \beta_t, u)|u\right) \tag{4.2}$$

A flow model for $X_t(u)$, denoted by (F, ϕ, β_t), is a triple satisfying the regularity conditions 4.8.11, 4.8.12, and 4.8.13 found in the Appendix and characterized by

(i) ***Stochastic Factors:*** *An $\mathscr{M}$-valued semi-martingale β_t. They are the dynamic-factors for the empirical factor model and $\mathscr{M}$ is their domain of definition,*

(ii) ***Empirical Factor Model****: A $\mathscr{F}_t$-predictable process $\{\phi(t,\beta,u)\}_{t\geq 0}$ taking values in the set of Borel-measurable maps from $\mathscr{M} \times \mathcal{U}$ to $\mathbb{R}$ which are twice differentiable in their $\mathscr{M}$ component. The predictable time-inhomogeneity of the factor model $\phi(t,\beta,u)$ represents the ability to update/recalibrate the factor model as new data is received,*

(iii) ***Encoding Functional:*** *A non-anticipative functional $F_t(\cdot|u) : D([0,t];\mathbb{R}) \times D([0,t];S_1^+) \times \mathcal{U} \to \mathbb{R}$. This encodes the dynamic factor model into the family of asset prices being modeled,*

(iv) ***Geometry's Factor:*** *The family of Riemannian metrics $\{g_t\}_{t\in[0,\infty)}$ on $\mathscr{M}$ capture the potentially non-Euclidean features present in the evolution of the stochastic-factors β_t.*

Remark 4.3.2. In the remainder of this chapter simplified dynamics are assumed on the stochastic factor process which rely on stochastic differential geometry to be stated (See Appendix 4.8.2 for details on stochastic differential geometry). For the remainder of this chapter the stochastic factor process will be assumed to be g-horizontal. When the context is clear, both the stochastic factor process and its g-stochastic anti-development will be denoted by β_t. The dynamics of the g-horizontal anti-development of the stochastic factor process will be assumed to solve the diffusion process

$$\beta_t = \beta_0 + \int_0^t \mu(s, \beta_s)ds + \int_0^t \sigma(s, \beta_s)dW_s, \tag{4.3}$$

where W_t is a Brownian motion on the Euclidean space of the same dimension as $\mathscr{M}$. The g-horizontal lift of the stochastic factor process β_t to the orthonormal frame bundle $\mathscr{O}(\mathscr{M})$ with initial frame $\xi_0 = \Xi$ will be denoted by ξ_t.

The set $\mathcal{U}$ is called the parameter space of the assets prices $X_t(u)$, $\mathscr{M}$ represents the factor's domain of definition, and the process $F(\phi(t, \beta_t, u)|u)$ is called the flow model's realization. The collection of all flow models with the same encoding functional F and stochastic factors β_t, but with possible different empirical factor models ϕ, is denoted by $C_\mu(F, \beta_t)$.

Definition 4.3.3 (Arbitrage-Free Flow) *If μ-a.e. member of the family of price processes $\{X_t(u)\}_{u\in\mathcal{U}}$ satisfies NFLVR for μ-a.e. u in $\mathcal{U}$, then (F, ϕ, β) is said to be an **Arbitrage-Free Flow**.*

Arbitrage-free regularization begins with a flow model and learns the closest arbitrage-free flow to it. Flow models can be used to model the price of zero-coupon bonds, call options, portfolios of stocks, amongst other asset prices.

Example 4.3.4 (Instantaneous Forward-Rate Curve)**.** The time t price of a zero-coupon bond with maturity T, denoted by $B(t, T)$, depends on the instantaneous interest rate in effect at that time. This interest-rate is called the short-rate r_t and is related to the bond price through

$$B(t, T) = \mathbb{E}\left[e^{-\int_t^T r_s ds} \mid \mathscr{F}_t^r\right],$$

where $\mathscr{F}_t^r$ is the filtration generated by the short rate r_t. Modeling bond prices through using short-rate models lacks the flexibility to easily calibrate to the realized initial term structure of interest to the bond price as well as incorporating the term-structure of interest into the bond price. This motives the framework of [58] which models $B(t, T)$ as a function of all the future instantaneous interest rates between times t and T, as observed from the current time t.

This family of future interest rates are denoted by $f(t,T)$ and the map $T \mapsto f(t,T)$ defines a stochastic process called the instantaneous forward-rate curve (FRC), which is related to the price of a zero-coupon bond and the short-rate by

$$B(t,T) = e^{-\int_t^T f(t,s)ds};\ f(t,t) = r_t. \tag{4.4}$$

Every maturity T defines a particular point on the FRC, and each individual FRC point's evolution is described by a SDE. This gives a description of $f(t,T)$ as system of infinitely many SDEs makes working directly with FRC models computationally intractable.

In practice, this is typically overcome by modeling $f(t,T)$ by a factor model $\phi(t,\beta,T)$. For example, consider the following flexible extension of the typically used Nelson-Siegel model, abbreviated NS

$$\begin{aligned}
\phi\left(t,\beta,T\right) =& \sum_{i=1}^{N} \beta_i \varphi_i \quad ; N \geq 3 \\
\varphi_1 =& 1 \\
\varphi_2 =& \frac{\left[1 - \exp\left(-T/\tau\right)\right]}{T/\tau} \\
\varphi_3 =& \left(\frac{\left[1 - \exp\left(-T/\tau\right)\right]}{T/\tau} - \exp\left(-T/\tau\right) \right) \\
\varphi_i =& \frac{\left[1 - \exp\left(-T^{k_i}/\tau\right)\right]}{T^{k_i}/\tau}; i > 3, k_i > 0.
\end{aligned} \tag{4.5}$$

The loadings β_1, β_2, β_3, and τ are interpreted as level, slope, curvature, and shape parameters, respectively [30]. To capture the dynamic nature of the market the factors $\beta_1, \beta_2, \beta_3$ are often taken to be stochastic. The additional factors $\{\phi_i\}_{i>3}$ capture various decay rates of the FRC.

The price of a zero-coupon bond modeled within the HJM framework, modeled by a factor model such as the extended Nelson-Siegel family of equation (4.5), has the following representation as a flow model:

(i) The manifold $(\mathscr{M}, g_t)$ is an open subset of the d-dimensional Euclidean space on which the factors β of ϕ are defined,

(ii) The set of possible parameters $\mathcal{U} = [0,\infty)$ are the possible times of maturity of the bond $B(t,T)$,

(iii) The stochastic factors β_t are the factors of ϕ which, in the extended Nelson-Siegel example, captured the stochastic evolution of level, slope and curvature parameters, and decay rates,

(iv) The factor model ϕ is a low-dimensional factor model for the FRC,

(v) The functional F encoding the FRC into the price of a zero-coupon bond $B(t,T)$ is

$$F(\cdot|u) \triangleq e^{-\int_t^u \cdot ds}.$$

Hence the FDR-HJM

$$B(t,T) = e^{-\int_t^u f(t,u)du} = e^{-\int_t^u \phi(t,\beta_t,T)ds} = F(\phi(t,\beta_t,u)|u). \tag{4.6}$$

Therefore the price of a zero-coupon bond $B(t,T)$ can be represented by the flow model $\left(e^{-\int_t^T \cdot ds}, \phi(t,T,\beta), \beta_t\right)$.

Remark 4.3.5 (Relationship to FDR-HJMs)**.** If ϕ is assumed to be deterministic and constant in time, then equation (4.6) is precisely the definition of an FDR-HJM as introduced in [9]. In this way, FDR-HJMs are particular cases of flow models.

Example 4.3.6 (Portfolio Value)**.** Let $(S_t^1, \ldots, S_t^d)$ be a set of risky assets, assume than an equivalent martingale measure $\mathbb{Q}$ exists and that the log-returns of the risky assets follow a d-dimensional diffusion process. Any self-financing portfolio with positions $w = (w_t^1, \ldots, w_t^d)$ on the risky assets is valued at a future time T via the risk-neutral pricing formula

$$V_T(w) \triangleq \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^d w_t^i S_t^i \mid \mathscr{F}_t^S\right],$$

where $\mathscr{F}_t^S$ is the filtration generated by S_t. The value process V_T can be represented by the following flow model:

(i) The manifold $(\mathscr{M}, g_t)$ is the d-dimensional Euclidean space,

(ii) The set of possible parameters $\mathcal{U} = [0,\infty)^d$ are taken to be the weights $u \triangleq w$,

(iii) The stochastic factors β_t are defined to be the log-returns

$$\beta_t^i \triangleq \ln\left(\frac{S_t^i}{S_0}\right),$$

(iv) The factor model ϕ will be taken to be the map aggregating the positions in each stock

$$\phi(t,\beta,u) \triangleq \sum_{i=1}^d w_t^i S_0^i e^{\beta^i},$$

(v) The functional F encoding the factor model ϕ into the value of the portfolio $V_T(u)$ is the path-dependent functional

$$F(\cdot|u) \triangleq \mathbb{E}_{\mathbb{Q}}\left[\cdot \mid \sigma(S_\star)_t\right].$$

Hence the portfolio value can be represented by the flow model $\left(\mathbb{E}_{\mathbb{Q}}\left[\cdot \mid \sigma(S_\star)_t\right], \sum_{i=1}^d w_t^i S_0^i e^{\beta_t^i}, (\ln\left(\frac{S_t}{S_0}\right))\right)$ with realization

$$V_T(w) = \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{d} w_t^i S_t^i \mid \sigma(S_\star)_t\right] = F(\phi(t, \beta_t, u)|u).$$

Example 4.3.7 (Stochastic Local Volatility)**.** In the Black-Scholes framework, the price of a European call option on a stock is characterized by the Black-Scholes formula

$$\begin{aligned} C(t, S_t, T, K, \sigma) &= N(d_1)S_t - N(d_2)Ke^{-r(T-t)} \\ d_1 &= \frac{1}{\sigma\sqrt{T-t}}\left[\ln\left(\frac{S_t}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)\right] \\ d_2 &= d_1 - \sigma\sqrt{T-t} \\ dS_t &= \sigma S_t dW_t. \end{aligned} \tag{4.7}$$

This formula depends only on the current price of the stock S_t, the time T at which the option matures, its pre-agreed upon strike price K, and the stock's volatility σ. All these quantities are known at time T except for the volatility σ which must be estimated. The volatility implied by the realized market option prices $\tilde{C}$, is typically found by first viewing $C(t, S_t, T, K, \sigma)$ solely a function of the volatility σ and subsequently solving the inverse problem

$$C(t, S_t, T, K, \sigma) - \tilde{C} = 0, \tag{4.8}$$

for the volatility that best explains the realized market prices $\tilde{C}$.

Solving the inverse problem of equation (4.8) yields different values of σ for different strikes and maturity times. Options with lower strike prices tend to have higher implied volatilities than their high stike price counterparts which leads to the well known *volatility smile* phenomenon. The surface obtained by solving this inverse problem for each strike and maturity time is called the implied volatility surface of S_t.

Models in which the volatility σ is allowed to be stochastic have superior empirical performance over their deterministic counterparts and their implied volatility surfaces takes on more natural shapes (see [60, Chapter 6.1] for a more in depth discussion on the subject) and provide a partial solution to this issue. Under a stochastic volatility model (SVM), the assets' risk-neutral dynamics is assumed to satisfy the following SDE

$$\begin{aligned} dS_t &= \sigma_t dW_T \\ d\sigma_t &= \mu(t, \sigma_t)dt + \nu(t, t, \sigma_t)dB_t \\ \mathbb{E}\left[[W, B]_t\right] &= \rho; \rho \in [-1, 1], \end{aligned} \tag{4.9}$$

where σ_t is the instantaneous volatility of S_t at time t called the instantaneous spot-volatility, analogously to r_t.

Just as the instantaneous spot-rate r_t discussed in Example (4.3.4), was inflexible and difficult to calibrate to daily observed market prices, the instantaneous spot-volatility σ_t

suffers from the similar shortcomings. In [32], both these issues are overcome by modeling the entire volatility surface, denoted by $\sigma(t, T, K)$, as a stochastic function of T and K. Analogously to equation (4.4), it is shown in [18] that for such a model the price of Call options, denoted henceforth by $C(t, T, K)$ is the unique solution to the initial value problem

$$\begin{aligned}
\partial_t C(\tau, K) =& \left(\frac{K^2 \nu(t, \tau, K)}{2}\right) \partial_K^2 C(\tau, K) \\
C(0, K) =& (S_t - K)^+ \\
dS_t =& \sigma(t, S_t, K) dW_t \\
\sigma =& \nu^2 \\
\tau \triangleq& T - t,
\end{aligned} \tag{4.10}$$

where $\nu = \sigma^2$, called the stochastic variance surface[1], is often used instead of σ for notational and computational convenience.

Analogously to the FRC setting, the dimensionality of the stochastic volatility surface $\sigma(t, T, K)$ leads to computational intractability. Analogously to the FRC setting, the curse of dimensionality motivates the use of factor models for the stochastic variance surface in place of modeling $\nu(t, T, K)$ directly as an infinite of SDEs. A common *globally parameterized* example is the SVI-JW described in [47] for the stochastic volatility surface or *locally parameterized* alternative that can be described following wavelet model for the variance surface

$$\begin{aligned}
\nu(t, T, K) \triangleq& \sum_{i,j=1}^{d} \beta_t^{i,j} \psi_{i,j}(T, K); \beta_t^{i,j} > 0 \\
\psi_{i,j}(t) =& \frac{1}{\sqrt{2^{-i}}} \psi\left(\frac{t - \frac{1}{j}}{2^{-i}}\right) \\
\psi(x, y) =& \frac{1}{\pi\sigma^2}\left(1 - \frac{1}{2}\left(\frac{x^2 + y^2}{\sigma^2}\right)\right) e^{-\frac{x^2+y^2}{2\sigma^2}}.
\end{aligned} \tag{4.11}$$

In Appendix 4.8.7, a natural Riemannian metric g^c is described on the space $(0, \infty)$, which ensures that the factor model $\sum_{i,j=1}^{d} \beta_t^{i,j} \psi_{i,j}(T, K)$ is well defined for all time by forcing the dynamic factors $\beta_t^{i,j}$ not to be able to escape $(0, \infty)$ in a finite amount of time. The price of a European call option $C(t, T, K)$ on the stock price can be represented by the following flow model:

(i) The manifold $(\mathscr{M}, g^c)$ is the subset $(0, \infty)^{d^2}$ of $\mathbb{R}^{d^2}$ equipped with the Riemannian metric g^c described in Lemma 4.8.5,

(ii) The set of possible parameters $\mathcal{U} = [0, \infty) \times [0, \infty)$ represents the set of all possible strikes and times of maturity of a European option on S_t,

(iii) The stochastic factors β_t are the loadings on the wavelet basis functions $\psi_{i,j}$,

[1] The variance surface tends to be numerically simpler to work with in general.

(iv) The factor model ϕ is the map

$$\phi(t,\beta,(T,K)) \triangleq \sum_{i,j=1}^{d} \beta_t^{i,j}\psi_{i,j}(T,K),$$

(v) The functional $F_t(\cdot|(T,K))$ encoding the stochastic variance surface ν into the price of a European call option is the solution operator $\Sigma_t(\cdot|T,K)$ mapping a positive number ς to the solution of the initial value problem

$$\begin{aligned} \partial_t c(\tau,K) =& \varsigma c(\tau,K) \\ c(0,K) =& (S_t-K)^+. \end{aligned} \tag{4.12}$$

Therefore the price of a European call option on a stock with price S_t can be represented by the flow model $\left(\Sigma_t(\cdot|T,K),\sum_{i,j=1}^{d}\beta^{i,j}\psi_{i,j}(T,K);\beta_t^{i,j}>0,\beta_t^c\right)$, with realization

$$C(t,T,K)=\Sigma_t\left(\nu(t,T,K)|(T,K)\right)=\Sigma_t\left(\phi(t,\beta,(T,K))|(T,K)\right)=F(\phi(t,\beta_t,u)|u).$$

Example 4.3.8 (Price Of An Option with Uncertain Stochastic Volatility)**.** Suppose μ_t is the density of the log price of a risky asset at a maturity time $t>0$. The value of an option at time t with payoff-function f, denoted by p_t^f, is computed in terms of the log-returns as

$$p_t^f \triangleq \int_{\mathbb{R}} f(x)\mu_t(x)dx. \tag{4.13}$$

Suppose μ_t itself is unknown, but has been estimated by the density w_t and let l_t be the likelihood ratio

$$l_t \triangleq \frac{g_t}{w_t}.$$

Under suitable conditions (see [2] for details), the time t option price p_t^f can be approximated by

$$\begin{aligned} p_t^{f,N} \triangleq& \sum_{n=0}^{N} f_n l_t^n \\ f_t^n \triangleq& \int_{\mathbb{R}} f(x)H_n(x)w_t(x)dx, \\ l_t^n \triangleq& \int_{\mathbb{R}} H_n(x)\mu_t(x)dx, \end{aligned} \tag{4.14}$$

where $\{H_n\}_{n\in\mathbb{N}}$ is an orthonormal polynomial basis of the weighted space $L^2_{w_t}$ and f_t^n as well as l_t^n are the projection of f and l_t onto the span of each function H_n.

In general, making assumptions on the dynamics of the volatility is more subtle than modeling those of the stock price since, since the volatility is unobservable. Therefore the exact specification of the volatility's evolution is subject to a certain amount of model uncertainty. It is therefore more realistic to consider a variety of dynamics for the asset's volatility, each of which describes a different model for the log-stock-price,

$$\begin{aligned}&\left.\begin{cases} dS_t^k &= \sigma_t^k dW_t \\ d\sigma_t^k &= \alpha^k(t, \sigma_t^k)dt + \beta^k(t, \sigma_t^k)dB_t \end{cases}\right\}_{k=1}^{d} \\ &\mathbb{E}\left[[W_\star, B_\star]_t\right] = \rho, \end{aligned} \tag{4.15}$$

where the correlation coefficient, denoted by ρ, is between -1 and 1, and S_t^k denotes the log-stock price under the assumption that the volatility process follows σ_t^k.

If $\mu_t^k(x)$ denotes the density log-price of the asset under the volatility specification σ_t^k, then it is more robust to select an optimally empirically performant mixture

$$\sum_{\substack{k=1 \\ \beta^1+\cdots+\beta^d=1 \\ \beta^k>0}}^{d} \beta^k \mu_t^k(x)$$

than modeling the log-stock price using a single density specification.

The constraint set

$$\left\{\beta \in \mathbb{R}^d : \beta^1 + \cdots + \beta^d = 1 \text{ and } \beta_k > 0, \right\} \tag{4.16}$$

may be difficult to work with. Instead working with the geometry provided by the largest open ball lying within this constraint set, defined by

$$B \triangleq \left\{ x \in \mathbb{R}^{d+1} : \sqrt{\sum_{i=0}^{d} x_i^2} = 1 \text{ and } x_0, \ldots, x_d > 0 \text{ and } d_\circ(\bar{x}, x) < \frac{\pi}{4} \right\},$$

provides a convenient proxy to this constraint set; here $d_\circ$ is the distance intrinsic to the d-sphere. This is because a geometry may be defined on B which provides a closed-form characterization of all continuous semi-martingales which do not leave B in a finite amount of time. See Appendix 4.8.8 for a detailed treatment of this technical point.

To summarize, the approximation $p_t^{f,N}$ to the value of the option at time t may be represented by the following flow model:

(i) The manifold $(\mathscr{M}, g)$ is the largest open ball B lying in the intersection of the d-dimensional sphere and the first orthant of $\mathbb{R}^d$, and g a Riemannian metric defined in Appendix 4.8.8.

(ii) The set of possible parameters $\mathcal{U} = \{0\}$ is a single dummy point,

(iii) The stochastically evolving mixture proportions of the hypothesized densities for the asset log-price movements, are described by a B-valued continuous semi-martingale β_t^c. The precise definition of β_t^c is given in Appendix 4.8.8.

(iv) The factor model ϕ is the map mixing the densities $g_1, \dots, g_d$:

$$\phi(t, \beta, u) \triangleq \sum_{k=1}^{d} \beta^k \mu_t^k(u),$$

(v) The functional F encoding a density $\mu_t(x)$ into approximate option value is

$$F(\cdot|u) \triangleq \sum_{n=0}^{N} f_t^n \left(\int_{\mathbb{R}} H_n(x) \cdot dx \right).$$

Hence the approximate price of an option with payoff f under uncertain stochastic volatility models can be represented by the flow model

$$\left(\sum_{n=0}^{N} f_t^n \left(\int_{\mathbb{R}} H_n(x) \cdot dx \right), \sum_{k=1}^{d} \left(\beta_t^c\right)^k \mu_t^k(u), \beta_t^c \right),$$

with realization

$$p_t^{f,N} = \sum_{n=0}^{N} f_t^n l_t^n = \sum_{n=0,k=1}^{N,d} \left(\beta_t^c\right)^k f_t^n \left(\int_{\mathbb{R}} H_n(x) \mu_t^k(x) dx \right) = F(\phi(t, \beta_t^c, u)|u).$$

Analogously to the FDR-HJM framework of [58] and its consistent analogue studied in [39], analogues to the HJM drift restriction and consistency conditions are given in the next section to characterize the non-existence of arbitrage opportunities. These will play an integral role in constructing an arbitrage-free penalty.

Remark 4.3.9. The terminology *"flow model"* comes from the fact that the factors β_t evolve on a manifold with time-dependent Riemannian metric, which in differential geometry is called a *flow*. Analogously to the static models of [9], if the mappings

$$\{\beta \mapsto (u \mapsto \phi(t, \beta, u))\}_{t \in [0,\infty)},$$

are invertible, they dynamically associates the manifold $(\mathscr{M}, g_t)$ to open subset of the Sobolev space $W^s(\mathcal{U})$ defined by their image at time t. The encoding functional terminology originates from the GHJM setting, where an unobservable process, or codebook is encoded into the price of an asset as introduced in [32].

Remark 4.3.10. In Lemma 4.8.9 it is shown that if $\mathscr{M} \subseteq \mathbb{R}^D$ can be interpreted as a suitable set of constraints on the factors β, then any continuous semi-martingale β_t on $\mathbb{R}^d$ can be transformed to a continuous semi-martingale β_t^c on $\mathscr{M}$ such that its infinitesimal tangential movements are β_t and β_t^c do not leave $\mathscr{M}$ in a finite amount of time. If $\mathbb{R}^d = \mathscr{M} = \mathbb{R}^D$, then β_t and β_t^c are indistinguishable. A time-dependent Riemannian metric g_t is allowed so that the structure of the factor's domain $\mathscr{M}$ may be updated.

By the tower law of conditional expectation, all the flow models of Example 4.3.6 are martingales and therefore cannot admit arbitrage opportunities. However, not all flow models are arbitrage-free flows and, in order to build an arbitrage penalty, we will first characterize the existence arbitrage for a flow model. First, recall that a local martingale measure (LMM) is a measure dominated by the reference measure $\mathbb{P}$, under which each $X_t(u)$ is a local martingale. Unlike an equivalent local martingale measure (ELMM), and LMM need not be equivalent to $\mathbb{P}$ (see [88] for details on LMMs and ELMMs).

Theorem 4.3.11. Let $\mathbb{Q} \ll \mathbb{P}$, ϕ be deterministic, and (F, ϕ, β_t, g_t) be a flow model with realization $X_t(u)$. For every $u \in \mathcal{U}$, the measure $\mathbb{Q}$ is a LMM for $X_t(u)$ if and only if

$$\begin{aligned}
&\int_0^t \mathscr{D}_s F(s, \varphi_s^u, [\varphi^u]_s | u) ds + \int_0^t \mathscr{V}_s F(s, \varphi_s^u, [\varphi^u]_s | u) \left[\frac{\partial \phi}{\partial s}(s, \beta_s, u) + \sum_{i=0}^d (\xi_s e_i) \phi(s, \beta_s, u) \mu(s, \beta_s) \right. \\
&\qquad \left. + \frac{1}{2} \sum_{i,j=1}^d Hess^{g_t} \left(\phi(s, \beta_s, u) \right) (\beta_s e_i, \beta_s e_j) [\beta^i, \beta^j]_s \right] ds \\
&\qquad + \int_0^s \left(\frac{1}{2} tr[{}^t \mathscr{V}_s^2 F(s, \varphi_s^u, [\varphi^u]_s)] \left[\sum_{i=0}^d (\xi_s e_i) \phi(s, \beta_s, u) \sigma(s, \beta_s) \right]^2 ds \right) = 0 \\
&\varphi_t^u \triangleq \phi(t, \beta_t, u)
\end{aligned} \tag{4.17}$$

is satisfied for $\mathbb{Q} \otimes m$-a.e. (ω, t) in $\Omega \times [0, \infty)$, here $\mathscr{D}$ and $\mathscr{V}$ are the horizontal and vertical derivatives of [33] (see [45] for details). In particular, if $\mathbb{Q} \sim \mathbb{P}$ then $\mathbb{Q}$ is simultaneously an ELMM for μ-a.e. $\{X_t(u)\}_{u \in \mathcal{U}}$ if and only if equation (4.17) holds for $\mathbb{Q} \otimes m \otimes \mu$-a.e. $(\omega, t, u) \in \Omega \times [0, \infty) \times \mathcal{U}$. Here $Hess^{g_t}$ is the Hessian on $(\mathscr{M}, g_t)$ at time t.

Proof. See Appendix B. □

Proposition 4.3.12 (Arbitrage-Free Characterization for the Forward-Rate Curve)**.** Let $\mathbb{Q} \ll \mathbb{P}$, ϕ be deterministic, and (F, ϕ, β_t, g_t) be the flow model of Example 4.3.4. The measure $\mathbb{Q}$ is a LMM for each $\{B(t, T)\}_{T \in [0, \infty)}$ for m-a.e. Maturity $T \geq 0$ if and only if, in local-coordinates,

$$\begin{aligned}
&\int_0^t \left[\frac{\partial \phi}{\partial t}(s, T, \beta_s) + \sum_{i=0}^d (\xi_s e_i) \phi(s, T, \beta_s) \mu(s, \beta_s) \right. \\
&\quad + \frac{1}{2} \sum_{i,j=1}^d \left(\frac{\partial^2 \phi}{\partial \beta_i \beta_j}(s, T, \beta_s) - \sum_{k=1}^d \Gamma_{ij}^k(t) \frac{\partial \phi}{\partial \beta_k}(s, T, \beta_s) \right) \sigma_i(s, \beta_s) \sigma_j(s, \beta_s) \\
&\quad \left. + \frac{1}{2} \sum_{i,j=1}^d \frac{\partial \phi}{\partial x_i}(s, T, \beta_s) \frac{\partial \phi}{\partial x_j}(s, T, \beta_s) \sigma_i(s, \beta_s) \sigma_j(s, \beta_s) \right] ds = 0,
\end{aligned} \tag{4.18}$$

holds, $\mathbb{Q} \otimes m \otimes m$-a.e. $\Gamma_{ij}^k(t)$ are the Christoffel symbols of the Riemannian metric g_t at time t.

If $(\mathscr{M}, g_t)$ is an Euclidean space and we assume that $\mu(t,z) = \mu(z)$ and $\sigma(t,z) = \sigma(z)$ are both smooth and deterministic, then Corollary 4.3.12 can be simplified to a PDE and we recover the consistency result of [40]. To see this, first consider the time reversal $\phi(t, \tau(t), T)$ where $\tau \mapsto T - t$. Then since the integral equation (4.18) must hold for all $t \leq T$ and all initial conditions of β_t it follows that we may let $t \to 0$. Doing so and replacing t by x, we obtain the PDE found in [40, Proposition 9.1].

Proposition 4.3.13 (Arbitrage-Free Characterization of the Stochastic Local Volatility Surface)**.** Let (τ, x) be a pair of time-to maturity and log-strike, let $\mathbb{Q} \sim \mathbb{P}$ be an ELMM for S_t, and consider the stochastic local volatility surface setting of Example 4.3.7. The measure $\mathbb{Q}$ is an ELMM for the call surface $C(t,\tau,x)$ for every pair $(\tau, x) \in [0,\infty) \times [0,\infty)$ if and only if, in local-coordinates on $(\mathscr{M}, g_t)$,

$$\begin{aligned}
&\int_0^t \mathbf{\Delta}_s(\tau,x) \left[\frac{\partial \varphi}{\partial x}(s,\tau,x,\beta_s) + \sum_{i=0}^d (\xi_s e_i) \varphi(s,\tau,x,\beta_s)\mu(s,\beta_s) \right. \\
&+ \frac{1}{2}\sum_{i,j=1}^d \left(\frac{\partial^2 \phi}{\partial \beta_i \beta_j}(s,\tau,x,\beta_s) - \sum_{k=1}^d \Gamma_{ij}^k(t) \frac{\partial \phi}{\partial \beta_k}(s,\tau,x,\beta_s) \right) \sigma_i(s,\beta_s)\sigma_j(s,\beta_s)\, ds \\
&\left. + \frac{1}{2}\mathbf{\Gamma}_s(\tau,x) \left[\sum_{i=0}^d (\xi_s e_i)\varphi(s,\tau,x,\beta_s)\sigma_i(s,\beta_s) \right]^2 \right] ds = 0
\end{aligned} \tag{4.19}$$

holds $\mathbb{P} \otimes m \otimes m$-a.e, where $\mathbf{\Delta}_s(\tau,x)$ and $\mathbf{\Gamma}_s(\tau,x)$ are the Greeks of the modified stock price $\tilde{S}_t$ defined by the SDE

$$d\tilde{S}_t = \frac{K\sigma(t,\tau,x)}{\sqrt{2}} dW_t.$$

The Greeks can be computed by

$$\begin{aligned}
\mathbf{\Delta}_t(\tau,x) &\triangleq \frac{1}{\tau z(\tilde{S}_t)} \mathbb{E}\left[\left(\tilde{S}_{\tau+t} - e^x \right)_+ \eta_t^\tau \mid \mathscr{F}_t^{\tilde{S}} \right] \\
\mathbf{\Gamma}_t(\tau,x) &\triangleq \mathbb{E}\left[(\tilde{S}_{\tau+t} - e^x)_+ \zeta_{s,\tau}^x \mid \mathscr{F}_t{}^S \right],
\end{aligned} \tag{4.20}$$

where the weight $\zeta_{s,T}^x$ is defined by

$$\begin{aligned}
\eta_t^\tau &\triangleq \int_t^{\tau+t} \frac{z_s}{\sigma(S_s)} dW_s \\
\zeta_t &\triangleq \frac{(\eta_0^{\tau+t} - \eta_0^s)^2}{\tau^2 z_t^2} - \frac{\mathscr{V}\sigma(S_t)}{\sigma(S_t)} \frac{(\eta_0^{\tau+t} - \eta_0^t)}{\tau z_t} - \frac{1}{\tau \sigma^2(S_t)},
\end{aligned}$$

and z_t is the first variation process of β_t defined by

$$dz_t = \mu(t, z_t)dt + \sigma'(S_t) z_t dW_t.$$

A consequence of equation (4.19) is that a necessary condition for a stochastic local volatility surface to be arbitrage-free is for the option's Gamma and Delta to evolve according to the ratio $\rho_t(\tau, x)$ defined by

$$\begin{aligned} \mathbf{\Delta}_t(\tau, x) &= -\rho_t(\tau, x)\mathbf{\Gamma}_t(\tau, x) \\ \rho_t(\tau, x) &\triangleq \frac{\left[\sum_{i=0}^{d} \varphi_t^i \sigma_s^i\right]^2}{2\dot{\varphi}_t + \sum_{i=0}^{d} \varphi_t^i \mu_t + \sum_{i,j=1}^{d} \left(\varphi_t^{i,j} - \sum_{k=1}^{d} \Gamma_{ij}^k(t)\varphi_t^k\right) \sigma_t^i \sigma_t^j}, \end{aligned} \tag{4.21}$$

whenever the denominator of $\rho_t(\tau, x)$ is defined. Here μ_t^i, σ_t^i, φ_t^i, $\varphi_t^{i,j}$, and $\dot{\varphi}_t^i$ abbreviate $\mu_i(t, \beta_t)$, $\sigma_i(t, \beta_t)$, $\frac{\partial \varphi(t,\tau,x,\beta_t)}{\partial \beta_i}$, $\frac{\partial^2 \varphi(t,\tau,x,\beta_t)}{\partial \beta_i \beta_j}$, and $\frac{\partial \varphi(t,\tau,x,\beta_t)}{\partial t}$ respectively. Note that Proposition 4.3.13 has different assumptions than the central result of [18]. Namely, the differentiability requirements for α and β are weakened and the prescription on the dynamics on ν is relaxed.

Example 4.3.14 (Uncertain Volatility Model)**.** Let (F, ϕ, β_t, g_t) be the flow model of Example 4.3.8, with $d = 2$. The maps ϕ and F are infinitely differentiable and constant in time. Therefore, their derivatives may be readily computed to be

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= 0 \\ \nabla \phi &= \left(g_t^1, g_t^2\right) \\ \nabla^2 \phi &= 0 \\ \mathscr{V} \sum_{n=0}^{N} f_n \left(\int_{\mathbb{R}} H_n(x) \cdot dx\right) &= \sum_{n=0}^{N} f_n H_n(x) \cdot dx \\ \mathscr{V}^2 \sum_{n=0}^{N} f_n \left(\int_{\mathbb{R}} H_n(x) \cdot dx\right) &= \sum_{n=0}^{N} f_n H_n(x) dx \\ \mathscr{D}^2 \sum_{n=0}^{N} f_n \left(\int_{\mathbb{R}} H_n(x) \cdot dx\right) &= 0. \end{aligned} \tag{4.22}$$

Substituting the quantities of equation (4.22) into equation (4.17) and noting that it is sufficient for the integrand to be zero for all values of β_t^c and of t, we conclude that an uncertain stochastic volatility model is arbitrage-free if

$$g^T \left(M(t, z) + \frac{1}{2}\Gamma(t, z) + \frac{\Sigma(t, z)}{2\beta(z)^T g(x)}\right) = 0 \tag{4.23}$$

holds for m-a.e. $(t, x, z) \in [0, \infty) \times \mathbb{R} \times \{z \in \mathbb{C} : Im(z) > 0\}$ where g, M, Σ, and β are:

$$g(x) \triangleq (g^1(x), g^2(x)); M(t, \beta) \triangleq (\mu^1(t, \beta^1), \mu^2(t, \beta^2)); \Sigma \triangleq (\sigma^1(t, \beta^1), \sigma^2(t, \beta^2)); \beta \triangleq (\beta^1, \beta^2)$$

$$\beta^k \triangleq \cos(\|\frac{z \overset{\rightarrow}{-} i}{z + i}\|)(\frac{1}{\sqrt{2}} \overset{\rightarrow}{,} \frac{1}{\sqrt{2}}) + \sin(\|\frac{z \overset{\rightarrow}{-} i}{z + i}\|)\frac{\frac{z \overset{\rightarrow}{-} i}{z+i}}{\|\frac{z \overset{\rightarrow}{-} i}{z+i}\|}; k = 1, 2. \tag{4.24}$$

Here Γ_{ij}^{k} are the Christoffel symbols of the hyperbolic upper-half plane and the expression for β is discussed in Appendix 4.8.10. Rearranging and integrating both sides over $\mathbb{R}$, we obtain an HJM-type drift restriction

$$\int_{x\in\mathbb{R}} g^T(x)M(t,z)dx = -\int_{x\in\mathbb{R}} g^T(x)\left(\frac{\Gamma(t,z)\beta(z)^T g(x)+\Sigma(t,z)}{2\beta(z)^T g(x)}\right)dx. \tag{4.25}$$

Theorem 4.3.11 characterizes arbitrage-free flows. Flow models which are not arbitrage-free may be minimally deforming them into arbitrage-free flows using Theorem 4.3.11 as well as a suitable measure of deviation from the initial flow model. This deviation is introduced now.

4.4 Arbitrage-Free Regularization

Following the introduction of flow models, the formalization of arbitrage-free regularization requires three components: a precise definition of what it means to deform a flow model, a rigorous way to measure how far a deformation is from the undeformed reference model, and a penalty detecting the existence of arbitrage opportunities permitted by the deformed model. This section introduces these three components in order and uses them to develop arbitrage-free regularization.

Definition 4.4.1 (Model Deviation) *Let (F,ψ,β_t,g_t), (F,ϕ,β_t,g_t) be in $C_\mu\left(F,\beta_t\right)$. Let $\mathbf{\Delta}_t^D$ be a non-anticipative functional from $C_\mu\left(F,\beta_t\right)\times C_\mu\left(F,\beta_t\right)$ to $\mathbb{R}$ satisfying*

(i) ***Non-negativity*** *$\mathbf{\Delta}_t^D$ is non-negative in both arguments.*

(ii) ***Identity of Indiscernibles*** *$\mathbf{\Delta}_t^D((F,\psi,\beta_t,g_t),(F,\phi,\beta_t,g_t)) = 0$ if $\psi(t,\beta_t,u) = \phi(t,\beta_t,u)$ for $\mathbb{P}\otimes m\otimes\mu$-a.e. (ω,t,u).*

(iii) ***Convexity*** *$\mathbf{\Delta}_t^D$ is convex in its first argument,*

(iv) ***Data-Driven*** *$\mathbf{\Delta}_t^D$ is $\mathscr{F}_t^X$-predictable,*

where $\mathscr{F}_t^X$ is the σ-algebra generated by $X_t(u)$. The non-anticipative functional $D_t \triangleq \mathbf{\Delta}_t^D(\cdot,\phi)$ is called a model deviation.

Lemma 4.4.2 (Effective Existence)**.** Let $g:\mathbb{R}\times\mathbb{R}\to[0,\infty]$ be a Borel-measurable function which is convex in its first argument and let μ_t be a $\mathscr{F}_t^X$-predictable cádlág process taking values in the set of σ-finite Borel measures on $\mathcal{U}$ equivalent to μ. The family of functionals

$$D_t(F,\psi,\beta_t,g_t) \triangleq \mathbb{E}\left[\int_0^t\int_{u\in\mathcal{U}} g\left(\psi(s,\beta_s,u),\phi(s,\beta_s,u)\right)\frac{d\mu_s}{d\mu}\mu(du)ds\right]$$

defines a model deviation.

Proof. By the monotonicity of integration (i) and (ii) hold. For every $(\omega, t, u) \in \Omega \times [0, \infty) \times \mathcal{U}$, the function $g(\cdot, \phi(t, \beta_t, u)(\omega))$ is strictly convex. Therefore, its integral is also convex by a result of [84]. □

Definition 4.4.3 (Deformation of a Flow Model) *Let (F, ψ, β_t, g_t), (F, ϕ, β_t, g_t) be in $C_\mu(F, \beta_t)$. The flow model (F, ψ, β_t, g_t) is said to be a deformation of (F, ϕ, β_t, g_t) if and only if there exists an $L^2_\mu(H^s(\mathcal{U})_\mu; H^s(\mathcal{U})_\mu)$-valued stochastic process Φ_t such that*

1. ***(Deformation)*** *$D_t((F, \psi, \beta_t, g_t), (F, \Phi_t(\phi), \beta)) = 0$, for $\mathbb{P} \otimes m \otimes \mu$-a.e. (ω, t, u) in $\Omega \times [0, \infty) \times \mathcal{U}$,*
2. ***(Predictability)*** *The process Φ_t is $\mathscr{B}_t \otimes \mathscr{F}_t$-predictable,*
3. ***(Square Integrability)*** $\mathbb{E}\left[\int_{u \in \mathcal{U}} \int_0^\infty (\Phi_t(\phi)(t, \beta_t, u))^2 \, dt\mu(du)\right] < \infty$

We say that Φ_t deforms (F, ϕ, β_t, g_t) and we will interchangeably denote the deformation (F, ψ, β_t, g_t) with Φ_t, and vice-versa depending on the context. The collection of all deformations of (F, ϕ, β_t, g_t) will be denoted by $D^2_\mu(F, \phi, \beta_t, g_t)$.

We will denote the family of price processes corresponding to the flow model (F, ψ, β_t, g_t) by X_t^Φ. For every u in $\mathcal{U}$, the member of X_t^Φ indexed by u will be denoted by $X_t^\Phi(u)$.

Definition 4.4.4 (Arbitrage Penalty)
Let $\left(AF_t : D^2_\mu(F, \phi, \beta_t, g_t) \to [0, \infty]\right)_{t \in [0, \infty)}$ be a non-anticipative functional such that for every $t \in [0, \infty]$ and every $(F, \phi, \beta_t, g_t) \in C_\mu(F, \beta_t)$,

$$AF_t(F, \phi, \beta_t, g_t)(t, \beta_t, u) = 0; \quad \mathbb{P} \otimes m \otimes \mu - a.e.$$

if and only if the subset of parameters in $\mathcal{U}$ for which $\mathbb{P}$ is not an ELMM for $X_t(u)$ has μ-measure 0. AF_t is called an arbitrage penalty for (F, ϕ, β_t, g_t).

Equation (4.17) characterizes the measures $\mathbb{Q} \ll \mathbb{P}$ under which $X_t(u)$ is a local-martingale as the measures under which LHS of equation (4.17) is equal to 0. Denote the LHS of equation (4.17) by $\Lambda(F, \phi, \beta_t, g_t)$. An arbitrage-penalty can be built by integrating the square of $\Lambda(F, \phi, \beta_t, g_t)$ over all the relevant states.

Theorem 4.4.5 (Effective Existence)**.** For any (F, ψ, β_t, g_t) in $D^2_\mu(F, \phi, \beta_t, g_t)$, the non-anticipative functional AF_t defined by

$$AF_t(F, \psi, \beta_t, g_t) \triangleq \mathbb{E}\left[\int_{u \in \mathcal{U}} \int_0^t \left[\Lambda(F, \phi, \beta_s, g_s)(\omega, s, u)\right]^2 ds\mu(du)\right] \tag{4.26}$$

is an arbitrage penalty on any subset $\mathscr{A}$ of $D^2_\mu(F, \phi, \beta_t, g_t)$.

Proof. Since the map $(F, \psi, \beta_t, g_t) \mapsto \Lambda(F, \psi, \beta_t, g_t)$ is non-anticipative, then AF_t is a non-anticipative functional. Let $\mathscr{A} \subseteq D^2_\mu(F, \phi, \beta_t, g_t)$. For any $(F, \psi, \beta_t, g_t) \in \mathscr{A}$ the process $\Lambda(F, \psi, \beta_t, g_t)(\omega, s, u)$ equals 0 $\mathbb{P} \otimes m \otimes \mu$-a.e. if and only if the SPDE of Theorem 4.3.11 holds for $\mathbb{P} \otimes m \otimes \mu$-a.e. $(\omega, t, u) \in \Omega \times [0, \infty) \times \mathcal{U}$, where ϕ is replaced with ψ. For μ-a.e. $u \in \mathcal{U}$ the process $\Lambda(F, \psi, \beta_t, g_t)$ is real-valued and therefore the process $[\Lambda(F, \psi, \beta_t, g_t)]^2$ takes values in $[0, \infty]$ and the monotonicity of integration implies that Definition 4.4.4 part (ii) holds. □

Including all the possible deformations of a flow model may be computationally intractable, thus limiting the problem defined in equation (4.1) to a more narrow subset of $D^2_\mu(F, \phi, \beta_t, g_t)$ is advantageous. Within such a subset of $D^2_\mu(F, \phi, \beta_t, g_t)$, it may happen that there is more than one optimizer of equation (4.1). All these optimizers exhibit equal model deviation from the empirical factor model ϕ and none of them admits arbitrage opportunities. Classes of deformations are introduced to address these two issues.

Lemma 4.4.6 (Equivalent)**.** The relation $\Phi_t \sim \Psi_t$ on elements of $D^2_\mu(F, \phi, \beta_t, g_t)$ defined by

$$D_t(\Phi_t) = D_t(\Psi_t) \text{ and } AF_t(\Phi_t) = AF_t(\Psi_t),$$

describes an equivalence relation on $D^2_\mu(F, \phi, \beta_t, g_t)$.

Proof. The equivalence property follows directly from the fact that equality is an equivalence, together with the properties of logical conjunctions. □

Let τ denote the topology on $C_\mu(F, \beta_t)$ generated by the functionals $\left\{\mathbf{\Delta}^D_t\right\}$ together with the sets $\{AF_t^{-1}(-\infty, \epsilon] : \epsilon \in \mathbb{R}\}$. The subspace topology of τ relative to $D^2_\mu(F, \phi, \beta_t, g_t)$ makes $D^2_\mu(F, \phi, \beta_t, g_t)$ into a topological subspace of $C_\mu(F, \beta_t)$. Since $\overset{D}{\underset{AF}{\sim}}$ defines an equivalence relation on $D^2_\mu(F, \phi, \beta_t, g_t)$, its set of equivalence classes inherits the quotient topology of $D^2_\mu(F, \phi, \beta_t, g_t)$ relative to the equivalence relation $\overset{D}{\underset{AF}{\sim}}$. This topological space will be denoted by $\mathbb{D}^2_\mu(F, \phi, \beta_t, g_t)$.

Remark 4.4.7. By construction of τ, AF_t is lower semi-continuous on $D^2_\mu(F, \phi, \beta_t, g_t)$.

Definition 4.4.8 (Class of Deformations) *Let $\mathbb{D}^2_\mu(F, \phi, \beta_t, g_t)$ denote, $D^2_\mu(F, \phi, \beta_t, g_t)/\sim$ where $\sim$ is the equivalence relation defined in Lemma 4.4.6.A subset $\mathscr{A}$ of $\mathbb{D}^2_\mu(F, \phi, \beta_t, g_t)$ is said to be a class of deformations for (F, ϕ, β_t, g_t) if and only if*

1. *The equivalence class of the trivial deformation $\tilde{I}_t$ defined by*

$$\tilde{I}_t(\psi) \mapsto \psi, \tag{4.27}$$

is an element of $\mathscr{A}$; $\psi \in W^{s,2}_{\mu_t}(\mathcal{U})$

2. *The subset $\{\Phi_t \in \mathscr{A} : AF_t(\Phi_t) = 0\}$is non-empty,*

3. *For every $\Phi^1, \Phi^2 \in \mathscr{A}$ there exists some $\Phi^{1,2} \in \mathscr{A}$ such that*

$$\Phi_t^2(\Phi_t^1(\phi)) = \Phi_t^{1,2}(\phi).$$

Remark 4.4.9. For simplicity of notation, unless unclear within its context, the equivalence relation $\overset{D}{\underset{AF}{\sim}}$ will be denoted simply by $\sim$. Likewise the trivial deformation $\tilde{I}_t$ may be denoted by the empirical factor model ϕ. Furthermore the equivalence relation $\sim$ ensures that AF_t and D_t are well-defined on $\mathbb{D}_\mu^2(F, \phi, \beta_t, g_t)$.

Definition 4.4.10 (Arbitrage-Free Regularization Operator) *Let D_t be a fixed model divergence and AF_t be a fixed an arbitrage penalty on $\mathbb{D}_\mu^2(F, \phi, \beta_t, g_t)$. Let $Dom(F, \phi, \beta_t, g_t)$ denote the collection of pairs $(\Psi_t, \mathscr{A})$, of a deformation Ψ_t in $\mathbb{D}_\mu^2(F, \phi, \beta_t, g_t)$ and a subset $\mathscr{A}$ of $\mathbb{D}_\mu^2(F, \phi, \beta_t, g_t)$ satisfying*

1. $\lim_{\lambda \mapsto 0^+} \underset{\Phi_t \in \mathscr{A}}{\operatorname{arginf}} D_t(\Phi_t(\Psi_t(\phi))) + \frac{1}{\lambda} AF_t(\Phi_t(\Psi_t(\phi))) \in \mathbb{D}_\mu^2(F, \phi, \beta_t, g_t),$

2. $\lim_{\lambda \mapsto 0^+} \inf_{\Phi_t \in \mathscr{A}} D_t(\Phi_t(\psi)) + \frac{1}{\lambda} AF_t(\Phi_t(\psi)) < \infty.$

The map $\mathbb{A}_\phi [\cdot|\cdot] : Dom(F, \phi, \beta_t, g_t) \to \mathbb{D}_\mu^2(F, \phi, \beta_t, g_t)$ is called the arbitrage-free regularization operator with domain $Dom(F, \phi, \beta_t, g_t)$. We will call $\mathbb{A}_\phi [\Phi_t|\mathscr{A}]$ the arbitrage-free regularization of $(F, \Phi_t(\phi), \beta_t)$ with respect to the class of deformations $\mathscr{A}$ and we will denote by

$$X_t^{\mathscr{A}}(u) \triangleq F\left(\mathbb{A}_\phi [\phi_t|\mathscr{A}] (t, \beta_t, u), [\mathbb{A}_\phi [\phi|\mathscr{A}] (\star, \beta_\star, u)]_t\right),$$

the price of asset under the arbitrage-free regularized model.

Theorem 4.4.11 (Arbitrage-Free Regularization)**.** Let $\mathbb{A}_\phi [\cdot|\cdot]$ be an arbitrage-free regularization operator. For any $(\Psi_t, \mathscr{A})$ in $Dom(F, \phi, \beta_t, g_t)$, $X_t^{\mathscr{A}}(u)$ is a $\mathbb{P}$-local martingale.

Proof. Definition 4.4.8 (iii) implies that the minimizers of $\mathbb{A}_\phi [\Psi|\mathscr{A}]$ and of $\mathbb{A}_\phi [\phi|\mathscr{A}]$ are the same. Therefore, without loss of generality Φ_t^λ, will denote a minimizer of $D_t(\Phi_t(\phi)) + \frac{1}{\lambda} AF_t(\Phi_t(\phi))$ over $\mathscr{A}$.

Suppose that $\lim_{\lambda \to 0^+} AF_t\left(\Phi_t^\lambda\right) \neq 0$. Since the non-anticipative functional AF_t is non-negative, there must exist a real number $c > 0$ and a sequence $\{\Phi_t^{\lambda_n}\}_{n \in \mathbb{N}}$ with $\lambda_n \mapsto 0^+$, such that $\lim_{n \mapsto \infty} AF_t\left(\Phi_t^{\lambda_n}\right) = c$.

The functional D_t is non-negative, therefore

$$\begin{aligned}
\infty &= \lim_{n \mapsto \infty} \frac{1}{\lambda_n} c && (4.28)\\
&= \lim_{n \mapsto \infty} \frac{1}{\lambda_n} AF_t(\Phi_t^\lambda) && (4.29)\\
&\leq \lim_{n \mapsto \infty} D_t(\Phi_t^\lambda) + \frac{1}{\lambda_n} AF_t(\Phi_t^\lambda), && (4.30)
\end{aligned}$$

contradicting Definition 4.4.10 (ii). Therefore $\lim_{\lambda \to 0^+} AF_t\left(\Phi_t^\lambda\right) = 0$. Since AF_t is non-negative and lower semi-continuous in the topology of $\mathbb{D}^2_\mu(F, \phi, \beta_t, g_t)$, it follows that

$$0 \leq AF_t\left(\lim_{\lambda \to 0^+} \Phi_t^\lambda\right) = \lim_{\lambda \to 0^+} AF_t\left(\Phi_t^\lambda\right) = 0.$$

Since $\mathbb{A}_\phi\left[\Psi | \mathcal{A}\right]$ is defined to be $\lim_{\lambda \to 0^+} \Phi_t^\lambda$, Definition 4.4.4 implies that $X_t^{\mathscr{A}}(u)$ is a $\mathbb{P}$-local martingale. □

This section introduced and justified the theoretical machinery needed to formalize the arbitrage-free regularization problem. The next section discusses examples, applications, and connections to other theories and practices in finance.

4.5 Extensions of Classical Risk-Neutral Pricing Theory

For a particular class of deformations, the Fundamental Theorem of Asset Pricing (FTAP) of [26], can be expressed in terms of the existence and uniqueness of the arbitrage-free regularization operator's output. In this section, we show how the FTAP, and the minimal martingale measure of [88] are both particular formulations of the arbitrage-free regularization problem. This fact used to motivate arbitrage-free regularization over more general classes of deformations than those corresponding to measure changes. Consequentially, arbitrage-free regularization extends the reach of classical risk-neutral pricing theory and techniques to flow models which permit arbitrage-opportunities, since they are minimally deformable into arbitrage-free flows using arbitrage-free regularization.

In [88, Corollary 3 and Theorem 1] it was shown that for a sufficiently well-behaved wealth process, there exists an equivalent local martingale measure (ELMM) to the real-world measure $\mathbb{P}$ which is most similar to $\mathbb{P}$. Dissimilarity between measures is quantified in terms of the lack of information, which is quantified by entropy (or Kullback-Leibler divergence divergence) defined by

$$H(\mathbb{Q} \| \mathbb{P}) \triangleq \begin{cases} \mathbb{E}_{\mathbb{Q}}\left[\log(\frac{d\mathbb{Q}}{d\mathbb{P}})\right] & \text{if } \mathbb{Q} \ll \mathbb{P} \\ \infty & \text{else,} \end{cases} \tag{4.31}$$

see [89] for a discussion between information and entropy. For this reason, the minimizer of the relative entropy over the collection of ELMMs measure is called the *minimal martingale measure* and will be denoted by $\hat{\mathbb{P}}$.

Consequentially, under the assumptions of [88, Theorem 7], the existence of the minimal martingale measure is equivalent to the set of equivalent martingale measures being non-empty. This property, as shown in [26], is equivalent to the NFLVR formulation of no-arbitrage holding for X_t (see Appendix 4.8.1 or [26] for details on NFLVR. Alternatively, see [44] for a discussion on various no-arbitrage conditions).

We view this argument as a special usage of the arbitrage-free regularization framework by considering the class of deformations $\mathscr{A}_{\mathbb{P}}$ defined by mapping ϕ to a factor model $\Phi_t(\phi)$ which can be represented as a measure change.

Definition 4.5.1 (Family of ELMMs) *A family of measures $\mathscr{Q} \triangleq \{\mathbb{Q}_u\}_{u\in\mathcal{U}}$ such that for μ-a.e. u in $\mathcal{U}$, $X_t(u)$ is a local-martingale under $\mathbb{Q}_u$, will be called a family of LMMs. If moreover for μ-a.e. u in $\mathcal{U}$, $\mathbb{Q}_u$ is equivalent to $\mathbb{P}$, we call $\mathscr{Q}$ a family of ELMMs to $\mathbb{P}$.*

Definition 4.5.2 (Equivalent Measure-Deformations) *A deformation of ϕ is in $\mathscr{A}_{\mathbb{P}}$ if and only if there exists a family $\mathscr{Q} \triangleq \{\mathbb{Q}_u\}_{u\in\mathcal{U}}$ of equivalent measure, to $\mathbb{P}$ whose cádlág versions of their density process $Z_t^{\mathbb{Q}_u}$ satisfy*

$$\Phi_t^{\mathscr{Q}}(\phi)(t,\cdot,u) = \phi(t, Z_t^{\mathbb{Q}_u}\cdot, u).$$

Lemma 4.5.3. The operator

$$\mathbb{A}_{\phi,\mu}^{H,\hat{AF}}\left[\phi|\cdot\right] \triangleq \underset{\Phi_t\in\mathscr{A}_{\mathbb{P}}}{\text{arginf}}\, \mathbb{E}\left[\int_0^T \hat{H}\left(\Phi_t\right)dt\right] + \hat{AF}\left(\Phi_t^Q\right), \tag{4.32}$$

where

$$\begin{aligned} \hat{H}\left(\Phi_t\right) &\triangleq \begin{cases} \int_{u\in\mathcal{U}} H(\mathbb{Q}_u\|\mathbb{P})^2\mu(du) & \text{if } (\exists\Phi_t^{\mathscr{Q}}\in\mathscr{A}_{\mathbb{P}})\Phi_t^{\mathscr{Q}}(\phi) = \Phi_t \\ \infty & \text{else} \end{cases} \\ \hat{AF}\left(\Phi_t^Q\right) &\triangleq \begin{cases} 0 & \text{if } \int_{u\in\mathcal{U}} \Lambda\left(\Phi_t(\phi)(t,\beta_t,u)\right)\mu(du) = 0 \\ \infty & \text{else} \end{cases}, \end{aligned} \tag{4.33}$$

defines an arbitrage-free regularization operator on the class of deformations $\mathscr{A}_{\mathbb{P}}$.

Proof. The relative entropy $H(\cdot\|\cdot)$ is a functional Bregman divergence, as shown in [46], and therefore is convex in the first argument, non-negative and zero if and only if $\mathbb{Q}$ and $\mathbb{P}$ are the same up to a set of measure 0. Since x^2 is convex, then $H(\cdot\|\cdot)$ is also convex. Setting μ_t to be Lebesgue measure m, it follows that μ_t is deterministic and therefore $\mathscr{F}_t^X$-predictable. Hence, Lemma 4.4.2 implies that $\hat{H}(\mathbb{Q}\|\mathbb{P})$ is a model deviation.

Since $\hat{H}$ only takes on finite values for deformations which are identifiable with measure changes we will only consider those for the following and subsequently identify $\hat{H}$ with H. The non-anticipative functional $\hat{AF}_t$ is non-negative and takes value 0 if and only if $\mathbb{Q}$ is an ELMM, then it defines an arbitrage-penalty. The non-anticipative functional $\hat{\mathrm{AF}}_t$ takes constant values 0 or ∞ therefore for any $\lambda > 0$, a minimizer of

$$\mathbb{E}\left[\int_0^T H\left(Q\|\mathbb{P}\right)dt\right] + \frac{1}{\lambda}\hat{AF}\left(Z_t^Q\right)$$

and of

$$\mathbb{E}\left[\int_0^T H\left(Q\|\mathbb{P}\right)dt\right] + \hat{AF}\left(Z_t^Q\right)$$

must be equivalent up to $\sim$ as defined in Lemma 4.4.6. Therefore

$$\begin{aligned}
&\lim_{\lambda\downarrow 0^+}\operatorname*{arginf}_{\Phi_t^Q\in\cdot}\mathbb{E}\left[\int_0^T H\left(Q\|\mathbb{P}\right)dt\right]+\frac{1}{\lambda}\hat{AF}\left(Z_t^Q\right)\\
&=\operatorname*{arginf}_{\Phi_t^Q\in\cdot}\mathbb{E}\left[\int_0^T H\left(Q\|\mathbb{P}\right)dt\right]+\hat{AF}\left(Z_t^Q\right)\\
&=\mathbb{A}_{\phi,\mu}^{H,\hat{AF}}\left[\phi|\cdot\right]
\end{aligned}$$

Hence $\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\cdot\right]$ defines an arbitrage-free regularization operator, with $\mathscr{A}_{\mathbb{P}}\in Dom(F,\phi,\beta_t,g_t)$. □

Theorem 4.5.4 (Arbitrage-Free Regularization Formulation of NFLVR)**.** Let β_t be a continuous $\mathscr{M}$-valued semi-martingale satisfying regularity conditions 4.8.11 as well as 4.8.12, and assume that μ, σ and F are deterministic functions. Then for μ-a.e. u in $\mathcal{U}$, $X_t(u)$ satisfies NFLVR if and only if the arbitrage-free regularization

$$\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]$$

exists. Moreover, if $\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]$ does exist, then for μ-a.e. u in $\mathcal{U}$

1. $\mathbb{P}$ is an ELMM for every $X_t^{\mathscr{A}_{\mathbb{P}}}(u)$,
2. $\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]=\phi(t,Z_t^{\hat{\mathbb{P}}_u}\beta_t,u)$ where for every $Z_t^{\hat{\mathbb{P}}_u}$ is the density process of the minimal martingale measure for $X_t(u)$ relative to $\mathbb{P}$.

Proof. The proof of Theorem 4.5.4 will be deferred to the appendix. □

We take a moment to discuss it and examine one of its consequences. To reformulate NFLVR for portfolios of a (finite) number of assets we refine the class of deformations $\mathscr{A}_{\mathbb{P}}$ to a smaller subclass. A subset $\bar{\mathscr{A}_{\mathbb{P}}}$ of $\mathscr{A}_{\mathbb{P}}$ defined by

$$\bar{\mathscr{A}_{\mathbb{P}}}\triangleq\left\{\Phi_t^{\mathscr{Q}}\in\mathscr{A}_{\mathbb{P}}:(\exists\mathbb{Q}\sim\mathbb{P})\mu\left(\{u\in\mathcal{U}:\mathbb{Q}_u\neq\mathbb{Q}\}\right)=0\right\}.$$

We denote the elements of $\bar{\mathscr{A}_{\mathbb{P}}}$ by $\Phi_t^{\mathbb{Q}}$, where $\mathbb{Q}$ is the μ-a.e. unique measure equating to all the members of the family $\mathscr{Q}$.

Corollary 4.5.5 (Arbitrage-Free Regularization Formulation of NFLVR)**.** Consider a sub-market $M\triangleq\{X_t(u_1),\ldots,X_t(u_N)\}$ of $\{X_t(u)\}_{u\in\mathcal{U}}$. Then NFLVR holds on M if and only if

1. $\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]$ exists,
2. $\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]=\mathbb{A}_{\phi,\mu}^{H,AFF}\left[\phi|\bar{\mathscr{A}_{\mathbb{P}}}\right]$,

where μ is taken to be the measure defined on subsets B of $\mathcal{U}$ by

$$\mu(B) \triangleq \#(B \cap \{u_1, \ldots, u_N\}).$$

Proof. By Theorem 4.5.4 for each u_i, $X_t(u_i)$ satisfies NFLVR if and only if there exists a unique $\Phi_t^{\mathscr{Q}}$ in $\mathscr{A}_{\mathbb{P}}$ solving the arbitrage-free regularization problem defining $\mathbb{A}^{H,AFF}_{\phi,\mu}[\phi|\mathscr{A}_{\mathbb{P}}]$. The Fundamental Theorem of Asset Pricing implies that any portfolio on $\{X_t(u_i)\}$ satisfies NFLVR (jointly) if and only if there exists an ELMM $\mathbb{Q} \sim \mathbb{P}$ simultaneously making every $\{X_t(u_i)\}$ a local-martingale. [88, Theorem 7] implies that if such an ELMM exists then there exists a unique minimal martingale measure $\hat{\mathbb{P}}$ minimizing $H(\cdot\|\mathbb{P})$, across equivalent measures such that *each* $X_t(u_i)$ are simultaneously local-martingales. By definition of $\mathscr{A}_{\mathbb{P}}^{\mu}$, this implies that the unique element of $\mathscr{A}_{\mathbb{P}}$ solving the arbitrage-free regularization problem defining $\mathbb{A}^{H,AFF}_{\phi,\mu}[\phi|\mathscr{A}_{\mathbb{P}}]$ is an element the element of $\mathscr{A}_{\mathbb{P}}^{\mu}$ of the form $\Phi_t^{\mathscr{Q}}; \quad \mathbb{Q}_{u_i} = \hat{\mathbb{P}}; \quad i = 1, \ldots, N$. □

In Remark 4.3.5, the FDR-HJM models of [9] were related to flow models representing the price of a zero-coupon bond for which the additional assumptions that σ, μ, and ϕ are deterministic and constant in time are made. In [39] it is shown that most FDR-HJM models fail to be arbitrage-free. Results such as [39, 38, 18] required that there exist a unique ELMM simultaneously making every $\{X_t(u)\}_{u\in\mathcal{U}}$ into a local-martingale. However as pointed out in Corollary 4.5.5, NFLVR is equivalent to this requirement holding for a finite number of the members of the infinitely large market $\{X_t(u)\}_{u\in\mathcal{U}}$. On the other hand, Theorem 4.5.4 showed that $\mathbb{P}$ can be viewed as an ELMM by considering the arbitrage-free regularization $\phi(t, Z_t^{\hat{\mathbb{P}}_u}\beta_t, T)$ in place of $\phi(t, \beta_t, T)$. However, $\phi(t, Z_t^{\hat{\mathbb{P}}_u}\beta_t, T)$ fails to be an FDR-HJM model, since the empirical factor model $\phi(t, Z_t^{\hat{\mathbb{P}}_u}\beta, T)$ is itself predictable and therefore is not deterministic. Therefore arbitrage-free regularization of FDR-HJM models, viewed within the flow model framework, may be an appropriate relaxation of the FDR-HJM formulation which freely allows for the existence of arbitrage-free factor models for the FRC.

The Fundamental Theorem of Asset Pricing of [26], states that every contingent claim can be replicated by a portfolio of market assets if and only if there exists a unique ELMM. In particular, this implies that there exists a unique $\mathbb{Q} \sim \mathbb{P}$ under which (F, ϕ, β_t, g_t) is arbitrage-free. In the language of arbitrage-free regularization, there must exist a unique $\Phi_t^{\mathbb{Q}} \in \mathscr{A}_{\mathbb{P}}$ such that (F, ϕ, β_t, g_t) is a local-martingale. However, this implies that market completeness is equivalent to the independence of the choice of model deviation or arbitrage-penalty when defining an arbitrage-free regularization operator on $\mathscr{A}_{\mathbb{P}}$.

Theorem 4.5.6 (Arbitrage-Free Regularization Formulation of Market Completeness)**.** Let (F, ϕ, β_t, g_t) be a flow model. For μ-a.e. u in $\mathcal{U}$, the market generated by $X_t(u)$ is complete if and only if for every pair of arbitrage-free regularization operators $\mathbb{A}^{D,AF}_{\phi,\mu}[\phi|\mathscr{A}_{\mathbb{P}}]$ and $\mathbb{A}^{\tilde{D},\tilde{AF}}_{\phi,\mu}[\phi|\mathscr{A}_{\mathbb{P}}]$,

$$\mathbb{A}^{D,AF}_{\phi,\mu}[\phi|\mathscr{A}_{\mathbb{P}}] \underset{AF}{\overset{D}{\sim}} \mathbb{A}^{\tilde{D},\tilde{AF}}_{\phi,\mu}[\phi|\mathscr{A}_{\mathbb{P}}].$$

Proof. If for μ-a.e. u in $\mathcal{U}$ the market generated by $X_t(u)$ is complete, then the family minimal martingale measures $\mathscr{Q} \triangleq \{\hat{\mathbb{P}}_u\}_{u\in\mathcal{U}}$ for $\{X_t(u)\}_{u\in\mathcal{U}}$ is the unique family of ELMMs for $\{X_t(u)\}_{u\in\mathcal{U}}$. Equivalently, there exists a unique $\Phi_t^{\mathscr{Q}} \in \mathscr{A}_{\mathbb{P}}$ such that $X_t^{\mathscr{A}_{\mathbb{P}}}(u)$ is a local-martingale. Assume that there exists and AF penalty and a deformation $\Phi_t^{\tilde{\mathscr{Q}}}$ of (F, ϕ, β_t, g_t) in $\mathscr{A}_{\mathbb{P}}$ minimizing $\mathbb{A}_{\phi}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]$ such that for μ-a.e. u in $\mathcal{U}$ $\mathbb{Q}_u$ are not the minimal-martingale measures. It follows that

$$\lim_{\lambda\mapsto 0^+} \inf_{\Phi_t^{\tilde{\mathscr{Q}}}\in\mathscr{A}} D_t(\Phi_t^{\tilde{\mathscr{Q}}}) + \frac{1}{\lambda}AF_t(\Phi_t) = \lim_{\lambda\mapsto 0^+} \quad D_t(\Phi_t^{\hat{Q}}) + \frac{1}{\lambda}AF_t(\Phi_t^{\hat{Q}}) \geq \lim_{\lambda\mapsto 0^+}\frac{1}{\lambda}AF_t(\Phi_t^{\hat{Q}}).$$

Since AF_t takes finite values if and only if $\{X_t^{\mathscr{A}_{\mathbb{P}}}(u)\}_{u\in\mathcal{U}}$ are local-martingales for μ-a.e u in $\mathcal{U}$, which by market completeness only happened if for μ-a.e. u in $\mathcal{U}$ $\tilde{\mathbb{Q}}_u$ is $\hat{\mathbb{P}}_u$. Therefore the LHS of equation (4.5.2) must be infinite for $\Phi_t^{\hat{Q}}$. This is a contradiction of the finiteness condition Definition 4.4.10 (ii). Therefore there exists a μ-a.e. unique minimizer of every arbitrage-free regularization with respect to $\mathscr{A}_{\mathbb{P}}$ in a complete market.

Conversely, assume that every arbitrage-free regularization operator on $\mathscr{A}_{\mathbb{P}}$ has a unique value (F, ψ, β_t, g_t), up to the equivalence relation $\sim$, of Lemma 4.4.6. In the first case that for μ-a.e. u in $\mathcal{U}$, $\{X_t(u)\}_{u\in\mathcal{U}}$ is a $\mathbb{P}$-local martingale, then, by Definition 4.4.4 (ii), it follows that for any arbitrage-penalty AF_t,

$$AF_t(\tilde{I}) = 0. \tag{4.34}$$

Similarly, by Definition 4.4.1 (ii), it follows that for any model deviation D_t,

$$D_t(\tilde{I}) = 0. \tag{4.35}$$

Since any model deviation and any arbitrage-penalty are non-negative, then Equations (4.34) and Equations (4.35) imply that for every model deviation D_t, every arbitrage-penalty AF_t, every $t \geq 0$ and every $\lambda > 0$

$$0 = D_t(\tilde{I}) + \frac{1}{\lambda}AF_t(\tilde{I}) \leq \underset{\Phi_t^Q\in\mathscr{A}_{\mathbb{P}}}{\operatorname{arginf}} D_t(\Phi_t^Q) + \frac{1}{\lambda}AF_t(\Phi_t^Q). \tag{4.36}$$

It follows that for every D_t and AF_t, $X_t(u)$ must be in the same equivalence class as $\mathbb{A}_{\phi}\left[\phi|\mathscr{A}_{\mathbb{P}}\right]$.

For the case where $\mathbb{P}$ itself is not an ELMM, assume that the market is not complete. Then, there exist families of ELMMs for which the set

$$\left\{u \in \mathcal{U} : \mathbb{Q}_u \neq \tilde{\mathbb{Q}}_u\right\}$$

has positive μ-measure. For a family of ELMMs $\mathscr{Q} \triangleq \{\mathbb{Q}_u\}_{u\in\mathcal{U}}$, define the non-anticipative functional

$$D_t^{\mathscr{Q}}(\Phi_t) \triangleq \begin{cases} 0 & \text{if } \Phi_t = \tilde{I}, \\ \int_{u\in\mathcal{U}} H(\tilde{\mathbb{Q}}_u\|\mathbb{Q}_u)^2\mu(du) & \text{if } (\exists \Phi_t^{\tilde{\mathscr{Q}}} \in \mathscr{A}_{\mathbb{P}})\Phi_t^{\tilde{\mathscr{Q}}}(\phi) = \Phi_t \text{ and } \tilde{\mathbb{Q}}_u \neq \mathbb{P}, \mu - a.e, \\ \infty & \text{else,} \end{cases}$$

where we have abbreviated $\{\tilde{\mathbb{Q}}_u\}_{u\in\mathcal{U}}$ by $\tilde{\mathscr{Q}}$.

The non-anticipative functional $D_t^{\mathscr{Q}}$ is non-negative, convex and has value 0 if (F,ψ,β_t,g_t) and (F,ϕ,β_t,g_t) define the same price processes $\mathbb{P}\otimes m\otimes\mu$-a.e. Hence, $D_t^{\mathscr{Q}}$ defines a model deviation. Moreover, any arbitrage-penalty AF_t such that Definition 4.4.10 (ii) holds,

$$\mathbb{A}_{\phi,\mu}^{\mathscr{Q},AF}\left[\Phi|\mathscr{A}\right] \triangleq \lim_{\lambda\mapsto 0^+} \underset{\Phi_t\in\mathscr{A}}{\operatorname{arginf}} D_t^{\mathscr{Q}}(\Phi_t) + \frac{1}{\lambda}AF_t(\Phi_t),$$

defines an arbitrage-free regularization operator. Moreover, by construction the only unique minimizers of $D_t^{\mathscr{Q}}$ are the deformations $\tilde{I}_t$ and $\Phi_t^{\mathscr{Q}}$. Since $\mathscr{Q}$ is a family of ELMMs and $\mathbb{P}$ is not, it follows that, for any arbitrage-penalty AF_t

$$AF_t\left(\tilde{I}_t\right) > 0; \quad AF_t\left(\Phi_t^{\mathscr{Q}}\right) = 0. \tag{4.37}$$

Therefore,

$$\begin{aligned}
\lim_{\lambda\mapsto 0^+} D_t^{\mathscr{Q}}\left(\tilde{I}_t\right) + \frac{1}{\lambda}AF_t\left(\tilde{I}_t\right) \geq \lim_{\lambda\mapsto 0^+}\frac{1}{\lambda}AF_t\left(\tilde{I}_t\right) &= \infty\\
\lim_{\lambda\mapsto 0^+} D_t^{\tilde{\mathscr{Q}}}\left(\Phi_t^{\tilde{\mathscr{Q}}}\right) + \frac{1}{\lambda}AF_t(\Phi_t^{\tilde{\mathscr{Q}}}) = D_t^{\tilde{\mathscr{Q}}}\left(\Phi_t^{\tilde{\mathscr{Q}}}\right) &> 0\\
\lim_{\lambda\mapsto 0^+} D_t^{\mathscr{Q}}\left(\Phi_t^{\mathscr{Q}}\right) + \frac{1}{\lambda}AF_t\left(\Phi_t^{\mathscr{Q}}\right) &= 0,
\end{aligned}$$

where $\tilde{\mathscr{Q}} \triangleq \{\tilde{\mathbb{Q}}_u\}_{u\in\mathcal{U}}$ is any family of ELMMs for which the set

$$\{u\in\mathcal{U} : \tilde{\mathbb{Q}}_u \neq \mathbb{Q}_u\}$$

has positive μ-measure. Hence, $\Phi_t^{\mathscr{Q}}$ is the unique minimizer of $\mathbb{A}_{\phi,\mu}^{\mathscr{Q},AF}\left[\cdot|\mathscr{A}\right]$. Therefore for distinct families of ELMMs $\mathscr{Q}$ and $\tilde{\mathscr{Q}}$,

$$\left(F,\Phi_t^{\mathscr{Q}}\left(\phi\right),\beta_t,g_t\right) \underset{AF_t}{\overset{D_t^{\mathscr{Q}}}{\nsim}} \left(F,\Phi_t^{\tilde{\mathscr{Q}}}\left(\phi\right),\beta_t,g_t\right),$$

contradicting the assumption that every arbitrage-free regularization operator of (F,ϕ,β_t,g_t) has a unique value up to $\sim$ on $\mathscr{A}_{\mathbb{P}}$. Therefore, there must exist a unique family of ELMMs. Hence, for μ-a.e. u in $\mathcal{U}$ the market generated by $X_t(u)$ must be complete by the Fundamental Theorem of Asset Pricing part 2. □

Theorem 4.5.4 reformulated the FTAP in terms of the existence and uniqueness of a particular arbitrage-free regularization problem which could be understood as deformations by measure change, for equivalent measures to $\mathbb{P}$. This intuition of the FTAP as particular types of deformations of a model may be extended to other types of deformations under which the flow model becomes arbitrage-free. This is explored in the next section.

The Fundamental Theorem of Asset Pricing gives mathematical meaning to risk-neutral pricing through conditions for the existence and uniqueness of the *risk-neutral* *value* of a contingent claim with payoff function f, on an asset whose price process follows X_t defined by

$$v_T \triangleq \mathbb{E}_{\hat{\mathbb{P}}}\left[f(X_T) \mid \mathscr{F}_t\right]. \tag{4.38}$$

When X_t can be represented by a flow model (F, ϕ, β_t, g_t), Theorems 4.5.5 and 4.5.6 state that, if $X_t^{\Phi}(u)$ satisfies NFLVR, then the risk-neutral pricing formula may be expressed as

$$v_T(u) = \mathbb{E}_{\hat{\mathbb{P}}}\left[f\left(X_T(u)\right) \mid \mathscr{F}_t\right] = \mathbb{E}_{\mathbb{P}}\left[f\left(X_T^{\mathscr{A}_{\mathbb{P}}}(u)\right) \mid \mathscr{F}_t\right], \tag{4.39}$$

with this formulation being unique if and only if the market is complete. Equation (4.39) implies that pricing $f(X_T(u))$ under the minimal martingale measure is equivalent to pricing the arbitrage-free regularization $f\left(X_T^{\mathscr{A}_{\mathbb{P}}}(u)\right)$ directly under $\mathbb{P}$, by minimally deforming the factor model ϕ according to $\mathscr{A}_{\mathbb{P}}$ instead of requiring a measure change from $\mathbb{P}$ to $\hat{\mathbb{P}}$.

More generally, Theorem 4.4.11 implies that for any class of deformations $\mathscr{A}$, $\mathbb{P}$ is always an ELMM for $X_t^{\mathscr{A}}(u)$. Therefore the right-hand side of equation (4.39) is always the risk-neutral price for the model $X_t^{\mathscr{A}}(u)$. Therefore the right-hand side of equation (4.39) provides an alternative to the classical risk-neutral pricing formula when the change of measure from $\mathbb{P}$ to $\hat{\mathbb{P}}$ is intractable, or does not exist. Moreover, there are other classes of deformations for which $\mathscr{A}_{\mathbb{P}}$ may provide better forecasts than pricing using $X_t^{\mathscr{A}_{\mathbb{P}}}(u)$. Hence, pricing derivatives on $X_t(u)$ is equivalent to the price of derivatives on the minimally deformed model $X_t^{\mathscr{A}}(u)$, is a consistent extension of the classical change-of-measure approach to risk-neutral pricing.

Definition 4.5.7 (Proximal Risk-Neutral Pricing Formula) *Let (F, ϕ, β_t, g_t) be a flow model, $\mathscr{A}$ be a class of deformations in Dom(F, ϕ, β_t, g_t), and $\mathbb{A}_\phi[\cdot|\cdot]$ an arbitrage-free regularization operator. Let f be a Borel-measurable function representing the payoff of a contingent claim at time T on the underlying asset whose price follows $X_t(u)$. If*

1. ***(Proximal No-Arbitrage)*** *the arbitrage-free regularization $\mathbb{A}_\phi[\phi|\mathscr{A}]$ exists and*

2. ***(Proximal Market Completeness)*** *the process defined by the arbitrage-free regularization is independent of choice of model deviation and arbitrage-penalty up to $\sim$,*

then the proximal risk-neutral price of the contingent claim $f(X_T(u))$ is defined to be

$$v_t(u|\mathscr{A}) \triangleq \mathbb{E}_{\mathbb{P}}\left[f\left(X_T^{\mathscr{A}}(u)\right) \mid \mathscr{F}_t\right]. \tag{4.40}$$

To motivate equation (4.40), a particular arbitrage-free regularization problem both admitting a closed form solution and not requiring the existence of an ELMM will now be developed. This arbitrage-free regularization problem's closed form expression will be the central component of efficient arbitrage-free estimation procedures introduced in the final section of this chapter.

Moving Beyond Measure Changes with Spread Deformations

When the dynamics factors β_t are assumed to follow an OU-process, a correction to this drawback is proposed in [21] by adding a deterministic spread $C(t,T)$ over the Nelson-Siegel curve $\phi_{NS}(t,\beta_t,T)$, the resulting model is called the Arbitrage-Free Nelson-Siegel model (AFNS). The addition of a spread over the factor model is a type of deformation which we illustrate in more generality here.

Definition 4.5.8 (Spread Deformations) *The set of maps*

$$\left\{\phi \mapsto \phi + C(t,u) : u \mapsto C(t,u) \in C^1(\mathcal{U};\mathcal{U}) \text{ and } t \mapsto C(t,u) \text{ is } \mathscr{F}_t^{\beta}\text{-predictable}\right\},$$

under the equivalence relation defined in Lemma 4.4.6, is called the class of spread deformations. It will be denoted by $\mathscr{A}_+$.

Lemma 4.5.9 (Least-Squares Arbitrage-Free Regularization)**.** The non-anticipative functional

$$D_t^2 \triangleq \mathbb{E}\left[\int_0^t \int_{u\in\mathcal{U}} (\psi(t,\beta_t,u) - \phi(t,\beta_t,u))\frac{d\mu_t}{dm}\mu(du)ds\right],$$

defines a model deviation. In particular, if AF_t is as in Theorem 4.4.5 and Λ_B represents the left-hand side equation (4.18), then

$$\begin{aligned}\mathbb{A}_{\phi,\mu}^{2,B}\left[\Phi|\mathscr{A}\right] &\triangleq \lim_{\lambda\mapsto 0^+}\inf_{\Phi_t\in\mathscr{A}} D_t^2(\Phi_t(\psi)) + \frac{1}{\lambda}AF_t^B(\Phi_t(\psi))\\ AF_t^B(F,\psi,\beta_t,g_t) &\triangleq \mathbb{E}\left[\int_{u\in\mathcal{U}}\int_0^t \left\{\Lambda_B(F,\psi,\beta_s,g_s)\right\}^2 ds\mu(du)\right] \qquad (4.41)\end{aligned}$$

defines an arbitrage-free regularization operator.

Proof. Since x^2 is strictly convex, Lemma 4.4.2 implies that D_t^2 must be a model deviation. Corollary 4.3.12 shows that Λ_B is zero if and only if Λ is zero when (F,ϕ,β_t,g_t) is the flow model of Example 4.3.4. Therefore, by Theorem 4.3.11 and Theorem 4.4.5, $\mathbb{A}_{\phi,\mu}^{2,B}\left[\Phi|\mathscr{A}\right]$ defines an arbitrage-free regularization operator for the flow model of Example 4.8.6. □

Theorem 4.5.10. Let (F,ϕ,β_t,g_t) be a flow model such that

1. For every $t \geq 0$, $\mathscr{D}_t F = 0$,
2. $\mathscr{V}F(t,\varphi_s^u,[\varphi_s^u]) \neq 0$, $m \otimes \mathbb{P}$-a.e.
3. $\phi(t,\beta,u) = \sum_{i=1}^N \beta^i\varphi_i(u)$ where $\{\varphi_i\}_{i=1}^N$ is a linearly-independent set in $W^s(\mathcal{U})$.

The class of spread deformations $\mathscr{A}_+$ is a class of deformation of the flow (F, ϕ, β_t, g_t). Moreover, if $X_t(u)$ is not arbitrage free, then the arbitrage-free regularization $\mathbb{A}^{2,B}_{\phi,\mu}\left[\phi|\mathscr{A}_+\right]$ is given by

$$\begin{aligned} \mathbb{A}^{2,B}_{\phi,\mu}\left[\Phi|\mathscr{A}\right] =&\phi(t,\beta_t,u) + \hat{C}(t,u) \\ \hat{C}(t,u) =& -\int_0^t \sum_{i=0}^{d} (\xi_s e_i)\phi(s,\beta,u)d\mu(s,\beta,u) \\ &+\frac{1}{2}\sum_{i,j=1}^{d} Hess^{g_t}\left(\phi(s,u)\right)(\beta e_i, \beta e_j)d[\beta^i,\beta^j]_s ds\|_{\beta=\beta_s}. \end{aligned} \tag{4.42}$$

Proof. Since $C(t,u) = 0$ is both predictable and smooth, then the trivial deformation is in $\mathscr{A}_+$. Therefore, Definition 4.4.8 part (i) holds.

If $\Phi_t(\phi) \mapsto \varphi(t,u,\beta^\varphi_t) + C_g(t,u)$ and $h_t(\phi) \mapsto \varphi(t,u,\beta^\varphi_t) + C_h(t,u)$ then the map

$$g_t \circ h_t(\phi) \mapsto \varphi(t,u,\beta^\varphi_t) + \left(C_g(t,u) + C_h(t,u)\right), \tag{4.43}$$

is in $\mathscr{A}_+$ and Definition 4.4.8 part (iii) holds.

If $X_t(u)$ is arbitrage-free, then (ii) of Definition 4.4.8 holds. Suppose that $X_t(u)$ fails to be arbitrage-free. Since ϕ is of the form $\phi(t,\beta,u) = \sum_{i=1}^N \beta^i\varphi_i(u)$, then

$$\left(\frac{1}{2}tr[{}^t\mathscr{V}^2_s F(s,\varphi^u_s,[\varphi^u]_s)]\left[\sum_{i=0}^{d}(\xi_s e_i)\phi(s,\beta_s,u)d\sigma(s,\beta_s,u)\right]^2 ds\right) = 0. \tag{4.44}$$

Plugging equation (4.44) into equation (4.17) and using assumption $(i)-(ii)$ it follows that for any $g_t \in \mathscr{A}_+$, $\Phi_t(\phi) = \phi + C(t,u)$ for any fixed $\beta \in \mathscr{M}$, equation (4.17) can only hold if

$$\begin{aligned} -\frac{\partial C(t,u)}{\partial t} = -\frac{\partial \Phi_t(\phi)(t,\beta,u)}{\partial t} = &\left[\sum_{i=0}^{d}(\xi_s e_i)\phi(s,\beta_s,u)d\mu(s,\beta_s,u)\right. \\ &\left. +\frac{1}{2}\sum_{i,j=1}^{d} Hess^{g_t}\left(\phi(s,u)\right)(\beta_s e_i, \beta_s e_j)d[\beta^i,\beta^j]_s\right]. \end{aligned} \tag{4.45}$$

Fixing β in equation (4.45) and integrating with respect to t, implies that $C(t,u)$ must be of the form described in equation (4.42). Therefore Definition 4.4.8 part (ii) holds.

Moreover, since $\Phi_t(\phi) \triangleq \phi + \hat{C}(t,u)$ is the only element of $\mathscr{A}_+$ on which the arbitrage penalty L is not infinite, then it follows that it must solve the minimization problem defining the arbitrage-free regularization operator $\mathbb{A}^{2,B}_{\phi,\mu}\left[\Phi|\mathscr{A}\right]$. □

The existence of the solution to the arbitrage-free regularization problem of Proposition 4.5.10 does not depend on the existence of an ELMM to $\mathbb{P}$ for $X_t(u)$. Therefore, like the NS model for the price of a bond, $X_t(u)$ is permitted to admit arbitrage-opportunities. However, unlike the AFNS correction to the NS model, the construction of $X_t^{\mathscr{A}_+}(u)$ does not require (F, ϕ, β_t, g_t) to represent the price of a bond or for the stochastic factors β_t to follow an OU process. Instead it works for any price process representable by a sufficiently time-homogeneous flow model.

Example 4.5.11 (Extended Arbitrage-Free Nelson-Siegel Models)**.** The extended Nelson-Siegel models satisfy the assumptions of Proposition 4.5.10. Therefore the arbitrage-free regularization problem of Proposition 4.5.10 applied to the Extended Nelson-Siegel admit the following closed-form solution

$$\begin{aligned}
\mathbb{A}^{2,B}_{\phi,\mu}\left[\phi|\mathscr{A}_+\right] =&\sum_{i=1}^{N}\beta_t^i\varphi_i(T) + \int_0^t\left[\sum_{i=1}^{d}\phi_i(T)\mu(s,T) + \frac{1}{2}\sum_{i,j=1}^{d}\phi_i(T)\phi_j(T)\sigma_i(s,T)\sigma_j(s,T)\right]ds\\
\phi_1 =&1\\
\phi_2 =&\frac{\left[1-\exp\left(-T/\tau\right)\right]}{T/\tau}\\
\phi_3 =&\left(\frac{\left[1-\exp\left(-T/\tau\right)\right]}{T/\tau} - \exp\left(-T/\tau\right)\right)\\
\phi_i =&\frac{\left[1-\exp\left(-T^{k_i}/\tau\right)\right]}{T^{k_i}/\tau}; i>3, k_i>0.
\end{aligned} \tag{4.46}$$

Equation (4.46) is the FRC formulation of the AFNS model, when $d = 3$ and β_t follows an OU-process.

In [38] it is shown that the Nelson-Siegel model of equation (4.5) admits arbitrage opportunities and therefore the Fundamental Theorem implies that there does not exist and ELMM to $\mathbb{P}$ making the Nelson-Siegel model arbitrage-free. In contrast, Lemma 4.5.9 showed that deforming the Nelson-Siegel model with respect to $\mathscr{A}_+$ gave a arbitrage-free bond model. Therefore, the following conclusion is taken.

Theorem 4.5.12 (Beyond Measure Changes)**.** There exist flow models (F, ϕ, β_t, g_t) for which there does not exist an ELMM to $\mathbb{P}$ making $X_t(u)$ a local-martingale, but there exists a class of deformations $\mathscr{A}$ for which $\mathbb{A}_\phi\left[\phi|\mathscr{A}\right]$ exists and is a $\mathbb{P}$-local-martingale.

In summary, arbitrage-free regularization with respect to $\mathscr{A}_{\mathbb{P}}$ was consistent with the Fundamental Theorem of Asset Pricing, as well as the risk-neutral pricing formula. Moreover, arbitrage-free regularization with respect to $\mathscr{A}_+$ provided a closed-form expression extending the classical risk-neutral pricing formula to flow models which may admit arbitrage. This provided an alternative procedure for constructing the AFNS correction to the NS model. Moreover, this procedure gave a closed-form arbitrage-free correction to solve a broad scope of models across many asset classes.

4.6 Empirical Performance

This section investigates the empirical performance of arbitrage-free regularization. It is observed that arbitrage-free regularization extracts subtle information not detectable by classical learning algorithms. Implementations are in the forward-rate curve setting.

The incorporation of no-arbitrage information into our estimates is desirable not only from a theoretical point-of-view, but also from a forecasting and curve-fitting perspective since they assimilate more information about the market. A meta-algorithm which takes an estimation procedure for the factors at time t_{i+1} given the current time t_i and returns an arbitrage-free estimate of the FRC at time t_{i+1} is introduced, implemented, and tested. It relies on a class of deformations, denoted by $\mathscr{A}^+$, which provide more flexibility than the class of spread deformation. The class of deformations $\mathscr{A}^+$ is defined, on the flow model of Example 4.5.11 by

$$\phi(t,\beta,u) \triangleq \sum_{i=1}^{d} \beta^i \varphi_i(T) \rightsquigarrow \sum_{i=1}^{d} \left(\beta^i + \alpha_t^i\right) \varphi_i(T) + C(t,T), \tag{4.47}$$

where α_t is $\mathscr{F}_t$-predictable. A key remark in Equation (4.47) is that the optimal spread $\hat{C}(t,T)$ of Proposition 4.5.10 is only a function of ϕ, μ, σ and the state of the dynamic factors β_t at time t. To emphasis this, for a given state of β_t, we will denote $\hat{C}(t,T,\beta_T)$ by $\hat{C}(t,T;\beta_t)$.

Meta-Algorithm 4.6.1 (Arbitrage-Free Estimation)**.**

```
Input: Bond data $B\left(t_i, T_j\right)_{i,j}$; an estimation algorithm $A$ for $\beta_t$; a strictly-convex
       penalty $P$ in $\mathbb{R}^d$
Output: An arbitrage-free FRC at time $t_{i+1}$
for every past time $t_i$ do
    1. (Stochastic Factor Estimation:)
        (a) Empirical Estimation Step: Estimate $\beta_{t_{i+1}}$ according to algorithm $A$.
        (b) (Spread Optimization:) Find $\alpha_{t_{i+1}}(\gamma) \in \mathbb{R}^N$ by minimizing (resp.
            heuristically reduce) the loss function:
                $\sum_{j=1} \left[C(t, T_j, \beta_{t_{i+1}} + \alpha_{t_{i+1}}(\gamma))\right]^2 + \gamma P\left(\alpha_t(\gamma)\right)^2 .$    (4.48)
            Minimize (resp. heuristically improve) $\alpha(\gamma)$ via cross-validation, to obtain
            $\hat{\alpha}_{t_{i+1}}$,
    2. (Arbitrage-Free Estimate:) Predict the curve at time $t_{i+1}$ to be
            $\sum_{i=1}^{N} \left(\beta^i_{t_{i+1}} + \hat{\alpha}^i_{t_{i+1}}\right) \varphi_i(T) + \hat{C}(t, T, \beta_{t_{i+1}} + \alpha_{t_{i+1}}).$
Choose $d$ with sequential-validation.
```

Every perturbation $\beta_t \rightsquigarrow \beta_t + \alpha_t$ impacts not only the loadings, but the optimal spread by perturbing it from $\hat{C}(t, T; \beta_t)$ into $\hat{C}(t, T; \beta_t + \alpha_t)$. This allows a method for computing (resp. heuristically approximating to an arbitrage-free solution) $\mathbb{A}^{2,B}_{\phi,\mu}\left[\phi|\mathscr{A}^+\right]$; thus providing an arbitrage-free estimate of the FRC once β_t is estimated according to some estimation scheme. This procedure is summarized in Meta-Algorithm 4.6.1.

Remark 4.6.2. Equation (4.48) balances the prediction quality with arbitrage-free correction by parsimoniously estimating dynamical factors which simultaneously predict well and decrease the magnitude of the optimal spread $C(t, T)$ over the curve ϕ.

For this empirical study, we consider successive 14054 business days of US interest-rate data, from January 1^{st}, 1970 until April 12^{th}, 2018. The observed maturities are at $1, 2, 3, 5, 7, 10, 20,$ and 30 years.

Estimation: Bi-Monthly

We investigate the proportion of monthly in-sample estimates for which the empirical factor model has a larger sum-of-squared errors than $\mathbb{A}^{2,B}_{\phi,\mu}\left[\phi|\mathscr{A}^+\right]$, in greater detail. The assumption that the instances when one of these models outperforms the other are i.i.d.

binomial random variables with proportion p is justified by the Wald-Wolfowitz, whose results are reported in Table 4.2 (see [101] for more details on this non-parametric test).

	1	2	3	5	7	10	20	30
Runs	8	6	7	16	4	12	6	8
p-value	$< e{-}5$	$< e{-}5$	$< e{-}5$	$< e{-}5$	$< e{-}5$	$< e{-}5$	$< e{-}5$	$< e{-}5$

Table 4.2: Wald-Wolfowitz Run Test Summary.

For this empirical investigation, we implement the Nelson-Siegel factor model of Example (4.46), with $N = 2$. The convex penalty function P in equation (4.48) is taken to be the ℓ^2 penalty on the vector α_t. The estimation step (1a), this time, is initialized using a moving window regression on two-month moving windows. The dynamics of β_t are assumed to follow an uncorrelated OU process

$$\beta_t = \beta_0 + \int_0^t A(K - \beta_s)ds + \int_0^t \Sigma dW_s,$$

where Σ is a diagonal matrix. A, K, Σ are estimated using maximum likelihood estimation. We find that, unlike the previous daily estimation procedure, the bi-monthly moving windows yields a significant increase in the proportion of times the arbitrage-free regularized model outperforms the Empirical factor model. We denote the sample proportion by $\hat{p}$.

Maturity	$\hat{p}$	St. Dev. of Estimate	99.9% Lower	99.9% Upper
1 Year	*0.826*	0.061	0.763	0.884
2 Year	0.829	*0.069*	0.757	0.894
3 Year	0.905	0.047	0.855	0.948
5 Year	0.956	0.033	0.920	0.984
7 Year	0.968	0.030	0.934	0.992
10 Year	**0.987**	**0.018**	0.967	1.000
20 Year	0.982	0.031	0.947	1.000
30 Year	0.918	0.053	0.860	0.965

Table 4.3: Estimated proportion when has lower squared error than ϕ does.

Table 4.3 presents 99.9% confidence intervals for the true probability that $\mathbb{A}_\phi\left[\phi|\mathscr{A}^+\right]$ outperforms ϕ. We make the *assumption* that this proportion follows a binomial distribution. Estimates are computed using Wilson's score interval which provides more accurate confidence intervals than the normal approximation where the population is binomial (see [102] for details on Wilson's score interval). The findings of Table 4.3 reinforce the fact that the arbitrage-free regularized model nearly always outperforms its empirical counterpart. However, a closer look a the statistics of Table 4.3 imply that there are indeed moments where the empirical factor model outperforms its arbitrage-free regularization.

Prediction: Daily

The one-day ahead forecasting performance of various arbitrage-free regularization operators will be compared against each other and benchmarked against their empirical factor model counterpart. For this implementation. We assume that ϕ is the extended Nelson-Siegel model and that β_t follows a multivariate OU process, as in Example 4.46. The stochastic factor estimation of Algorithm 4.6.1 step $(1a)$, is performed using a Kalman smoother. The Kalman smoother is initialized using the following sequence of algorithms. First the dynamics factors β_t are estimated using regression. The regression estimates are used to initialize the maximum likelihood method. The maximum likelihood estimates are then used to initialize the Kalman filter. Finally the Kalman smoother is used to estimate the one-day-ahead parameters, it is initialized with the estimates of the Kalman smoother. Step $(1b)$ is performed using heuristic methods to estimate α.

The parameters $0.5 \leq \gamma \leq 1.5$ and $0 \leq N \leq 20$ are chosen through cross-validation on a grid of possible values; the initial bounds for the grid where chosen empirically. The optimal values for γ and d were found to be 0.7 and 10, respectively.

Table 4.4 shows that for nearly every maturity, the absolute mean error and standard deviation of the errors is lower for the $\mathbb{A}_{\phi,\mu}^{2,B}\left[\phi|\mathscr{A}^{+}\right]$ model than they are for the Empirical and the extended AFNS model $\mathbb{A}_{\phi,\mu}^{2,B}\left[\phi|\mathscr{A}_{+}\right]2$ models. In general, $\mathbb{A}_{\phi,\mu}^{2,B}\left[\phi|\mathscr{A}_{+}\right]$ performs poorer than $\mathbb{A}_{\phi,\mu}^{2,B}\left[\phi|\mathscr{A}^{+}\right]$ and ϕ. This is because for every value of β_t, there exists exactly one spread $C(t,T,\beta_t)$ correcting for the existence of arbitrage in the bond price. This makes $\mathbb{A}_{\phi,\mu}^{2,B}\left[\phi|\mathscr{A}_{+}\right]$ inflexible, not allowing it to simultaneously describe the data and meet the no-arbitrage requirement. On the other hand, $\mathbb{A}_{\phi,\mu}^{2,B}\left[\phi|\mathscr{A}^{+}\right]$ regularized does not suffer from this limitation. The performance is significantly better than the alternatives for all maturities below 10 years and competitively better for those on and above 10 years. The empirical factor model does exhibit a lower variance for long-end maturities.

	1 Year			**2 Year**		
Error Statistics	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical
Mean	-0.704	**0.00000**	-0.704	**0.141**	0.845	-0.704
St.dev	0.073	**0.006**	0.073	0.063	**0.028**	0.074
95% Upr	-0.560	0.013	-0.560	0.265	0.900	-0.559
95% Lwr	-0.847	-0.013	-0.847	0.017	0.790	-0.848
99% Upr	-0.515	0.017	-0.515	0.303	0.917	-0.514
99% Lwr	-0.892	-0.017	-0.892	-0.022	0.773	-0.893
AIC	11,059.310	**-37,283.540**	11,059.310	**-4,602.155**	12,888.850	11,058.020

	3 Year			**4 Year**		
Error Statistics	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical
Mean	**0.449**	1.152	-0.704	**0.449**	1.152	-0.704
St.dev	0.051	**0.050**	0.072	0.051	**0.050**	0.072
95% Upr	0.549	1.251	**0.562**	0.549	1.251	-0.562
95% Lwr	0.348	1.054	-0.846	0.348	1.054	-0.846
99% Upr	0.581	1.281	-0.517	0.581	1.282	-0.517
99% Lwr	0.317	1.023	-0.891	0.317	1.023	-0.891
AIC	**6,452.629**	16,117.440	11,081.030	**6,439.378**	16,089.490	11,063.310

	5 Year			**7 Year**		
Error Statistics	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical
Mean	**0.618**	1.321	-0.702	**0.643**	1.347	-0.707
St.dev	**0.034**	0.079	0.073	**0.034**	0.100	0.073
95% Upr	0.684	1.476	-0.559	0.710	1.543	-0.563
95% Lwr	0.552	1.167	-0.846	0.577	1.152	-0.850
99% Upr	0.704	1.524	-0.514	0.731	1.604	-0.518
99% Lwr	0.531	1.118	-0.891	0.556	1.090	-0.895
AIC	**9,674.864**	17,505.120	11,039.740	**10,096.060**	17,714.500	11,101.500

	10 Year			**20 Year**		
Error Statistics	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical
Sample Mean	**0.666**	1.370	-0.699	**0.673**	1.377	-0.711
St.dev	**0.048**	0.118	0.077	0.083	0.154	**0.071**
95% Upr	0.759	1.601	-0.549	0.836	1.679	-0.572
95% Lwr	0.572	1.138	-0.850	0.511	1.075	-0.850
99% Upr	0.789	1.674	-0.502	0.887	1.774	-0.529
99% Lwr	0.543	1.065	-0.897	0.460	0.980	-0.894
AIC	**10,459.950**	17,894.160	11,002.570	**10,628.570**	17,977.620	11,163.980

	30 Years		
Error Statistics	$\mathscr{A}^+$	$\mathscr{A}_+$	Empirical
Sample Mean	**0.692**	1.396	-0.699
St.dev	0.085	0.158	**0.075**
95% Upper	0.860	1.706	-0.551
95% Lower	0.525	1.087	-0.846
99% Upper	0.913	1.803	-0.505
99% Lower	0.472	0.989	-0.893
AIC	**10,915.270**	18,119.850	10,989.520

Table 4.4: One day ahead FRC prediction errors.

In Table 4.4, $\mathscr{A}^+$ denotes the model $\mathbb{A}^{2,B}_{\phi,\mu}[\phi|\mathscr{A}^+]$, $\mathscr{A}_+$ denotes the model $\mathbb{A}^{2,B}_{\phi,\mu}[\phi|\mathscr{A}_+]$, and Empirical denotes ϕ. Recall that $\mathbb{A}_{\phi,\mu}^{2,B}[\phi|\mathscr{A}_+]$ reduces to the AFNS model when $N = 2$. Table 4.4 shows that $\mathbb{A}_{\phi,\mu}^{2,B}[\phi|\mathscr{A}^+]$ contains more information about the data, in the sense of information theory than its naive or extended AFNS counterparts with the exception of the 1 year maturity bond, as is reflected in the lower AIC score. Expressed differently, the low AIC of $\mathbb{A}_{\phi,\mu}^{2,B}[\phi|\mathscr{A}^+]$ implies that it is the most parsimonious amongst the naive model and extended AFNS model $\mathbb{A}_{\phi,\mu}^{2,B}[\phi|\mathscr{A}_+]$.

In [38] it was shown that a large class of factor models for the FRC admit arbitrage and contain the extended Nelson-Siegel models as a particular case. Therefore, there are instances in which a model admitting arbitrage better explains the data than the closest arbitrage-free model to it.

In the frame of behavioral finance, if market frictions are overlooked, this is empirical evidence that there are instances where the market is acting more *irrational* than *rational* by admitting potential mis-pricing opportunities. These behavioral anomalies can be quantified by first investigating when there is a spread between the empirical factor model and its arbitrage-free regularization, and secondly asking if that spread is statistically significant in terms of the data.

Definition 4.6.3 (Irrational Overpricing (resp. Underpricing)) *Let T be a fixed maturity time and consider that either*

1. $\mathbb{A}_{\phi}[\phi|\mathscr{A}](t,T,\beta_t) > \phi(t,T,\beta_t)$,
2. $\phi(t,T,\beta_t) > \mathbb{A}_{\phi}[\phi|\mathscr{A}](t,T,\beta_t)$,

where $I(\cdot)$ is the indicator random variable.

In the case where (ia) holds, we say that T is irrationally under-priced. Similarly, if (ib) holds, then we say that T is irrationally over-priced. If either (ia) and (ib) do not hold, then we say that the market is in a rational state.

Since β_t is not directly observable, the irrational overpricing (resp. underpricing) will be modeled using the following observable process.

Definition 4.6.4 (Potential Overpricing (resp. Underpricing)) *Let $\hat{\beta}_t$ be the parameters of the factor model $\phi(t,T,\beta_t)$ for the FRC estimated using the Kalman smoother. Let $t_0 < \cdots < t_M$ be a sequence of times on which the vector of forward-rates $\{f(t,T_j)\}_{j=1}^{N_t}$ was observed for a zero-coupon Bond with price $B(t_\star,T_\star)$ were observed. Here $N_t \geq 1$. Fix the thresholds $\epsilon,\delta > 0$. Let T be a fixed maturity time and consider that*

1. *Potentially there is a price inconsistent with the no-arbitrage pricing theory, in the sense that either*
   (a) $\mathbb{A}_{\phi}[\phi|\mathscr{A}](t,T,\hat{\beta}_t) > \phi(t,T,\hat{\beta}_t) + \epsilon$,
   (b) $\phi(t,T,\hat{\beta}_t) > \mathbb{A}_{\phi}[\phi|\mathscr{A}](t,T,\hat{\beta}_t) + \epsilon$.

2. *The tail mass of the empirical error distribution for the arbitrage-free regularized model is less than that of the empirical factor model. That is*

$$(1+\delta) < \frac{\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_{t_i}} I\left(\left[err^N(t_i,T_j)-\overline{err}^N\right]^2 > \left[err^N(t_i,T)-\overline{err}^N\right]^2\right)}{\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_{t_i}} I\left(\left[err^A(t_i,T_j)-\overline{err}^A\right]^2 > \left[err^A(t_i,T)-\overline{err}^A\right]^2\right)}$$

$$err^A(t,T) \triangleq \mathbb{A}_\phi\left[\phi\middle|\mathscr{A}^+\right](t,\beta_t,T) - f(t,T); \quad err^N(t,T) \triangleq \phi(t,\beta_t,T) - f(t,T),$$

$$\overline{err}^A \triangleq \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_{t_i}} \frac{err^A(t_i,T_j)}{N_{t_i}}; \quad \overline{err}^N \triangleq \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_{t_i}} \frac{err^N(t_i,T_j)}{N_{t_i}},$$

where $I(\cdot)$ is the indicator random variable.

In the case where (ia) and (ii) hold, we say that T is potentially under-priced. Similarly, if (ib) and (ii) hold, then we say that T is potentially over-priced. If either (ii) or (i) fails to hold, we say that the market is in a potentially rational state.

As expected, when estimating the empirical factor model and its arbitrage-free regularization on a daily basis, the arbitrage-free regularization performs better, but its error distribution has a greater sample average than when these models are estimated on a bi-monthly basis. This is consistent with economic theory, which stipulates that mispricings exist, but are corrected very quickly by the market due to market efficiency.

The proportion of times that that a point of the FRC is either potentially over-priced, or potentially under-priced varies it may be large or small depending on the chosen threshold. We denote this proportion by $\hat{\pi}_\delta^\epsilon$. Its estimates are illustrated by the following Table 4.5, ϵ and δ are in basis points.

Thresholds: $\epsilon = 0.1, \delta = 0$

Maturity	$\hat{\pi}^{\epsilon}_{\delta}$	St. Dev. of Estimate	Lower-Bound	Upper-Bound
US1Y	0.429	0.080	0.350	0.509
US2Y	0.458	0.088	0.370	0.546
US3Y	0.484	0.080	*0.404*	0.565
US5Y	0.500	0.080	0.420	0.580
US7Y	0.404	0.081	0.323	0.486
US10Y	**0.547**	**0.080**	**0.467**	**0.627**
US20Y	*0.303*	0.090	0.216	0.395
US30Y	0.365	0.090	0.277	0.456

Thresholds: $\epsilon = 1, \delta = 0.1$

Maturity	$\hat{\pi}^{\epsilon}_{\delta}$	St. Dev. of Estimate	Lower-Bound	Upper-Bound
US1Y	0.225	0.067	0.160	0.294
US2Y	0.377	0.089	0.289	0.467
US3Y	0.303	0.074	0.231	0.379
US5Y	0.374	0.078	0.298	0.453
US7Y	0.316	0.080	0.238	0.397
US10Y	**0.410**	**0.079**	0.331	0.489
US20Y	*0.018*	0.031	0.000	0.053
US30Y	*0.307*	*0.090*	0.220	0.399

Thresholds: $\epsilon = 2, \delta = 0.8$

Maturity	$\hat{\pi}^{\epsilon}_{\delta}$	St. Dev. of Estimate	Lower-Bound	Upper-Bound
US1Y	0.193	0.064	0.133	0.259
US2Y	**0.346**	0.087	0.260	0.435
US3Y	0.256	0.070	0.188	0.329
US5Y	0.300	0.074	0.228	0.375
US7Y	0.262	0.076	0.189	0.339
US10Y	**0.327**	0.075	0.253	0.404
US20Y	*0.018*	**0.031**	0.000	0.053
US30Y	0.278	*0.087*	0.194	0.368

Table 4.5: Probability that a point is potentially mispriced.

As in Table 4.3, italicized maturities are least likely and exhibit greatest variance, while boldface maturities exhibit the lowest variance and highest estimated probability of being potentially priced.

Table 4.5 shows that the 2 and 10 year bonds have the highest probability of being potentially mispriced, according to the conservative thresholds $\epsilon = 2$, and $\delta = .8$. These mispricings do not take liquidity or market frictions into account and may not present arbitrage opportunities for these reasons. If in a pair of bonds with different maturities one is frequently potentially overpriced while the other is potentially underpriced and the two rapidly switch between these two states, relative to one another, then this pair of bonds can be used to form a statistical arbitrage strategy.

Define the states $(2 > 10)*$, $(2 < 10)*$, and $(2|10)*$ by

- **$(2 > 10)*$:**
  - $B(t, 2)$ is potentially rationally priced and $B(t, 10)$ is potentially underpriced,
  - $B(t, 2)$ is potentially overpriced and $B(t, 10)$ is potentially rationally priced,
  - $B(t, 2)$ is potentially overpriced and $B(t, 10)$ is potentially underpriced,
- **$(2 < 10)*$:**
  - $B(t, 10)$ is potentially rationally priced and $B(t, 2)$ is potentially underpriced,
  - $B(t, 10)$ is potentially overpriced and $B(t, 2)$ is potentially rationally priced,
  - $B(t, 10)$ is potentially overpriced and $B(t, 2)$ is potentially underpriced,
- **$(2|10)*$:** $(2 > 10)*$ and $(2 < 10)*$ are false.
- **$(2 > 10)$:**
  - $B(t, 2)$ is rationally priced and $B(t, 10)$ is irrationally underpriced,
  - $B(t, 2)$ is irrationally overpriced and $B(t, 10)$ is rationally priced,
  - $B(t, 2)$ is irrationally overpriced and $B(t, 10)$ is irrationally underpriced,
- **$(2 < 10)$:**
  - $B(t, 10)$ is irrationally rationally priced and $B(t, 2)$ is irrationally underpriced,
  - $B(t, 10)$ is irrationally overpriced and $B(t, 2)$ is irrationally rationally priced,
  - $B(t, 10)$ is irrationally overpriced and $B(t, 2)$ is irrationally underpriced,
- **$(2|10)$:** $(2 > 10)$ and $(2 < 10)$ are false.

The observable states $(2 > 10)*$, $(2 < 10)*$, and $(2|10)*$ are modeled as reflecting the hidden states $(2 > 10)$, $(2 < 10)$, and $(2|10)$. The transitions between the hidden states is assumed to follow a time-homogeneous Markov process. Likewise, the transition between the observable states is modeled as following a time-homogeneous Markov process.

The matrix whose entries describe the probabilities that any observable state is seen given that the hidden Markov process is in any one of the hidden states is called the emission probability matrix. The matrix describing the probability that the hidden Markov process transitions between any two of its states, as well as the emission probability matrix are estimated in Table 4.5. The transition probability matrix between the states $(2 > 10)$, $(2|10)$, and $(2 < 10)$ and the emission probability matrix, are estimated using the Baum-Welch formulation of the EM algorithm.

We initialize the estimate of the EM algorithm using a maximum likelihood estimate of the transition between the observed states. Our findings are recorded within the following tables.

(a) Transition Probabilities

	$(2<10)$	$(2\|10)$	$(2>10)$
$(2<10)$	0.968	0.015	0.017
$(2\|10)$	0	0.975	0.025
$(2>10)$	0.041	0.046	0.912

(b) Emission Probabilities

	$(2<10)^*$	$(2\|10)^*$	$(2>10)^*$
$(2<10)$	0.532	0.468	0
$(2\|10)$	0	0.993	0.007
$(2>10)$	0	0.264	0.736

Figure 4.1: Estimated Transition Probability Matrix for Pair of Maturities $(2,10)$

In Figure 4.1a, We denote by $(2<10)^*$ (resp. $(2|10)^*$, resp. $(2>10)^*$) the estimated analogue of $(2<10)$ (resp. $(2|10)$, resp. $(2>10)$) estimated using $\hat{\pi}^{\epsilon}_{\delta}$ as in Table 4.5. As anticipated, Figure 4.1a shows that the pair $B(t,2)$ and $B(t,10)$ infrequently transition between the states $(2<10)$, $(10<2)$ and $(2|10)$. Since transitions do indeed occur making a pairs trading strategy possible. The nearly diagonal nature of the transition probability matrix implies that round trips will take some time to complete.

The first and third entries of the emission probabilities matrix's middle column show that many the estimated times when the probability that a point is indeed irrationally mispriced, given that it is potentially mispriced can be significantly lower than the probability that it is mispriced. Therefore the probabilities of potential mispricings reported in Table 4.5 are much higher than the estimated probabilities that any of the US treasury bonds are indeed mispriced.

We discuss the strategy in more detail here before evaluating its performance.

Strategy 4.6.5 (Statistical Fixed-Income Arbitrage Strategy)**.**

1. **Identify High Arbitrage Potential Maturities:** At the current time t_1, identify maturities $\{T_1 < \cdots < T_n\}$ which have a high positive probability of being potentially mispriced,

2. **Identify Most Actively Fluctuating Pair:** Select two maturities T_a and T_b amongst $\{T_1 < \cdots < T_n\}$, maximizing

$$(T_a,T_b) \triangleq \max_{(a,b)} \sum_{\substack{i,j=1\\ i\neq j}}^{3} \left(p_{i,j}^{a,b}\right)^2,$$

   where $(p_{i,j}^{a,b})_{i,j}$ is the transition probability matrix between states $(T_a<T_b)$, $(T_a>T_b)$, and $(T_a|T_b)$, represented by $j=1,2,3$ respectively.

3. **Buy low sell high:** Go short $B(t_1,T_a)$ and go long on $B(t_1,T_b)$ by K units if in state $(T_a>T_b)$ at time t,

4. **Close positions:** For a future time point $t_2 > t_1$, go short $B(t_2, T_b)$ and go long on $B(t_2, T_a)$ by K units if in state $(T_a < T_b)$, and if $B(t_2, T_b) > B(t_1, T_b)$ as well as $B(t_2, T_a) > B(t_1, T_a)$.

Repeat steps 3 and 4 until desired capital is reached.

We implement strategy 4.6.5 across a 15 year time horizon ending on December 2014 on US bonds. As expected, transitions rarely occur in pairs and 2 round trips takes about 15 years. The performance of the strategy is benchmarked against two strategies, one which invests in the 2 year bond and reinvests immediately when the bond matures, and another which does the same for the 10 year bond. Each of the three implemented strategies is self-financing with initial portfolio value of 1m USD and short-selling is allowed with the constraint that the total portfolio value is non-negative.

Portfolio Metrics	Pairs	US2	US10	Best	Worst
Terminal Wealth	**2,019,419.000**	$1,794,979.000$	$2,006,301.000$	Pairs	US10
Min. Excess Wealth	**605,300.300**	0	0	Pairs	US2 & US10
Max. Excess Wealth	**1,019,419.000**	$794,979.000$	$1,006,301.000$	Pairs	US10
$ES_{0.5}$	10.894	14.731	**5.127**	US10	US2
$ES_{0.9}$	3.475	5.319	**1.670**	US10	US2
$ES_{0.95}$	3.128	4.774	**1.499**	US10	US2
$ES_{0.99}$	2.883	4.396	**1.380**	US10	US2
$ES_{0.999}$	3.497	4.318	**1.355**	US10	US2
Prop. Time Active	**0.558**	1	1	Pairs	US2 & US10

Table 4.6: Strategy Comparisons and Metrics

In Table 4.6, Maximum (resp. Minimum) Excess Wealth denote the most (resp. least) the portfolio were worth, minus the initial capital of 1m USD, during the trading period. ES denotes the historical expected shortfall, and Prop. Time Active is the proportion of the time window on which the portfolio is non-empty. Portfolios evaluated here as self-financing with initial capital of 1m USD.

From Table 4.6 we see that the pairs strategy spends the least time actively trading in the market, has the highest portfolio value and the lowest minimum portfolio value. The Pairs trading strategy seems to take on less risk than investment in the 10 Year bond but more risk than investment in the 2 Year bond, this is due to the short and long positions of both bonds partially of setting each-other. The expected shortfall does not capture the risk avoided by not participating in the market. As is illustrated in the last row of Table 4.6, the pairs trading strategy must have a lower market risk since the trader spends less time actively holding assets than with the other two a-priori seemingly passive strategies. Overall, our pairs trading strategy has a higher payout and relatively lower risk when benchmarked against these two low-risk passive bond investing strategies.

In summary, our low-risk pairs strategy based on the detection and classification of potential mispricings in the market provides an alternative to classical pairs strategies where the assets are required to be co-integrated (see [19, Chapter 11] for details). Instead, here, we require that the pair of assets' pricing states fluctuate often and above a preset threshold. This is the final application of arbitrage-free regularization presented in this chapter, and serves as a natural endpoint wherein our new tools were used to obtain a tangible and novel trading signal. We now close this chapter by taking the time to summarize our findings.

4.7 Summary

We have introduced an unsupervised learning algorithm, arbitrage-free regularization, which optimally removes arbitrage opportunities from factor models for a wide range of asset classes. Its definition resided on the introduction of a new class of semi-parametric models providing an alternative finite dimensional generalization of the FDR-HJM models other than to the GHJM framework. This new modeling framework have factor models at their core, but allows for their constant predictable deformation, a feature which allowed us to meet all of our modeling principles. Since flow models are a predictably deforming factor model, and are not statically chosen factor model such as FDR-HJM models, the no-arbitrage condition can be enforced with arbitrage-free regularization while the interpretability of the original factor model. Lastly, their re-usability is a result of the generality and flexibility provided by the functional F, allowing flow models and their arbitrage-free regularization to model many asset classes. Additionally, flow models naturally are able to incorporate path-dependent and non-Euclidean features adding to their flexibility and generality.

The arbitrage-free regularization operator, within the framework of flow models, allowed to give equivalent and extended formulations of many classical results from risk-neutral pricing theory such as NFLVR, market completeness, minimal martingale measure, and the risk-neutral pricing formula. Arbitrage-free regularization allowed for finding the risk-neutral price directly under the objective measure $\mathbb{P}$ by optimally deforming the factor model's structure instead of changing measure. This formulation was shown to extend the classical risk-neutral pricing theory to models which admit arbitrage. The Nelson-Siegel model was used as a familiar example on this theme and the Arbitrage-Free Nelson-Siegel was shown to be the result of a particularly formulation of the Nelson-Siegel model's arbitrage-free regularization.

Arbitrage-free regularization was used as an integral part of the arbitrage-free estimation meta-algorithm, which allowed any estimation procedure for the dynamic factors to simultaneously extract further arbitrage-free information from the market data and the empirical factor model. The arbitrage-free estimation analogues of the Kalman-smoother and MLE algorithms were shown to provide superior one-day ahead forecasts and in-sample bi-monthly estimates of forward-rate curves. These improvements were quantified in terms of SSE, standard-deviation of errors, AIC, and probability that the arbitrage-free regularization outperforms its empirical counterpart. The Nelson-Siegel, AFNS, and their extensions provided benchmarks to make concrete understanding of this performance gain. In nearly every case arbitrage-free regularization was found to outperform their empirical and AFNS counterparts. The predictive advantage of arbitrage-free regularization of the NS model over the empirical factor model was due to the flexible incorporation of the arbitrage-free evolution of the bond price. The out-performance of the AFNS model by the arbitrage-free regularized NS model was explained as the added flexibility the arbitrage-free estimated models provided since they admit many arbitrage-free deformations and not only a single spread deformation.

The mismatch between certain models admitting arbitrage in a model and their arbitrage-free regularization gave way to an mispricing-detection technique. This misprice detection methodology was used to construct a mispricing classification algorithm which in-turn formed the cornerstone of a pairs trading strategy relying on methods from hidden-Markov model theory. The disadvantage of our arbitrage-free regularization is that regularization with respect to more general classes of deformations that $\mathscr{A}^+, \mathscr{A}_+$, or $\mathscr{A}_{\mathbb{P}}$ may have less straightforward solutions and potentially requires the introduction of new estimation techniques. However, arbitrage-free regularization and arbitrage-free estimation with respect to either of the three aforementioned classes of deformations is solved and provides numerically encouraging results.

In closing, arbitrage-free regularization and estimation of flow models provides a novel extension of the classical foundation of risk-neutral pricing theory. This extension provides many theoretical and predictive advantages as well as new types of problems and applications to explore.

4.8 Appendix

This section contains a number of appendices which deal either with technical proofs or with related background material.

Definition 4.8.1 (Admissibility) *Let I be a finite subset of $\mathcal{U}$. A H_t strategy is said to be (α, I)-admissible, for $\alpha > 0$ if*

1. *H_t is predictable with respect to the filtration generated by $\{X_t(u_i)\}_{i=1}^N$,*
2. *$\lim\limits_{t\mapsto\infty} \sum_{i\in I} \int_0^t H_s(u_i) dX_s(u_i) \geq -\alpha$,*
3. *$H(u) = 0$ for $u \notin I$. .*

If H_t is (α, I)-admissible with respect to some $\alpha > 0$ and then H_t is said to be I-admissible.

Remark 4.8.2. The FTOAP, as formulated in [27] assumes that the market there are only a finite number of underlying assets in the market; which in the notation of this chapter means that $\mathcal{U}$ is finite. If $\mathcal{U}$ is finite, then the definition of admissibility in [44, Definition 2.1] is the same as Definition 4.8.1 if $\mathcal{U}$ is finite. However, Definition 4.8.1 does not require that $\mathcal{U}$ be finite, but instead it requires that assets from the market generated by a finite collection $\{X_t(u_i)\}_{i\in I}$ of pre-determined assets may be traded. As illustrated in [4], allowing for an infinite number of assets to be traded at once may become intractable through conventional tools.

Definition 4.8.3 (No Free Lunch with Vanishing Risk) *Let I be a finite subset of $\mathcal{U}$. A sequence $\{H_t^n\}_{n\in\mathbb{N}}$ of I-admissible strategies is said to be an Free Lunch with Vanishing risk if there exists and increasing sequence $\{\delta_n\}_{n\in\mathbb{N}}$ in $[0,1)$ converging to 1 and $\epsilon > 0$ such that*

1. $\mathbb{P}\left(\lim_{t\mapsto\infty} \sum_{i\in I} \int_0^t H_s(u_i)dX_s(u_i) > \delta_n - 1\right) = 1,$

2. $\mathbb{P}\left(\lim_{t\mapsto\infty} \sum_{i\in I} \int_0^t H_s(u_i)dX_s(u_i) > \epsilon\right) \geq \epsilon.$

If there are no free lunches with vanishing risk, then we say that $\{X_t(u_i)\}_{i\in I}$ satisfies NFLVR. If for every finite subset I of $\mathcal{U}$, $\{X_t(u_i)\}_{i\in I}$ satisfies NFLVR then we say that $\{X_t(u)\}_{u\in\mathcal{U}}$ satisfies NFLVR.

Remark 4.8.4. Since the FTOAP of [27] only requires a finite number of underlying assets in the market then Definition 4.8.3 is equivalent to [44, Definition 2.2 (iii)]. If $\mathcal{U}$ is infinite, then Definition 4.8.3 is exactly NFLVR with the restriction that the trader first selects a sub-market generated by the finite collection of assets $\{X_t(u_i)\}_{i\in I}$ and subsequently only trades using that sub-market. This is realistic since, for example, portfolios of call options are finite even though any number of strikes and maturity times may be mathematically modeled or since bond portfolios are finite but any number of maturity times may be mathematically modeled.

Therefore once the set I is selected the definition of NFLVR in Definition 4.8.3 becomes equivalent to [44, Definition 2.2 (iii)]. Therefore by the FTOAP, NFLVR holds for the market generated by $\{X_t(u_i)\}_{i\in I}$ if and only if there exists an ELMM $\mathbb{Q} \sim \mathbb{P}$ for which $\{X_t(u_i)\}_{i\in I}$ are $\mathbb{Q}$-local martingales. However, the FTOAP does not make claims about $\{X_t(u)\}_{u\in\mathcal{U}}$ if $\mathcal{U}$ is not finite.

Therefore the requirement that $\mathbb{Q}$ is an ELMM for each $\{X_t(u)\}_{u\in\mathcal{U}}$ is a strictly stronger claim. This greater restrictiveness can be understood as the reason that ELMMs for the large class of factor models studied in [38] fail to exist. This restriction may be relaxed by allowing the measure μ in Definition 4.4.1 be the counting measure supported on a finite subset of $\mathcal{U}$.

The focus of this Appendix is to show that, on any suitable constraint set, any continuous semi-martingale has a geometric analogue which does not leave the constraint set in a finite amount of time $\mathbb{P}$-a.s. This geometric process's tangential movements are precisely those of the original process. We first take a moment to review the stochastic differential geometry of [35] is first taken.

A Primer on Stochastic Riemannian Geometry

To every tangent space to a point on $\mathscr{M}$, there can be attributed a collection of orthogonal basses. Gluing each of these collection of basses to every point on $\mathscr{M}$ defines a manifold $\mathscr{O}(\mathscr{M})$ called the orthogonal frame bundle on $\mathscr{M}$. Riemannian geometry can be formalized as the study of movement on $\mathscr{M}$ which is describable by smooth transitions of these orthogonal basses on $\mathscr{O}(\mathscr{M})$. More formally once a basis is chosen at the path's starting point, it can be shown that any smooth path on $\mathscr{M}$ corresponds to a unique path in the frame bundle $\mathscr{O}(\mathscr{M})$ describing the infinitesimal evolutions of its orientations (see [91] for details on this discussion).

The same remains true for the path of any continuous semi-martingale β_t on $\mathscr{M}$, once an initial frame Ξ is chosen in the tangent space of β_0. The resulting lifted process, denoted by U_t^{β}, is called the *horizontal lift* of β_t to $\mathscr{O}(\mathscr{M})$ for the initial frame Ξ, and can be shown to itself be an $\mathscr{O}(\mathscr{M})$-semi-martingale. Since every tangent space of a point in $\mathscr{M}$ can be identified with $\mathbb{R}^d$, the movement is described by the horizontal lift U_t^{β} of β_t. This procedure traces out a process in $\mathbb{R}^d$ called the Riemannian *stochastic anti-development* of b_t in $\mathbb{R}^d$. This procedure can be inverted and it can be used to trace out a continuous semi-martingale on $\mathscr{M}$ beginning with a semi-martingale on $\mathbb{R}^d$; the inverse procedure of the Riemannian *stochastic anti-development* is called the Riemannian *stochastic development* in $\mathscr{M}$ of a continuous semi-martingale. For more detail, the reader is directed to [61] or [97]. In this chapter, the Riemannian metric g_t is allowed to smoothly vary in time. All the analogous time-dependent results remain true in this context, a detailed treatment of which is found in [51].

Proof of Remark 4.3.10

A central difficulty when working intrinsically on $\mathscr{M}$ is that the process can leave $\mathscr{M}$ in finite time. We construct a Riemannian metric which extends the usual Euclidean metric to $\mathscr{M}$ which ensures that any stochastic process does not leave the set $\mathscr{M}$ in finite time. We would like to note that the work of [59] on the Riemannian metric of the SABR volatility model uses some of the same tools we employ in our framework.

Lemma 4.8.5. Let $\mathscr{M}$ be an d-dimensional sub-manifold of $\mathbb{R}^D$ such that there exists natural numbers $0 \leq d_1, d_2 \leq d$ and there is a C^2-diffeomorphism Φ from $\mathscr{M}$ to the product manifold:

$$\begin{aligned} & A^{d_1} \times B \times \mathbb{R}^{d-d_1-d_2} \\ A \triangleq & \{(x,y) \in \mathbb{R}^2 : y > 0\} \\ B \triangleq & \{x = (x_1, \dots, x_{d_2-1}, 0) \in \mathbb{R}^{d_2} : \|x\| < 1\} \\ E \triangleq & \mathbb{R}^{d-d_1-d_2}. \end{aligned} \tag{4.49}$$

There exists a Riemannian metrics $g^{c,\star}$ and g^c on $A^{d_1} \times B \times \mathbb{R}^{D-d_1-d_2}$ and $\mathscr{M}$, respectively such that

1. Both $\left(A^{d_1} \times B \times \mathbb{R}^{d-d_1-d_2}, g^{c,\star}\right)$ and $(\mathscr{M}, g^c)$ are geodesically and stochastically complete,

2. The diffeomorphism Φ is an isometric embedding .

Proof. Since A is the upper-half plane, the hyperbolic metric whose Riemannian metric tensor ds^A is defined by

$$(ds^A)^2(p,v) \triangleq \frac{d_E(p,v)}{\|p\|},$$

where d_E is the Euclidean metric tensor on A, p is a point in A, and v is a tangent vector in A, defines a Riemannian metric on A. In [50] it is proven that the hyperbolic metric is both geodesically and stochastically complete. Similarly the Riemannian metric tensor ds^B on B defined by

$$(ds^B)^2(p,v) \triangleq \frac{d_E(p,v)}{\|p\|^2 - 1},$$

is the Poincaré-disc model for hyperbolic space (see [71]), which is isometrically-isomorphic to A, and therefore is geodesically and stochastically complete. In [50], it is shown that E with the usual Euclidean metric tensor d_E is both geodesically as well as stochastically complete. Since the product of geodesically complete Riemannian manifolds is geodesically complete then $A \times B \times E$ is geodesically complete under the product Riemannian metric [63].

The Laplacian on a product manifold, such as $A \times B \times E$, can be written as the sum of the Laplacian Δ of its parts

$$\Delta = \Delta_{A_1} + \cdots + \Delta_{A_{d_1}} + \Delta_B + \Delta_E, \tag{4.50}$$

where $A_1 \times \dots A_{d_1}$ are d_1 distinct copies of making up A^{d_1}. Since $A_1, \dots, A_{d_1}, B$ and E where stochastically complete the equations

$$\Delta_i v = av; i \in \{A_1, \dots, A_{d_1}, B, E\}\, a > 0,$$

have precisely one solution ([49, page 74]). By the linear relationship in equation (4.50), it follows that there exists a unique solution to the equation

$$\begin{aligned} \Delta v &= \sum_{i=0}^{d_1} \Delta_{A_i} v_{A_i} + \Delta_B v_B + \Delta_E v_E \\ &= \sum_{i=0}^{d_1} a v_{A_i} + a v_B + a v_E. \end{aligned} \tag{4.51}$$

Therefore, $A \times B \times E$ is stochastically complete by [50, Theorem 1.1], where $v_{A_1}, \dots, v_{A_{d_1}}, v_B$ and v_E are the projections of v onto the subspaces A, B and E respectively. Therefore

$A^{d_1} \times B \times E$ is both geodesically and stochastically complete when equipped with the product Riemannian metric defined y $g^{c,\star} \triangleq \bigotimes_{i=1}^{d_1}(ds^A) \otimes ds^b \otimes ds^E$. Define the Riemannian metric g^c on $\mathscr{M}$ to be the pull-back of g^c, that is

$$g^c(x, y) \triangleq g^{c,\star}(d\Phi^{-1}(x), d\Phi^{-1}(y)),$$

where $d\Phi^{-1}$ is the pushforward of $g^{c,\star}$ along Φ (see [63] for details); makes Φ into an isometry. □

Example 4.8.6 (Extended Nelson-Siegel Loadings)**.** In equation (4.5), the factors of the Nelson-Siegel model could take on arbitrary values. Therefore the process β_t constrained to the manifold $\mathscr{M} = \mathbb{R}^d$ on which the factors are defined must be β_t itself. In this case the Riemannian metric g^c on $\mathscr{M}$ is the Euclidean metric and the isometry Ψ is the identity map.

Example 4.8.7 (Stochastic Volatility Surface Factors)**.** In equation (4.11), the factors β of the wavelet model for the stochastic variance surface were required to live on $(0, \infty)^{d^2} = B^{d^2}$. By Lemma 4.8.9, any continuous semi-martingale β_t on $\mathbb{R}^{d^2}$ may be constrained to $(0, \infty)^{d^2}$. In this case, the Riemannian metric g^c on $\mathscr{M}$ is non-Euclidean, but the isometry Φ is the identity map.

Example 4.8.8 (Mixed Densities for Stochastically Uncertain Models)**.** The set

$$\mathscr{M} \triangleq Ball(\bar{x}; \frac{\pi}{4});\ \bar{x} \triangleq \frac{1}{\sqrt{3}}(1, 1, 1),$$

is the open ball contained within the set $\{x \in \mathbb{R}^{d+1} : \|x\| = 1 \text{ and } x_i > 0\}$ with maximal measure. The Riemannian exponential map

$$\begin{aligned} Exp_{\bar{x}} :& Ball\left(\bar{x}; \frac{\pi}{4}\right) \to \mathbb{R}^d, \\ Exp_{\bar{x}}(\vec{v}) \triangleq& \cos(\|\vec{v}\|)\vec{\bar{x}} + \sin(\|\vec{v}\|)\frac{\vec{v}}{\|\vec{v}\|} \end{aligned} \tag{4.52}$$

is a radial isometry (see [63] for details) therefore it defines an isometric isomorphism from the set $\mathscr{M}$ onto the open set $B_0 \triangleq \left\{x \in \mathbb{R}^d : \|x\| < \frac{\pi}{4}\right\}$. The map $z \mapsto \frac{z+i}{z-i}$ is an isometry between the hyperbolic metric equipped with the hyperbolic metric and the upper half plane $\mathbb{C}_+ \triangleq \{z \in \mathbb{C} : Im(z) > 0\}$. Since it will be simpler to work in the upper-half plane directly, the map

$$\begin{aligned} &\Phi : \mathbb{C}_+ \to \mathscr{M} \\ &\Phi(z) = \cos(\|\frac{z \vec{-} i}{z + i}\|)\vec{\bar{x}} + \sin(\|\frac{z \vec{-} i}{z + i}\|)\frac{\frac{z\vec{-}i}{z+i}}{\|\frac{z\vec{-}i}{z+i}\|}; \\ &\bar{x} \triangleq \frac{1}{\sqrt{2}}(1, 1), \end{aligned}$$

defines a diffeomorphism between the hyperbolic upper-half plane and the points on the largest intrinsic ball on the sphere of radius 1 contained entirely within the first orthant of $\mathbb{R}^3$. By Lemma 4.8.5, the $\mathbb{C}_+$ is of the form $A^2 \times B^0 \times \mathbb{R}^{2-0-2} = B$, therefore the Riemannian metric g^c on $\mathscr{M}$ is the defined by the pull-back across Φ.

We are now in a position to formalize and prove Remark 4.3.10.

Lemma 4.8.9 (Canonical Constrained Process)**.** Let $\mathscr{M}$ be a d-dimensional sub-manifold of $\mathbb{R}^D$ and suppose there exists non-negative integers d_1 and d_2, such that there exists a C^2-diffeomorphism from $\Phi : \mathscr{M} \to A^{d_1} \times B \times \mathbb{R}^{D-d_1-d_2}$, where A, B are as in Lemma 4.8.5. Let β_t be an $\mathbb{R}^d$-valued continuous semi-martingale.

Then the process β_t^c defined by

$$\beta_t^c \triangleq \Phi\left(\beta_t^{g^c}\right),$$

is a g^c-semi-martingale; it does not leave the manifold $\mathscr{M}$ in a finite amount of time $\mathbb{P}$-a.s. Moreover, if $d_1 = d_2 = 0$ then $\beta_t = \beta_t^c$.

Proof. In [61] it is shown that the anti-development of a continuous semi-martingale onto a Riemannian manifold $(\mathscr{M}, g)$ is itself a g-semi-martingale. Lemma 4.8.5 shows that $(\mathscr{M}, g^c)$ is stochastically complete therefore any g^c-semi-martingale does not leave $\mathscr{M}$ in a finite amount of time; in particular this is the case for β_t^c.

If $\mathbb{R}^d = \mathscr{M}$, then the left action by U_t is the identity map $x \mapsto Ix$. Therefore its derivative is the map I and the Hessian $Hess^{g_t}(U_t e^i, U_t e^j) = 0$. Hence,

$$d\beta_t^c = \sum_{i=1}^{D} e_i \cdot d\beta_t + 0 = d\beta_t.$$

□

Example 4.8.10. Let β_t be an $\mathbb{R}^2$-valued diffusion solving the SDE

$$d\beta_t = \mu(t, \beta_t)dt + \sigma(t, \beta_t)dW_t.$$

Let Φ be the diffeomorphism described in Example 4.8.8. In local coordinates on A, the canonical constrained process β_t^{Φ} is given by

$$\begin{aligned} d\beta_t^c =& \mu(t, \beta_t^c)dt - \Gamma dt + \sigma(t, \beta_t^c)dW_t \\ \Gamma \triangleq& \begin{pmatrix} \frac{1}{(\beta_t^c)^2} & \frac{-1}{(\beta_t^c)^2} \\ \frac{-1}{(\beta_t^c)^2} & \frac{-1}{(\beta_t^c)^2} \end{pmatrix} \begin{pmatrix} (\sigma(t, \beta_t^c)^1) \\ (\sigma(t, \beta_t^c)^2) \end{pmatrix}, \end{aligned}$$

see [35, Section 9.2.4] for local descriptions of g-martingales, in terms of Christoffel symbols; here the Christoffel symbols used are those of the hyperbolic space.

Proofs whose length may detract from the flow of the chapter are recorded in this appendix. The proofs are divided into two groups, proofs of No-Arbitrage related results and proofs related to the extensions of classical risk-neutral pricing theory.

Proofs of No-Arbitrage Theorems

Proof of Theorem 4.3.11. For legibility, abbreviate $\phi(t,u,\beta_t)$ by $f(t,u)$ and let

$$f(t,u) = f(0,u) + \int_0^t \alpha(s,u,f(t,u))ds + \int_0^t \gamma(s,u,f(t,u))dW_s.$$

If $F(t,f(t,u),[f(t,u)])$ is a realization of the flow model then, by [45, Theorem 3.1], it follows that

$$\begin{aligned}F(t,f_t,[f]_t) =& F(0,f_0,[f]_0) + \int_0^t \mathscr{D}_s F(s,f_s,[f]_s)ds + \int_0^t \mathscr{V}_s F(s,f_s,[f]_s)\alpha(s,u,f)ds \\ &+ \int_0^t \frac{1}{2} tr[{}^t\mathscr{V}_s^2 F(s,f_s,[f]_s)]\gamma(s,u,f)^2 ds \\ &+ \int_0^t \frac{1}{2} tr[{}^t\mathscr{V}_s^2 F(s,f_s,[f]_s)]\gamma(s,u,f)dW_s.\end{aligned} \tag{4.53}$$

By the (constructive) Martingale Representation Theorem (see [45, Theorem 3.2]), it follows that $F(t,f_t,[f]_t)$ is a local-martingale if and only if

$$\begin{aligned}\int_0^t \mathscr{D}_s F(s,f_s,[f]_s)ds + \int_0^t \mathscr{V}_s F(s,f_s,[f]_s)\alpha(s,u,f)ds& \\ + \int_0^t \frac{1}{2} tr[{}^t\mathscr{V}_s^2 F(s,f_s,[f]_s)]\gamma(s,u,f)^2 ds = 0.&\end{aligned} \tag{4.54}$$

Since (F,ϕ,β_t,g_t) is a flow model, then, by an Ito's formula for β_t on the Riemannian manifold $(\mathscr{M},g)$, it follows that

$$\begin{aligned}\alpha(t,u,f)dt =& \frac{\partial\phi}{\partial t}(t,u,\beta_t) + \sum_{i=0}^d (\xi_t e_i)\phi(t,u,\beta_t)\mu(t,\beta_t)dt \\ &+ \frac{1}{2}\sum_{i,j=1}^d Hess^{g_t}\left(\phi(t,u)\right)(\beta_t e_i, \beta_t e_j)d[\beta_t^i,\beta_t^j]_t \\ \gamma(t,u,f)dW_t =& \sum_{i=0}^d (\xi_t e_i)\phi(t,u,\beta_t)\sigma(t,\beta_t)dW_t.\end{aligned} \tag{4.55}$$

Moreover, the Martingale Representation Theorem guarantees that α and β are unique up to indistinguishability. By Lemma 4.8.5 we have a Riemannian metric on $\mathscr{M}$ and from the Itô formula of [51, Corollary 3.6], we obtain

$$\begin{aligned}\int_0^t \mathscr{D}_s F(s, f_s, [f]_s) ds + &\int_0^t \mathscr{V}_s F(s, f_s, [f]_s) \left[\frac{\partial \phi}{\partial t}(t, u, \beta_t) + \sum_{i=0}^{d} (\xi_t e_i) \phi(t, u, \beta_t) \mu(t, \beta_t) dt \right. \\ &\left. + \frac{1}{2} \sum_{i,j=1}^{d} Hess^{g_t} \left(\phi(t, u)\right) (\beta_t e_i, \beta_t e_j) d[\beta_t^i, \beta_t^j]_t \right] ds \\ &+ \int_0^t \left(\frac{1}{2} tr[{}^t\mathscr{V}_s^2 F(s, f_s, [f]_s)] \left[\sum_{i=0}^{d} (\xi_t e_i) \phi(t, u, \beta_t) \sigma(t, \beta_t) \right]^2 ds \right) = 0.\end{aligned} \tag{4.56}$$

Since the process $F(t, \phi(t, u, \beta_t))$ is a local-martingale if and only if its drift term is indistinguishable from 0, it follows that $F(t, \phi(t, u, \beta_t))$ is a local-martingale if and only if the functional SPDE of equation (4.56) holds. □

Proof of Corollary 4.3.12. If we take $F(t, x, a)$ in Theorem 4.3.11 to be

$$F(t, f(t, u), [f(t, u)]) \triangleq \exp\left(- \int_t^T f(t, u) du \right),$$

then F is a C^∞ function of only the current state $f(t, \tau)$ and not of the entire path $t \mapsto f(t, u)$. Therefore, [45, Example 3.1] implies that

$$\begin{aligned}\mathscr{D}_t F = &(\partial_t \exp(x))|_{x = -\int_t^T f(t,u) du} = 0, \\ \mathscr{V}_t F = &(\partial_x \exp(x))|_{x = -\int_t^T f(t,u) du} = \exp\left(- \int_t^T f(t, u) du \right).\end{aligned} \tag{4.57}$$

Define the process $\phi(t, u, \beta_t) \triangleq - \int_t^T \varphi(t, u, \beta_t) du$ and substitute equations (4.57) into equation (4.17) to obtain

$$\begin{aligned}0 + \exp\left(- \int_t^T f(t, u) du \right) &\left[\int_0^t \left[\frac{\partial \phi}{\partial t}(t, u, \beta_t) + \sum_{i=0}^{d} (\xi_t e_i) \phi(t, u, \beta_t) \mu(t, \beta_t) dt \right. \right. \\ &\left. + \frac{1}{2} \sum_{i,j=1}^{d} Hess^{g_t} \left(\phi(t, u)\right) (\beta_t e_i, \beta_t e_j) d[\beta_t^i, \beta_t^j]_t \right] ds \\ &\left. + \int_0^t \left(\frac{1}{2} \left[\sum_{i=0}^{d} (\xi_t e_i) \phi(t, u, \beta_t) \sigma(t, \beta_t) \right]^2 ds \right) \right] = 0.\end{aligned} \tag{4.58}$$

Since $\exp\left(-\int_t^T f(t,u)du\right) > 0$, we have

$$\begin{aligned}
0 =& \int_0^t \left[\frac{\partial \phi}{\partial t}(t,u,\beta_t) + \sum_{i=0}^d (\xi_t e_i)\phi(t,u,\beta_t)\mu(t,\beta_t)dt\right.\\
&\left. + \frac{1}{2}\sum_{i,j=1}^d Hess^{g_t}\left(\phi(t,u)\right)(\beta_t e_i, \beta_t e_j)d[\beta_t^i,\beta_t^j]_t\right] ds\\
&+ \int_0^t \left(\frac{1}{2}\left[\sum_{i=0}^d (\xi_t e_i)\phi(t,u,\beta_t)\sigma(t,\beta_t)\right]^2 ds\right)\\
=& \int_0^t \left[\frac{\partial \phi}{\partial t}(t,u,\beta_t) + \sum_{i=0}^d (\xi_t e_i)\phi(t,u,\beta_t)\mu(t,\beta_t)dt\right.\\
&\left. + \frac{1}{2}\sum_{i,j=1}^d Hess^{g_t}\left(\phi(t,u)\right)(\beta_t e_i, \beta_t e_j)d[\beta_t^i,\beta_t^j]_t\right] ds\\
&+ \int_0^t \left(\frac{1}{2}\sum_{i,j=1}^d (\xi_t e_i)\phi(t,u,\beta_t)\sigma(t,\beta_t)(\xi_t e_j)\phi(t,u,\beta_t)\sigma(t,\beta_t)ds\right).
\end{aligned} \tag{4.59}$$

In local coordinates, equation (4.59) takes the form

$$\begin{aligned}
&\int_0^t \left[\frac{\partial \phi}{\partial t}(t,T,\beta_t) + \sum_{i=0}^d (\xi_t e_i)\phi(t,T,\beta_t)\mu(t,\beta_t)dt\right.\\
&+ \frac{1}{2}\sum_{i,j=1}^d \left(\frac{\partial^2 \phi}{\partial \beta_i \beta_j}(t,T,\beta_t) - \sum_{k=1}^d \Gamma_{ij}^k(t)\frac{\partial \phi}{\partial \beta_k}(t,T,\beta_t)\right)\sigma_i(t,\beta_t)\sigma_j(t,\beta_t)\\
&\left. + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial \phi}{\partial x_i}(t,T,\beta_t)\frac{\partial \phi}{\partial x_j}(t,T,\beta_t)\sigma_i(t,\beta_t)\sigma_j(t,\beta_t)\right] ds,
\end{aligned} \tag{4.60}$$

which established equation (4.18). □

Proof of Proposition 4.3.13. Let $F(t,\cdot)$ be the solution operator $\Sigma_t(\cdot|(T,K))$ defined by equation (4.12). Theorem 4.3.11 implies

$$\begin{aligned}
&\int_0^t \mathscr{V}\Sigma_t(\varphi_t^{(T,K)}|(T,K))\left[\frac{\partial \varphi}{\partial x}(t,\beta_t,\tau,K) + \sum_{i=0}^d (\xi_t e_i)\varphi(t,\beta_t,\tau,K)\mu(t,\beta_t,)dt\right.\\
&\left. + \frac{1}{2}\sum_{i,j=1}^d Hess^{g_t}\left(\phi(t,\beta_t,\tau,K)\right)(\beta_t e_i,\beta_t e_j)d[\beta_t^i,\beta_t^j]_t\right] ds\\
&+ \int_0^t \left(\frac{1}{2}\mathscr{V}^2\Sigma_t(\varphi_t^{(T,K)}|(T,K))\left[\sum_{i=0}^d (\xi_t e_i)\varphi(t,\beta_t,\tau,K)\sigma(t,\beta_t)\right]^2 ds\right) = 0.
\end{aligned} \tag{4.61}$$

Since $\Sigma_t(\varphi_t^{(T,K)}|(T,K))$ takes its input to the solution of the parabolic PDE with Borel-measurable initial condition given by the function $(S_T-K)_+$, then the Feynman-Kac formula implies that

$$\begin{aligned}\Sigma_t(\varphi_t^{(T,K)}|(T,K)) =&\mathbb{E}_{\mathbb{Q}}\left[\left(\tilde{S}_T-K\right)_+ \mid \sigma(S_t)\right]\\ d\tilde{S}_t =& \frac{K\nu(t,\tau,K)}{\sqrt{2}}.\end{aligned} \tag{4.62}$$

As discussed in [62], the call-option price verifies [62, Assumptions 4.7 and 4.8]; therefore by [62, Corollary 4.14 and Theorem 4.17] the first and second vertical derivatives of $\mathbb{E}_{\mathbb{Q}}\left[\left(\tilde{S}_T-K\right)_+ \mid \sigma(S_t)\right]$ are equal to the quantities described in equation (4.20). Hence, substituting $\mathscr{V}\mathbb{E}_{\mathbb{Q}}\left[\left(\tilde{S}_T-K\right)_+ \mid \sigma(S_t)\right]$ and $\mathscr{V}^2\mathbb{E}_{\mathbb{Q}}\left[\left(\tilde{S}_T-K\right)_+ \mid \sigma(S_t)\right]$ into equation (4.61) and making the change of variables $x \triangleq log(K)$ and $\tau \triangleq T-t$, yields equation (4.19). □

Proofs of Risk-Neutral Pricing Theory Results

Proof. Proof of Theorem 4.5.4

1. Assume that $\mathbb{A}_{\phi,\mu}^{H,\hat{AF}}[\phi|\cdot]$ exists. Therefore there exists a minimizer of the equation (4.33) in $\mathscr{A}_{\mathbb{P}}$. Hence there exists a family of measures $\{\hat{\mathbb{P}}_u\}$ such that for μ-a.e. u in $\mathcal{U}$, $X_t(u)$ is a local-martingale. Therefore, for μ-a.e. u in $\mathcal{U}$, the Fundamental Theorem implies that $X_t(u)$ satisfies NFLVR.

2. Assume now that for μ-a.e. u in $\mathcal{U}$, NFLVR holds if and only if, for μ-a.e. u in $\mathcal{U}$, there exists equivalent martingale measures $\hat{\mathbb{P}}_u$ to $\mathbb{P}$ and $X_t(u)$ admits a strict martingale density.

   Since μ, σ and F are deterministic, then the functional Ito formula of [45, Theorem 6.2.1] implies that the drift and volatility of $X_t(u)$ are deterministic for every $u \in \mathcal{U}$. Therefore, the definition of the mean-variance trade-off process described in [88, Equations 1.1-1.3], which we will denote by $\hat{K}_t^{X_t(u)}$, must be deterministic for every $u \in \mathcal{U}$. Hence, for μ-a.e. u in $\mathcal{U}$ the conditions [88, Theorem 7] are met, thus the minimal martingale measure $\hat{\mathbb{P}}_u$ is the unique solution to the problem

$$\underset{Q \in ELMM(\mathbb{P}|X_t(u))}{\operatorname{arginf}} \quad H(\mathbb{Q}\|\mathbb{P}), \tag{4.63}$$

   where $ELMM(\mathbb{P}|X_t(u))$ is the collection of equivalent local-martingale measures for $X_t(u)$ to $\mathbb{P}$ with $\mathbb{P}$-square integrable density processes. Since $\int_0^T \int_{u\in\mathcal{U}} \cdot^2 \mu(du)dt$

is monotone and convex, then equation (4.63) implies that its unique minimizer of

$$\underset{\{\hat{\mathbb{P}}_u\}_{u\in\mathcal{U}}\in\mathscr{E}(\mathbb{P})}{\operatorname{arginf}} \int_0^T \int_{u\in\mathcal{U}} H(\mathbb{Q}_u\|\mathbb{P})^2\mu(du)dt \tag{4.64}$$

must be the family of measures $\{\hat{\mathbb{P}}_u\}_{u\in\mathcal{U}}$; here $\mathscr{E}(\mathbb{P})$ is the collection of all families of measures $\{\mathbb{Q}_u\}_{u\in\mathcal{U}}$ such that for μ-a.e. u in $\mathcal{U}$, $\mathbb{Q}_u$ is an ELMM for $X_t(u)$. Since the $\mathbb{Q}_u$-dynamics of β_t are $Z_t^{\mathbb{Q}_u}\beta_t$, then by theorem 4.3.11, the condition that $X_t(u)$ is a $\mathbb{Q}_u$-local martingale is equivalent to

$$0 = \Lambda\left(\phi(t, Z^{\mathbb{Q}_u}\beta_t, u)\right) = \Lambda\left(\Phi_t^{\mathscr{Q}}(\phi)(t,\beta_t,u)\right),$$

which is in turn equivalent to $\hat{AF}\left(\Phi_t^Q\right)$ being finite and thus taking value 0. Therefore, equation (4.63) is equivalent to

$$\underset{\Phi_t^Q\in\mathscr{A}_{\mathbb{P}}}{\operatorname{arginf}} \int_0^T \int_{u\in\mathcal{U}} H(\mathbb{Q}\|\mathbb{P})^2\mu(du)dt + \hat{AF}\left(\Phi_t^Q\right). \tag{4.65}$$

Since H is deterministic and constant in time, $H(\mathbb{Q}\|\mathbb{P})$ achieves its minimum only if $\mathbb{E}\left[\int_0^T H(\mathbb{Q}|\mathbb{P})dt\right]$ achieves its minimum. Therefore, equation (4.65) is equivalent to minimizing Equation (4.33).

By construction, $\phi(t, Z_t^{\hat{\mathbb{P}}_u}\beta_t, u)$ is a local martingale, for every u. Therefore, whenever $X_t^{\mathscr{A}_{\mathbb{P}}}(u)$ exists the measure $\mathbb{P}$ is a LMM for it. □

Following [45], denote the class of boundedness-preserving functionals of a path by $\mathcal{B}$. Denote the horizontal and vertical derivative operators of the path by $\mathbb{D}([0,T];\mathbb{R}^d)$ and $\mathscr{V}$ respectively (see Definitions 2.6 and 2.8 of [45]). Write $\mathbb{C}^{1,2}$ for the class of once continuously horizontally and twice continuously vertically differentiable functionals of a path. Finally, $\mathbb{F}_l^\infty$ denotes the class of all left-continuous functionals of the path by.

Regularity Condition 4.8.11 (Encoding Functional Regularity)

1. *F is predictable in its second argument,*
2. *$\mathscr{V}F$, $\mathscr{V}^2F$, $\mathscr{D}F$ are all in $\mathbb{B}$,*
3. *$F, \mathscr{V}F, \mathscr{V}^2F$ are all in $\mathbb{F}_l^\infty$, and*
4. *$\mathscr{V}F$ is horizontally Lipschitz ([45]).*

Regularity Condition 4.8.12 (Factor Model Regularity)

1. $\mathbb{E}\left[\int_{u\in\mathcal{U}}\int_0^\infty \left(\phi(t,\beta_t,u)\right)^2 dt\mu(du)\right] < \infty.$

2. *For every* $(u, \beta) \in \mathcal{U} \times \mathscr{M}$ *the map* $t \mapsto \varphi(t, u, z)$ *is continuously differentiable,*
3. *There exists an* $s > (D + 1)/2$*, such that for every* $\beta \in \mathscr{M}$*, the map*

$$(t, u) \mapsto \phi(t, \beta, u) \tag{4.66}$$

   is an element of the Sobolev space $W_2^s(\mathbb{R} \times \mathcal{U})$*,*

4. *For every* $(t, u) \in [0, \infty) \times \mathcal{U}$*, the map* $\beta \mapsto \phi(t, \beta, u)$ *is a* C^2*-diffeomorphisms its image.*

Regularity Condition 4.8.13 (Geometric Regularity)

$$\mathbb{P} \otimes m\left(\{(\omega, t) \in \Omega \times [0, \infty) : \beta_t(\omega) \notin \mathscr{M}\}\right) = 0.$$

5. Conclusions and Future Work

This thesis consists of three main projects: the Non-Euclidean Upgrading meta-algorithm, a computational characterization and proof of the existence of non-Euclidean conditional expectation, and the introduction of arbitrage-free regularization and flow modeling framework. Our main contributions to each of these projects will be summarized here and related future research directions will also be discussed.

5.1 The NEU meta-algorithm

The contributions of this project are summarized below.

- The introduction of the universal reconfiguration property,
- The proof that the universal reconfiguration property implies the universal approximation property,
- The discovery of a class of algorithms other than neural networks having the universal approximation property,
- The introduction of the NEU meta-algorithm, which incorporated the universal reconfiguration property into any objective learning algorithm,
- The proof that the NEU version of any algorithm outperforms the original algorithm if the dataset is not already ideal,
- Numerical illustration of the performance gained by the NEU meta-algorithm as applied to regression analysis and principal component analysis,
- Applications to stock tracking and low-dimensional term-structure modeling,
- Parallels and contrasts to geodesic regression, principal geodesic analysis algorithms of [41, 42], and the geometric methods of [53].

Future work for the NEU meta-algorithm project will be to apply NEU to other classes of objective learning algorithms such as classification or clustering, use those to detect subtle trends in financial data, and the development of trading strategies exploiting those trends. The investigation of new properties the NEU algorithm may gain by considering reconfiguration maps other than rapidly decaying rotations presents another future research direction for this project.

5.2 Non-Euclidean Conditional Expectation and Efficient Portfolio Filtering

The contributions of the non-Euclidean conditional expectation and efficient portfolio filtering project are now summarized.

- Two rigorous formulations and characterizations of intrinsic conditional expectation were introduced,
- Both formulations were shown to exist and be equivalent on $L^2_{\mathbb{P}}(\mathscr{G}_t; \mathscr{M})$,
- Non-Euclidean filtering equations were derived as a consequence of the characterizations of non-Euclidean conditional expectation as dynamic conditional expectation,
- A Non-Euclidean Kalman filtering methodology was used to forecast efficient portfolios, these were benchmarked against competing algorithms from the electrical engineering and mathematical imagining literature, and found to outperform their Euclidean and non-Euclidean competitors,
- The spaces $\mathbb{L}^p_{\mathbb{P}}(\mathscr{G}_\cdot; \mathscr{M})$ were introduced and used to formulate Γ-convergence based proof techniques which are novel to applied probability theory.

A future line of work for the non-Euclidean conditional expectation project is to further localize the results and relax the assumption that non-Euclidean signal is defined on a Cartan-Hadamard manifold. This can either be accomplished by generalizing the results to arbitrary Riemannian manifolds or to general non-positive curvature spaces. Connections with the Cartan-Hadamard manifold structure of two-factor stochastic volatility models, discussed in [60], will also be explored.

5.3 Arbitrage-Free Regularization

The contributions of the arbitrage-free regularization project will now be summarized.

- An finite-dimensional, non-Euclidean, path-dependent alternative to the general HJM framework of [18] was introduced,
- Equivalent local martingale measures (ELMMs) for flow models were characterized,
- ELMMs for stochastic local volatility surfaces were characterized in terms of the Greeks,
- The arbitrage-free regularization methodology which deforms a model until the objective measure becomes its risk-neutral measure was introduced,

- Classical arbitrage-free pricing theory results such as the NFLVR, market completeness, and the minimal martingale measure were reformulated as the existence, uniqueness, and the solution to a particular arbitrage-free regularization problem,
- The generality of arbitrage-free regularization was used to extend classical arbitrage-pricing theory to models which admit arbitrage opportunities but are deformable into arbitrage-free models; this gave meaning to pricing under models such as the Nelson-Siegel model,
- The correction of a model admitting arbitrage into one that does not, such as the arbitrage-free Nelson-Siegel correction of the Nelson-Siegel model, was addressed in for general flow models and closed form solutions were found in the fixed-income setting,
- Implementations confirmed the predictive gain of arbitrage-free regularized models over their empirical counterparts,
- An arbitrage-detection methodology and pair trading strategy was introduced and shown to provide high payoffs at a low risk.

A future direction for the arbitrage-free regularization project is to further explore the numerical performance of arbitrage-free regularization for volatility surfaces and options. Another direction for the arbitrage-free regularization project is to apply the novel minimal deformation methodology to meet financial objectives other than risk-neutrality, such as aggregate risk minimization.

Bibliography

[1] 10 Major Companies Tied to the Apply Supply Chain (AAPL). https://www.investopedia.com/articles/investing/090315/10-major-companies-tied-apple-supply-chain.asp. Accessed: 2018-07-25.

[2] D. Ackerer and D. Filipovic. Option pricing with orthogonal polynomial expansions. *arXiv e-prints*, 2017.

[3] L. I. Allerhand and U. Shaked. Robust stability and stabilization of linear switched systems with dwell time. *IEEE Trans. Automat. Contr.*, 56(2):381–386, 2011.

[4] K. Back and S. R. Pliska. On the Fundamental Theorem of Asset Pricing with an infinite State Space. *J. Math. Econom.*, 20(1):1–18, 1991.

[5] W. Ballmann. *Lectures on spaces of nonpositive curvature*, volume 25 of *DMV Seminar*. Birkhäuser Verlag, Basel, 1995.

[6] F. Barbaresco. Innovative tools for radar signal processing based on cartans geometry of spd matrices and information geometry. In *Radar Conference*, pages 1–6. IEEE, 2008.

[7] M. J. Best. *Portfolio Optimization*. Chapman and Hall/CRC, first edition, 2010.

[8] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.*, 31(1):1–29, 2003.

[9] T. Björk and B. J. Christensen. Interest rate dynamics and consistent forward rate curves. *Math. Finance*, 9(4):323–348, 1999.

[10] T. Björk and R. M. Gaspar. Interest rate theory and geometry. *Port. Math.*, 67(3): 321–367, 2010.

[11] O. Blanchard and J. Simon. The long and large decline in us output volatility. *Brookings papers on economic activity*, 2001(1):135–164, 2001.

[12] S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM J. Matrix Anal. Appl.*, 31(3):1055–1070, 2009.

[13] A. Braides. A handbook of γ -convergence. *Handbook of Differential Equations: Stationary Partial Differential Equations*, 3:101–213, 2006.

[14] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[15] D. C. Brody and L. P. Hughston. Chaos and coherence: a new framework for interest–rate modelling. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 460, pages 85–110. The Royal Society, 2004.

[16] H. Bühler, L. Gonon, J. Teichmann, and B. Wood. Deep Hedging. *ArXiv e-prints*, 2018.

[17] R. Carmona and S. Nadtochiy. Local volatility dynamic models. *Finance Stoch.*, 13(1):1–48, 2009.

[18] R. A. Carmona. HJM: A unified approach to dynamic models for fixed income, credit and equity markets. In *Paris-Princeton Lectures on Mathematical Finance 2004*, volume 1919 of *Lecture Notes in Math.*, pages 1–50. Springer, Berlin, 2007.

[19] A. Cartea and S. Jaimungal. Algorithmic trading of co-integrated assets. *Int. J. Theor. Appl. Finance*, 19(6):1650038, 18, 2016.

[20] X. Chen, Z. Y. Dong, K. Meng, Y. Xu, K. P. Wong, and H. Ngan. Electricity price forecasting with extreme learning machine and bootstrapping. *IEEE Transactions on Power Systems*, 27(4):2055–2062, 2012.

[21] J. H. E. Christensen, F. X. Diebold, and G. D. Rudebusch. The affine arbitrage-free class of Nelson-Siegel term structure models. *J. Econometrics*, 164(1):4–20, 2011.

[22] S. N. Cohen and R. J. Elliott. *Stochastic calculus and applications*. Probability and its Applications. Springer, Cham, second edition, 2015.

[23] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

[24] E. De Giorgi. Sulla convergenza di alcune successioni d'integrali del tipo dell'area. *Rend. Mat.*, 8:277–294, 1975.

[25] A. De Gregorio and S. M. Iacus. Adaptive LASSO-type estimation for multivariate diffusion processes. *Econom. Theory*, 28(4):838–860, 2012.

[26] F. Delbaen and W. Schachermayer. A general version of the fundamental theorem of asset pricing. *Math. Ann.*, 300(3):463–520, 1994.

[27] F. Delbaen and W. Schachermayer. The fundamental theorem of asset pricing for unbounded stochastic processes. *Math. Ann.*, 312(2):215–250, 1998. ISSN 0025-5831.

[28] S. Devin, B. Hanzon, and T. Ribarits. A Finite-Dimensional HJM Model: How Important is Arbitrage-Free Evolution? *Int. J. Theor. Appl. Finance*, 13(8):1241–1263, 2010.

[29] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3): 189–228, 1996. With comments and a rejoinder by the authors.

[30] F. X. Diebold and C. Li. Forecasting the term structure of government bond yields. *J. Econometrics*, 130(2):337–364, 2006.

[31] T. E. Duncan. Some filtering results in Riemann manifolds. *Information and Control*, 35(3):182–195, 1977.

[32] B. Dupire. Pricing with a Smile. *Risk Magazine*, 7(1):18–20, 1994.

[33] B. Dupire. Functional Itô calculus. *Bloomberg Portfolio Research Paper No. 2009-04-FRONTIERS*, 2009. URL https://ssrn.com/abstract=1435551.

[34] S. Eckstein and M. Kupper. Computation of optimal transport and related hedging problems via penalization and neural networks. *ArXiv e-prints*, 2018.

[35] K. D. Elworthy. *Stochastic differential equations on manifolds*, volume 70 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge-New York, 1982.

[36] N. B. Erichson, P. Zeng, K. Manohar, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin. Sparse principal component analysis via variable projection. *arXiv preprint arXiv:1804.00341*, 2018.

[37] E. F. Fama and K. R. French. Industry costs of equity. *Journal of financial economics*, 43(2):153–193, 1997.

[38] D. Filipović. Exponential-polynomial families and the term structure of interest rates. *Bernoulli*, 6(6):1081–1107, 2000.

[39] D. Filipović. *Consistency Problems for Heath-Jarrow-Morton Interest Rate Models*, volume 1760 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2001.

[40] D. Filipović, S. Tappe, and J. Teichmann. Term structure models driven by Wiener processes and Poisson measures: existence and positivity. *SIAM J. Financial Math.*, 1(1):523–554, 2010.

[41] P. T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. volume 105, pages 171–185, 2013.

[42] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imaging*, 23(8):995–1005, 2004.

[43] P. T. Fletcher, S. M. Pizer, and S. C. Joshi. Shape variation of medial axis representations via principal geodesic analysis on symmetric spaces. In *Statistics and analysis of shapes*, Model. Simul. Sci. Eng. Technol., pages 29–59. Birkhäuser Boston, Boston, MA, 2006.

[44] C. Fontana, M. Jeanblanc, and S. Song. On arbitrages arising with honest times. *Finance Stoch.*, 18(3):515–543, 2014.

[45] D.-A. Fournie. *Functional Ito Calculus and Applications.* PhD thesis, 2010. Thesis (Ph.D.)–Columbia University.

[46] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Inform. Theory*, 54(11): 5130–5139, 2008.

[47] J. Gatheral. A parsimonious arbitrage-free implied volatility parameterization with application to the valuation of volatility derivatives. Presentation at Global Derivatives and Risk Management, Madrid, 2004.

[48] I. M. Gel'fand and G. E. Shilov. *Generalized functions. Vol. 2.* AMS Chelsea Publishing, Providence, RI, 2016.

[49] A. Grigor'yan. *Heat kernel and analysis on manifolds*, volume 47 of *AMS/IP Studies in Advanced Mathematics.* American Mathematical Society, Providence, RI; International Press, Boston, MA, 2009.

[50] A. Grigor'yan. Stochastic Completeness of Symmetric Markov Processes and Volume Growth. *Rend. Semin. Mat. Univ. Politec. Torino*, 71(2):227–237, 2013.

[51] H. Guo, R. Philipowski, and A. Thalmaier. Martingales on Manifolds with Time-Dependent Connection. *J. Theoret. Probab.*, 28(3):1038–1062, 2015.

[52] C. Han and F. C. Park. A geometric GARCH framework for covariance dynamics. *SSRN Preprints*, 2016.

[53] C. Han, F. C. Park, and J. Kang. A geometric treatment of time varying volatilities. *Rev Quant Finance Account*, 49:1121–1141, 2017.

[54] P. Harms, D. Stefanovits, J. Teichmann, and M. Wüthrich. Consistent Recalibration of Yield Curve Models. *Math. Finance*, 28(3):757–799, 2018.

[55] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, New York, second edition, 2009.

[56] A. Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.

[57] S. Hauberg, F. Lauze, and K. S. Pedersen. Unscented kalman filtering on riemannian manifolds. *J. Math. Imaging Vis.*, 46(1):103–120, 2013.

[58] D. Heath, R. Jarrow, and A. Morton. Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. *Econometrica*, pages 77–105, 1992.

[59] P. Henry-Labordère. A general asymptotic implied volatility for stochastic volatility models. *ArXiv e-prints*, 2005.

[60] P. Henry-Labordère. *Analysis, Geometry, and Modeling in Finance.* Chapman & Hall/CRC Financial Mathematics Series. CRC Press, Boca Raton, FL, 2009.

[61] E. P. Hsu. *Stochastic Analysis on Manifolds*, volume 38 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[62] S. Jazaerli and Y. F. Saporito. Functional Itô calculus, Path-Dependence and the Computation of Greeks. *Stochastic Process. Appl.*, 127(12):3997–4028, 2017.

[63] J. Jost. *Riemannian Geometry and Geometric Analysis.* Universitext. Springer, Heidelberg, sixth edition, 2011.

[64] X.-M. L. a. K. David Elworthy, Yves Le Jan. *The Geometry of Filtering.* Frontiers in Mathematics. Birkhuser Basel, 2010.

[65] O. Kallenberg. *Foundations of modern probability.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[66] J. Kallsen and P. Krühner. On a Heath-Jarrow-Morton Approach for Stock Options. *Finance Stoch.*, 19(3):583–615, 2015.

[67] J. L. Kelley. *General topology.* Springer-Verlag, New York-Berlin, 1975.

[68] A. W. Knapp. *Lie groups beyond an introduction*, volume 140. Springer, 2002.

[69] N. J. Korevaar and R. M. Schoen. Sobolev spaces and harmonic maps for metric space targets. *Comm. Anal. Geom.*, 1(3-4), 1993.

[70] A. Kratsios and C. B. Hyndman. The NEU meta-algorithm for geometric learning with applications in finance. 2018.

[71] J. M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2013.

[72] P. Liu. Approximation capabilities of multilayer feedforward regular fuzzy neural networks. *Appl. Math. J. Chinese Univ. Ser. B*, 16(1):45–57, 2001.

[73] S. MacLane. *Categories for the working mathematician.* Springer-Verlag, New York-Berlin, 1971. Graduate Texts in Mathematics, Vol. 5.

[74] H. M. Markowitz. Portfolio selection: Efficient diversification of investments. 1959.

[75] G. D. Maso. *An introduction to G-convergence.* Progress in Nonlinear Differential Equations and Their Applications. Birkhuser Boston, 1 edition, 1993.

[76] M. Moakher and M. Zéraï. The riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *J. Math. Imaging Vis.*, 40(2):171–187, 2011.

[77] J. Nash. The imbedding problem for Riemannian manifolds. *Ann. of Math. (2)*, 63: 20–63, 1956.

[78] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7(4):308–313, 1965.

[79] S. Ng and P. Caines. Nonlinear filtering in riemannian manifolds. *IFAC Proceedings Volumes*, 17(2):817–821, 1984.

[80] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[81] B. Pfaff. *Analysis of integrated and cointegrated time series with R.* Use R! Springer, New York, second edition, 2008.

[82] P. Piccione and D. V. Tausk. On the banach differential structure for sets of maps on non-compact domains. *Nonlinear analysis*, 46(2):245–265, 2001.

[83] F. Riesz. Elementarer beweis des egoroffschen satzes. *Monatshefte für Mathematik und Physik*, 35(1):243–248, 1928.

[84] R. T. Rockafellar. Integrals which are convex functionals. *Pacific J. Math.*, 24: 525–539, 1968.

[85] S. Said and J. H. Manton. On filtering with observation in a manifold: reduction to a classical filtering problem. *SIAM J. Control Optim.*, 51(1):767–783, 2013.

[86] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, 1990.

[87] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[88] M. Schweizer. On the minimal martingale measure and the Föllmer-Schweizer decomposition. *Stochastic Anal. Appl.*, 13(5):573–599, 1995.

[89] C. E. Shannon. A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[90] M. B. Shapiro, S. S.and Wilk. An analysis of variance test for normality: Complete samples. *Biometrika*, 52:591–611, 1965.

[91] R. W. Sharpe. *Differential Geometry*, volume 166 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[92] J. Sirignano and R. Cont. Universal features of price formation in financial markets: perspectives from deep learning. *ArXiv e-prints*, 2018.

[93] S. T. Smith. Covariance, subspace, and intrinsic cramer-rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005.

[94] H. Snoussi. Particle filtering on Riemannian manifolds. Application to covariance matrices tracking. In *Matrix information geometry*, pages 427–449. Springer, Heidelberg, 2013.

[95] E. H. Spanier. *Algebraic topology.* Springer-Verlag, New York, 1995.

[96] K. Stoller. The world's largest tech companies 2018: Apple, samsung take top spots again. *Frobes*, Jun 2018.

[97] D. W. Stroock. *An Introduction to the Analysis of Paths on a Riemannian Manifold*, volume 74 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2000.

[98] N. R. Swanson. Money and output viewed through a rolling window. *Journal of monetary Economics*, 41(3):455–474, 1998.

[99] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B. Stat. Methodol*, pages 267–288, 1996.

[100] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Dokl. Akad. Nauk. SSSR*, 151:501–504, 1963.

[101] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *Ann. Math. Statistics*, 11:147–162, 1940.

[102] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.*, 22(158):209–212, 1927.

[103] H. Xu, C. Caramanis, and S. Mannor. Robust regression and LASSO. *IEEE Trans. Inform. Theory*, 56(7):3561–3574, 2010.

[104] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

[105] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B. Stat. Methodol*, 67(2):301–320, 2005.

[106] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J. Comput. Graph. Statist.*, 15(2):265–286, 2006.