Kernel methods and Model predictive approaches for Learning and Control

Thèse N° 9524

Présentée le 2 septembre 2019

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire d'automatique 3
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Sanket Sanjay DIWALE

Acceptée sur proposition du jury

Prof. A. Ijspeert, président du jury
Prof. C. N. Jones, directeur de thèse
Prof. M. Zeilinger, rapporteuse
Prof. J. Suykens, rapporteur
Prof. G. Trecate, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Dedicated to my parents

Acknowledgments

Throughout the writing of this dissertation, I have benefited from the support and assistance from a great many people and thank them all for their guidance. In particular, I would like to thank Prof. Colin Jones for his support and several hours of productive discussions on the scientific and research content of this thesis. I would also like to thank Prof. Giancarlo Trecate, Prof. Phillippe Meullhaupt, Prof. Alireza Karimi, Prof. Dominique Bonvin and Prof. Victor Panaretos for their time and discussions, and for pointing me to various references for my research. I want to thank Prof. Salzmann for his help with various technical and non-technical topics over the years.

The presentation of this dissertation was also greatly enhanced by the reviews provided by Prof. Johan Suykens, Prof. Melanie Zeilinger and Prof. Giancarlo Trecate, and I thank them for their time in reviewing the dissertation and their helpful suggestions. I would also like to thank Dr. Arka Mallick for his help with proofreading some of the mathematical aspects of my work.

I thank Dr. Timm Faulwasser and Dr. Ioannis Lymperopoulos for their collaboration and guidance during the initial years of my work and for guiding me through the years. I also thank the various members of our lab, Milan, Georgios, Altug, Faran, Tomasz, Luca, Harsh, Pulkit, Sriniketh, Martand, Petr, Tafarel, Ivan, Ye Pu, Yingzhao, Emilio, Jean-Hubert and others, for creating such a vibrant learning atmosphere in the lab and for the various discussions held over the years. A special thanks to the masters and bachelors students who worked with me on several projects in the lab. Also my gratitude to Ruth, Francine, Eva and Nicole for helping me with administrative work at the lab.

Finally, I would like to thank my parents, family members, and friends for their support and encouragement over the years and for helping me get through this passionate period of my studies.

Lausanne, June 2019 S.D.

Abstract

Data-driven modeling and feedback control play a vital role in several application areas ranging from robotics, control theory, manufacturing to management of assets, financial portfolios and supply chains. Many such problems in one way or another are related to variational problems in optimal control and machine learning.

The following work first presents, a generalized representer theorem for solving such variational problems when closed, densely defined operators, like the differential operators, are involved. Furthermore, loss functionals on infinite dimensional Hilbert spaces are considered to allow for greater freedom in problem formulations. The statement of the theorem presents a necessary and sufficient condition for the existence of linear representer for optimal solutions of such problems. Finally, examples, applying the theorem to neural networks, stochastic regression, and sparsity-inducing regularization problems are presented.

The second part of the thesis deals with applications of variational optimization in control problems. Examples from optimal control and model predictive control are presented for applications in the domain of autonomous vehicles and airborne wind energy systems. First, a combination of manifold learning and model predictive control is presented for obstacle avoidance in autonomous driving. Manifold learning is presented as a means to describe boundaries of star-shaped sets for which a single inequality constraint is sufficient to check containment of a point in the set's interior. The approach presented, learns the largest star-shaped set within a circular range such that all obstacle points remain outside the set. The inequality condition for checking containment in such sets is incorporated into a multi-phase, free-end-time optimal control problem to plan trajectories and control inputs moving the vehicle from one point to another while remaining within a given collection of star-shaped sets. The multi-phase, free-end-time problem is adapted to a moving horizon form to give a model predictive path following controller that avoids obstacles by virtue of the manifold learning scheme. A real-time, dynamically updated manifold is learned using point cloud data from a lidar-like sensor on the vehicle to avoid any apriori unknown or moving obstacles. Convergence and recursive feasibility guarantees for the MPC scheme are provided under mild assumptions on the behavior of the obstacles and dynamics of the vehicle. An automated parking scenario in the presence of static and dynamic obstacles is demonstrated in simulation for the complete process of optimal trajectory planning and path following.

Next, a continuous time, path following model predictive control scheme is shown for an Airborne Wind Energy (AWE) system. Here stability and convergence guarantees are provided by combining the model predictive controller with terminal constraints inspired from a convergent vector field design problem. A formal stability proof relying on Lyapunov stability arguments is presented to show that for such a design of vector field terminal constraints the path following controller converges to a zero tracking error on the desired path.

The last part of the thesis deals with uncertainty in AWE systems arising due to uncertain wind conditions and unknown aerodynamic characteristics. Two different methods are presented for controlling the system in such environments. First, a Gaussian process based, data-driven approach is presented to optimize a closed-loop AWE system under unknown objective, constraint, and dynamic functions. Transient measurements are used to learn surrogate models for these quantities, and a safe online optimization scheme is presented to maximize the performance of an underlying controller while satisfying the safety constraints.

The second approach presents a nonlinear adaptive control scheme for path following in AWE systems in the presence of parametric uncertainties in the dynamics. Instead of using an approach that learns such model parameters from data, a direct adaptive scheme is considered, where obtaining accurate parameter estimates is not the goal of the parameter update scheme, but rather to cooperate with a nonlinear controller towards driving the path following error towards zero. Approximate feedback linearization of the AWE system is presented, and online parameter update laws are designed such that the path following error is driven into a small neighborhood of zero and the parameter estimates are guaranteed to be bounded for all time.

Finally, the appendix of the thesis is used to present some mathematical preliminaries.

Résumé

Le contrôle et la modélisation à partir des données jouent un rôle vital dans plusieurs domaines d'application comme par exemple la robotique, la théorie du contrôle, la fabrication à la gestion des actifs, les portefeuilles financiers et les chaînes d'approvisionnement. Beaucoup de ces problèmes, d'une manière ou d'une autre, sont liés à des problèmes variationnels du contrôle optimal et d'apprentissage.

Le travail suivant présente d'abord, une approche basée sur un théorème de représentants généralisés pour résoudre des problèmes variationnels fermés lorsque des opérateurs densément définis comme par exemples des opérateurs différentiels, sont impliqués. De plus, les fonctions de perte sur des espaces de Hilbert de dimensions infinies sont considérées pour permettre une plus grande liberté dans la formulation des problèmes. L'énoncé du théorème présente une condition nécessaire et suffisante pour l'existence d'un représentant linéaire pour des solutions optimales de ces problèmes. Finalement, des exemples d'application du théorème aux réseaux neuronaux, à la régression stochastique et aux problèmes de régularisation impliquant la rareté sont présentés.

La deuxième partie de la thèse porte sur les applications de l'optimisation variationnelle dans les problèmes de contrôle. Des exemples de commande optimale et de commande prédictive sont présentés pour des applications dans le domaine des véhicules autonomes et des systèmes d'énergie éolienne aéroportés. Tout d'abord, une combinaison d'apprentissage multiple et de commande prédictive est présentée pour éviter les obstacles en conduite autonome. L'apprentissage multiple est présenté comme un moyen de décrire les limites d'ensembles en forme d'étoile pour lesquels une seule contrainte d'inégalité est suffisante pour vérifier le confinement d'un point à l'intérieur de l'ensemble. L'approche présentée, apprend le plus grand ensemble en forme d'étoile à l'intérieur d'une plage circulaire de sorte que tous les points d'obstacle restent en dehors de l'ensemble. La condition d'inégalité pour la vérification du confinement dans de tels ensembles est incorporée dans un problème de contrôle optimal à phases multiples et à temps de fin libre pour planifier les trajectoires et les entrées de contrôle permettant de déplacer le véhicule d'un point à un autre tout en restant dans une collection donnée d'ensembles en étoile. Le problème à phases multiples de temps libre est adapté à une forme d'horizon mobile pour donner un chemin prédictif de modèle suivant un contrôleur qui évite les obstacles en vertu du schéma d'apprentissage multiple. Pour éviter tout obstacle inconnu ou en mouvement, une variété mis à jour dynamiquement en temps réel est apprise à l'aide d'un nuage de points provenant d'un capteur de type lidar monté sur le véhicule. Les garanties de convergence et de faisabilité récursive pour le schéma des PPM sont fournies sous des hypothèses sur le comportement des obstacles et la dynamique du véhicule. Le processus complet de planification de trajectoire optimale et de suivi de trajectoire est démontré en simulation pour un scénario de stationnement automatisé en présence d'obstacles statiques et dynamiques.

Ensuite, un schéma de commande prédictive de temps et de trajet continu est montré pour un système d'énergie éolienne aéroporté (Airborne Wind Energy - AWE). Les garanties de stabilité et

de convergence sont fournies en combinant le contrôleur prédictif avec des contraintes terminales inspirées d'un problème de conception d'un champ vectoriel convergent. Une preuve formelle de stabilité s'appuyant sur les arguments de stabilité de Lyapunov est présentée pour montrer que pour une telle conception des contraintes terminales de champ vectoriel, le contrôleur de suivi de trajectoire converge vers une erreur de suivi nulle sur la trajectoire souhaitée.

La dernière partie de la thèse traite de l'incertitude des systèmes AWE due à des conditions de vent incertaines et à des caractéristiques aérodynamiques inconnues. Deux méthodes différentes sont présentées pour contrôler le système dans de tels environnements. Tout d'abord, une approche gaussienne basée sur les données est présentée pour optimiser un système AWE en boucle fermée avec des contraintes, des objectifs et des fonctions dynamiques inconnus. Des mesures transitoires sont utilisées pour apprendre les modèles de substitution pour ces grandeurs, et un schéma d'optimisation en ligne est présenté pour maximiser la performance d'un contrôleur sous-jacent tout en satisfaisant les contraintes de sécurité.

La deuxième approche présente un schéma de contrôle adaptatif non linéaire pour le suivi de trajectoire dans les systèmes AWE en présence d'incertitudes paramétriques dans la dynamique. Au lieu d'utiliser une approche qui apprend de tels paramètres à partir des données, on envisage un schéma adaptatif direct, où l'obtention d'estimations précises des paramètres n'est pas le but du schéma de mise à jour des paramètres, mais plutôt de coopérer avec un contrôleur non-linéaire pour conduire l'erreur vers zéro. L'approximation de la linéarisation de rétroaction du système AWE est présentée, et les lois de mise à jour des paramètres sont conçues de manière à ce que l'erreur de suivi de trajectoire soit poussée dans le voisinage de zéro et que les estimations des paramètres soient garanties bornées dans le temps.

L'annexe de la thèse présente des préliminaires mathématiques utilisés lors de cette thèse.

Contents

List of Figures

Chapter 1

Introduction

Variational optimization problems in infinite dimension Hilbert spaces are fairly common in machine learning and control algorithms. In learning problems these appear as an optimization over a space of functions (denoted here as $\mathcal{H}$) to minimize a loss functional given some training data and a regularizer used to guide the optimal solution towards a certain bias. Given specific forms of the loss functional and regularizers, kernel methods and representer theorems have a long history [1, 2, 3, 4, 5] of being used to reformulate such problems to an equivalent optimization over finite dimensional spaces. The equivalent finite dimensional solution to the infinite dimensional problem is referred to as a representer for the optimal solution. Generalized representer theorems [6, 7, 8, 9] address the concern of when such an equivalent representer solution exists in a setting that is mostly agnostic to the specific form of the loss function and regularizer. Instead, necessary and sufficient conditions are given in terms of general properties required from the loss function and regularizer, for such a representer to exists. [6, 7, 8, 9] present generalized representer theorems for problems with a finite collection of bounded linear functionals mapping the original Hilbert space $\mathcal{H}$ into an euclidean space $\mathbb{R}^m$ and with loss functionals on such an $\mathbb{R}^m$. Chapter 2 presents an extension of this work to the case where we have possibly unbounded, closed, densely defined operators mapping the Hilbert space $\mathcal{H}$ to another separable Hilbert space $\mathcal{Z}$ and for loss functionals defined on such a $\mathcal{Z}$. Thus, equivalently, we allow for an infinite collection of linear functionals (possibly unbounded) mapping $\mathcal{H}$ to $\mathcal{Z}$ with loss functionals on $\mathcal{Z}$. Necessary and sufficient conditions for the existence of linear representers is presented for this setting. The utility of such an extension is shown for the case of learning stochastic processes where $\mathcal{Z}$ can in general be a space of measurable functions. A technical assumption of “r-regularity” from the previous counter-part of the generalized theorem [9] is dropped and its implications are shown with an example of ℓ_1-regularization in function spaces. Also an application of the theorem is presented to the case of neural networks involving closed, densely defined operators.

In the second part of the thesis, variational problems for optimal control and model predictive control (MPC) schemes are presented. Here finite dimensional approximations for the problems are considered through a sampled time approach, discretizing the time horizon for the controller into finitely many time periods. Integrator approximations are used to solve the continuous time ordinary differential equations governing the dynamics for a system and the initial and end states of the these discrete time segments are treated as decision variables in a finite dimensional optimization problem. The optimal control approaches are presented as their continuous time formulation for the theoretical results, while the discrete time approximation is used for numerical results.

Chapter 3 presents a combination of manifold learning with optimal control for obstacle avoidance in autonomous driving vehicles. The manifold learning algorithm is used to learn star shaped sets in the environment within which the vehicle can move without encountering any obstacles. A single inequality constraint can be used to check the containment of any point within a star shaped set and thus including such a constraint in the optimal control problems allows for planning and control of the vehicle within a collection of star shaped sets, thereby avoiding any obstacles. A multiphase free-end-time optimal control formulation is presented for planning trajectories within such sets and a moving horizon version of the multiphase free-end-time problem is used to formulate a path following MPC controller capable of controlling the vehicle in presence of dynamic obstacles. Under mild assumptions on vehicle dynamics and obstacle movement, convergence and recursive feasibility guarantees are provided for the MPC formulation. Numerical studies for the trajectory planning and MPC control scheme are presented in presence of static and dynamic obstacles.

Chapter 4 presents a path following MPC formulation for an Airborne wind energy system. A different, fixed-end-time formulation for path following is presented compared to the one used in Chapter 3 to allow for faster control computation. A terminal constraint, inspired by a convergent vector field design problem, is incorporated in the MPC scheme to ensure stability and convergence of the controller to a zero tracking error along the reference path. Proof for recursive feasibility is presented, under a reachability assumption for the designed vector field constraint. Finally, numerical results for the approach are presented under nominal and perturbed simulation settings for the AWE system. The path following controller shows robust convergence to the desired path under perturbed conditions and model mismatch, and the vector field based terminal constraints included in the problem, show an improved convergence rate for the error, while providing formal guarantees for the convergence.

The final part of the thesis, presents solutions for optimization and control in an AWE system in presence of uncertain wind conditions and unknown aerodynamic characteristics for the vehicle. Chapter 5 presents a data based optimization scheme for a closed-loop AWE system. Given a simple low-level controller for the system, the functions mapping the set-points of the controller to the closed-loop performance for a given objective, constraint and dynamics are unknown. Gaussian process (GP) surrogate models are used to represent the unknown mappings and a Gaussian process optimization scheme is presented to select the candidate points at which the performance is expected to improve while satisfying the unknown constraints with high confidence. Surrogate GP models for the closed-loop dynamics are included in the optimization scheme to allow for faster learning and optimization in the system with transient measurements. The approach is compared to a GP optimization scheme without the dynamics, using only steady state measurements and the inclusion of the transient measurements and dynamics, is shown to result in more robust solutions under varying wind conditions and in faster convergence to the optimal performance.

Chapter 6 presents a nonlinear direct adaptive control approach for path following in AWE systems under parametric uncertainties in the open loop dynamics for the system. The wind and aerodynamic characteristics are shown to be representable as unknown, affine parameters in a nonlinear dynamical system. A nonlinear controller is designed for approximate feedback linearization of the system, assuming the parameters as known quantities. The effects due to the unknown affine parameters are canceled out by the design of an augmented dynamical system, updating parameter estimates in such a way that the feedback linearization controller using these parameter estimates in place of the real, unknown parameters results in an asymptotically stable controller. The tracking error for the path following problem is shown with a Lyapunov function based argument to converge to a small neighborhood of zero and the parameter dynamics are shown to result in bounded estimates for the parameters.

The Appendix is used to present some mathematical preliminaries used for results in Chapter 2 and useful for further extensions to the work.

Part I

Representer theorems for variational problems in learning and control

Chapter 2

Generalized Representer Theorems

The necessary and sufficient conditions for existence of a generalized representer theorem are presented for learning Hilbert space - valued functions. Representer theorems involving explicit basis functions and Reproducing Kernels are a common occurrence in various machine learning algorithms like generalized least squares, support vector machines, Gaussian process regression, and kernel-based deep neural networks to name a few. Due to the more general structure of the underlying variational problems, the theory is also relevant to other application areas like optimal control, signal processing and decision making. The following presents a generalized representer theorem using the theory of closed, densely defined linear operators and subspace valued maps as a means to address variational optimization problems in learning and control. The implications of the theorem are presented with examples of multi-input - multi-output problems from kernel-based deep neural networks, stochastic regression and sparsity learning problems.

2.1 Introduction

The development of kernel-based methods for regression and machine learning has a long history with several algorithms basing themselves on the Reproducing Kernel Hilbert Space (RKHS) theory. Some of the early works in the field include [1, 4, 2], which looked at problems of spline interpolation and smoothing in the RKHS setting. Several practical learning algorithms like linear regression, support vector machines, Bayesian regression were also developed in their kernel forms to allow more complex nonlinear representations of data (see [10] for some examples). Kernel-based stochastic models are also popular in the form of Gaussian Process models (see [11]). Kernel based neural networks are investigated in [12, 13, 14, 15, 16].

Most such problems (see example 2.6) from learning and control in their general form can be written as a variational optimization problem of the following form,

$$f_{opt} = \arg\min_{f \in \mathcal{H}} \quad C(Lf) + \Omega(f) \tag{2.1}$$

where $\mathcal{H}$ and $\mathcal{Z}$ are some separable Hilbert spaces (possibly infinite dimensional, e.g. spaces of square integrable functions), $L : \mathcal{H} \to \mathcal{Z}$ is a closed, densely defined linear operators, and $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are general nonlinear functionals encoding the cost functions, regularizers and constraints in the problem. The functionals C and Ω are written separately as different properties are assumed to hold for the two functionals (see Section 2.3). We also hide the fact (in the notation of $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$) that the functional can be dependent on additional inputs like the data set used for learning which are fixed during the optimization and thus not explicitly shown in the notation.

Let $\mathcal{V}(\mathcal{H})$ denote the collection of all closed vector subspaces in $\mathcal{H}$ and $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ denote a map from a vector in $\mathcal{H}$ to a closed subspace of $\mathcal{H}$. Also let S have a subspace valued extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ given by the union operation, i.e, for any $A \in \mathcal{V}(\mathcal{H})$, $S(A) = \cup_{a \in A} S(a)$, must belong to $\mathcal{V}(\mathcal{H})$. Let $L^* : \mathcal{Z} \to \mathcal{H}$, defined on a dense subset $\text{dom}(L^*) \subseteq \mathcal{Z}$, denote the adjoint operator to the closed, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$. Let $\text{range}(L^*)$ denote the range of the operator L^* given by the set $\{L^*(z) : z \in \text{dom}(L^*)\}$. The generalized representer theorem (Theorem 2.3) then states under certain assumptions on C, Ω and S, that an optimal solution for (2.1) can be found in a subspace of $\mathcal{H}$ (often finite dimensional) given by $S(\text{range}(L^\star))$, i.e.,

$$f_{opt} \in S(\text{range}(L^*)) \tag{2.2}$$

Representer theorems thus provide a means to reduce infinite dimensional optimization problems for learning in the Hilbert space $\mathcal{H}$ to an equivalent and often tractable finite dimensional optimization in $\mathcal{Z}$ of the form,

$$\begin{aligned} f_{opt} &= L^* z_{opt} \\ z_{opt} &= \underset{f \in S(\text{range}(L^*))}{\arg\min} \quad C \circ Lf + \Omega(f) \end{aligned} \tag{2.3}$$

If $\mathcal{Z}$ and $S(\text{range}(L^*))$ are finite dimensional then (2.3) is a finite dimensional optimization.

A key underlying tool in the use of RKHS methods is the Riesz Representer Theorem [17, Theorem 3.3.1] and the existence and uniqueness of adjoint operators for bounded linear operators given by [17, Theorem 5.4.2]. The above two theorems combined with restrictions on the forms of the objective and constraint functionals in learning problems have led to several variants of representer theorems. Early variants of representer theorems are presented in [2] for variational problems in learning real valued functions with least squares regularization. Representer theorems for kernel versions of different learning algorithms like SVM, PCA, CCA, ICA can be found in [3]. Works like [18, 19, 20] present representer theorems for kernel based learning methods for vector valued functions in Hilbert spaces. While these works cover a large set of learning algorithms, the representer theorem needed to be proven individually for each problem. This has prompted investigation into unifying representer theorems into a single generalized theorem and characterizing the class of problems for which a representer theorem can be guaranteed to exist.

The first such results appear to have come from [6], where the problem is addressed for learning real valued functions with functionals of the form (2.4).

$$f_{opt} = \underset{f \in \mathcal{H}}{\arg\min} \quad C(f(x_1), \dots, f(x_m)) + \Omega(||f||_{\mathcal{H}}) \tag{2.4}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space of $\mathbb{R}$-valued functions with kernel K, $f(x_1), \dots, f(x_m)$ are function evaluations for f at given points $x_1, \dots, x_m$. The functional C is of the form, C :

$\mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, and $\Omega : [0, \infty) \to \mathbb{R}$ is a strictly monotonically increasing function. The strictly increasing monotonic property of Ω was shown to be a sufficient condition for the existence of a representer such that,

$$f_{opt} \in \left\{ \sum_{i=1}^{m} c_i K(\cdot, x_i) : c_i \in \mathbb{R} \right\} \tag{2.5}$$

The regularizers (written as a function of the norm of f) showed how kernel versions of the least squares algorithms in linear regression, SVMs and others are covered by a single generalized theorem.

[8] relaxed the restriction on the regularizer further and provided necessary and sufficient conditions for the existence of representer theorems. [8] considers problems of the form

$$f_{opt} = \arg\min_{f \in \mathcal{H}} \quad C(\langle w_1, f \rangle_{\mathcal{H}}, \ldots, \langle w_m, f \rangle_{\mathcal{H}}) + \Omega(f) \tag{2.6}$$

where $\mathcal{H}$ is a separable Hilbert space, $w_1, \ldots, w_m \in \mathcal{H}$ are given vectors corresponding to bounded functionals on $\mathcal{H}$ and functionals $C : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are lower semi-continuous functionals. If for all orthogonal vectors $f, g \in \mathcal{H}$ ($\langle f, g \rangle_{\mathcal{H}} = 0$), $\Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$, the functional Ω is called "orthomonotone". It was also shown that this orthomonotone property is necessary and sufficient for the existence of a representer in the form,

$$f_{opt} \in \left\{ \sum_{i=1}^{m} c_i w_i : c_i \in \mathbb{R} \right\} \tag{2.7}$$

[6, 8] restricted the scope of their theorem to learning $\mathbb{R}$-valued functions. The generalized theorem was extended to learning multi-output functions in [9] with the help of subspace valued maps $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. [9] considers problems of the form,

$$f_{opt} = \arg\min_{f \in \mathcal{H}} \quad C(\langle w_1, f \rangle_{\mathcal{H}}, \ldots, \langle w_m, f \rangle_{\mathcal{H}}) + \Omega(f) \tag{2.8}$$

where $\mathcal{H}, w_1, \ldots, w_m$, $C : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are as before from [8]. However, Ω satisfies the orthomonotone property with respect to a subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, defined as, for any $f \in \mathcal{H}$ and $g \in S(f)^{\perp}$, $\Omega(f + g) \geq \Omega(f)$. The representer theorem then provides that,

$$f_{opt} \in \sum_{i=1}^{m} S(w_i) \tag{2.9}$$

(the summation over sets $S(w_i) + S(w_j)$ being considered as the pairwise addition $a + b$ of all possible pairs $(a, b) \in S(w_i) \times S(w_j)$).

The results from [6, 8] can be viewed under this framework as Ω being orthomonotone with respect to a trivial map $S_{\mathbb{R}}(w_i) = \{\lambda w_i : \lambda \in \mathbb{R}\}$. The inclusion of orthomonotonicity with respect to non trivial subspace valued maps allows the consideration of a larger class of regularizers for Ω including regularizers like the ℓ_1-norm, Frobenius norm, trace norm and general spectral norms in matrix learning problems [7]. For learning a matrix $f \in \mathcal{H} = \mathbb{R}^{m \times n}$, with Ω being a monotonically increasing penalty on $f^T f$, [9, Example 4.2] shows the representer is of the form $\sum_{i=1}^{m} S(w_i)$ with

$S(w_i) = \{w_i c_i : c_i \in \mathbb{R}^{n\times n}\}$ for given $w_i \in \mathcal{H} = \mathbb{R}^{m\times n}$, thus showing the role of S in extending the result from [8] to a multi-output scenario.

[9] however makes an assumption of "r-regularity" (see appendix for definition) on the allowed subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, which requires for all $w \in \mathcal{H}$, the dimension of $S(w) \leq r$ for some finite $r \leq m$. We show in Section 2.4.3 that for ℓ_1-regularization on function spaces the functional Ω is orthomonotone with respect to a non r-regular subspace valued map S, i.e. no such finite r exists for all $w \in \mathcal{H}$. Theorem 2.3 eliminates the r-regularity assumption and enables the generalized representer theorem to be applied to such problems.

The prior counter parts of Theorem 2.3 [6, 7, 8, 9] also consider functionals $C : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ defined on $\mathbb{R}^m$ instead of an arbitrary separable Hilbert space $\mathcal{Z}$. In Section 2.4.2, we show with an example of stochastic process regression the utility of considering loss functionals $C : \mathcal{Z} \to \mathbb{R}\cup\{\infty\}$ over an infinite dimensional Hilbert space $\mathcal{Z}$. The learning problems for stochastic processes require loss functionals to be defined over a Hilbert space of measurable functions (not isomorphic to $\mathbb{R}^m$) and were thus outside the scope of previous generalized representer theorems from [6, 7, 8, 9].

We thus present here an extension for the generalized representer theorem where the functional $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is a lower semi-continuous non-linear functional over an arbitrary Hilbert space $\mathcal{Z}$, in terms of non r-regular subspace valued maps and adjoints of closed, densely defined linear operators.

The chapter is structured as follows. Section 2.2 presents some preliminary definitions and results of existing notions required to establish the generalized representer theorem. Section 2.2.1 presents some background material on linear operators and their adjoints. Section 2.2.2 presents the notion of a subspace valued map and Section 2.2.3 presents the notion of orthomonotone functionals with respect to a subspace valued map. The generalized representer theorem giving necessary and sufficient conditions for the existence of a representer is then presented in Section 2.3. Section 2.4 presents examples of some simple learning problems to highlight extensions made by the representer theorem. The appendix provides proofs for some lemmas and discussion with regards to the subspace valued maps considered in the chapter and their relation to properties of quasilinear, idempotent and r-regular subspace valued maps used in previous works.

2.2 Preliminaries

The notions of adjoints and closed operators are known to be crucial in determining solutions to linear inverse problem of the form $Lx = y$ (find x given y) [21]. It is thus natural for them to be important in the theory for a generalized representer theorem (which cover problems of the form $Lx = y$ as a special case). Section 2.2.1 presents some preliminary, well known results that will be useful in proving the generalized representer theorem.

2.2.1 Closed linear operators and adjoint operators

Let $\mathcal{H}$ and $\mathcal{Z}$ be two arbitrary separable Hilbert spaces. Let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, $\langle\cdot,\cdot\rangle_{\mathcal{Z}}$ be the inner products defined on $\mathcal{H}$ and $\mathcal{Z}$ respectively. A closed linear operator from $\mathcal{H}$ to $\mathcal{Z}$ is defined as follows.

Definition 2.1. *(Closed linear operator)*
Let $\mathcal{H}, \mathcal{Z}$ be two separable Hilbert spaces. Let $\operatorname{dom}(L) \subseteq \mathcal{H}$ *be the domain for a linear operator*

$L : \mathrm{dom}(L) \to \mathcal{Z}$. *$L$ is called a closed operator if the graph of the operator,* $\mathrm{graph}(L) = \{(x, Lx) : x \in \mathrm{dom}(L)\}$ *is a closed subset of* $\mathcal{H} \times \mathcal{Z}$.

An operator L is called closable if there exists an extension to L that is closed.

A linear operator (not necessarily closed) is said to be densely defined on $\mathcal{H}$ if $\mathrm{dom}(L)$ is a dense subset of $\mathcal{H}$. Let $L : \mathrm{dom}(L) \to \mathcal{Z}$ be a linear operator, densely defined on $\mathcal{H}$. Then an adjoint operator can be defined as follows,

Definition 2.2. *(Adjoint for densely defined operators)*
Let $\mathrm{dom}(L)$ *be a dense subset of* $\mathcal{H}$ *and* $L : \mathrm{dom}(L) \to \mathcal{Z}$ *be a densely defined operator (also denoted as* $L : \mathcal{H} \to \mathcal{Z}$*). Let* $\mathrm{dom}(L^*) := \{z \in \mathcal{Z} : f(h) = \langle Lh, z\rangle_{\mathcal{Z}}$ *is bounded linear functional on* $\mathrm{dom}(L)\}$*. The adjoint* $L^* : \mathrm{dom}(L^*) \to \mathcal{H}$ *is defined as the operator mapping* $z \in \mathrm{dom}(L^*)$ *to a dual in* $\mathcal{H}$ *such that,*

$$\forall f \in \mathrm{dom}(L), z \in \mathrm{dom}(L^*) \qquad \langle Lf, z\rangle_{\mathcal{Z}} = \langle f, L^*z\rangle_{\mathcal{H}} \tag{2.10}$$

By [22, Chapter 10, Proposition 1.6], if the operator $L : \mathrm{dom}(L) \to \mathcal{Z}$ is closable and densely defined then the adjoint L^* is a closed, densely defined operator, i.e., $\mathrm{dom}(L^*)$ is a dense subset of $\mathcal{Z}$. For a closed densely defined operator $L : \mathcal{H} \to \mathcal{Z}$, $L^*L : \mathrm{dom}(L^*L) \subseteq \mathcal{H} \to \mathcal{H}$ and $LL^* : \mathrm{dom}(LL^*) \subseteq \mathcal{Z} \to \mathcal{Z}$ are closed, densely defined, self-adjoint operators [23]. Also, for a closed and bounded operator $L : \mathrm{dom}(L) \subseteq \mathcal{H} \to \mathcal{Z}$ the domain is the entire space, i.e. $\mathrm{dom}(L) = \mathcal{H}$ and the adjoint L^* is also closed and bounded.

Further, by Banach's closed range theorem [24, Chapter 7.5], the null space of a densely defined, closed linear operator $\mathcal{N}_L = \{f \in \mathrm{dom}(L) : Lf = 0\}$ is a closed subset in $\mathcal{H}$ and can be characterized in terms of the orthogonal complementary space $\mathcal{N}_L^\perp$ and the adjoint operator $L^* : \mathrm{dom}(L^*) \to \mathcal{H}$ as follows,

Lemma 2.1. *Let* $\mathcal{N}_L$ *be the null space of a closed, densely defined operator* $L : \mathrm{dom}(L) \to \mathcal{Z}$ *and* $\mathcal{N}_L^\perp$ *be its orthogonal complementary space, then,*

$$\mathcal{N}_L^\perp = \mathrm{range}(L^*) = \{L^*z : z \in \mathrm{dom}(L^*)\}$$

The above lemma is a direct result of the closed range theorem and we refer the reader to [24, Chapter 7.5] for the proof.

Corollary 2.1. *For some finite* $m \in \mathbb{N}$, *let* $\{L_i : \mathrm{dom}(L_i) \to \mathcal{Z}_i : i = 1, \ldots, m\}$ *be a set of closed, densely defined operators with* $\mathcal{Z}_i$ *being separable Hilbert spaces and* $\mathrm{dom}(L_i) \subseteq \mathcal{H}$ *for some separable Hilbert space* $\mathcal{H}$*. Let* $\cap_{i=1}^m \mathrm{dom}(L_i)$ *be a dense subset of* $\mathcal{H}$*. The joint null space is* $\mathcal{N}_{L_1,\ldots,L_m} = \mathcal{N}_{L_1} \cap \cdots \cap \mathcal{N}_{L_m}$ *and* $\mathcal{N}^\perp_{L_1,\ldots,L_m} = \mathcal{N}^\perp_{L_1} + \cdots + \mathcal{N}^\perp_{L_m} = \{\sum_{i=1}^m L_i^* z_i : z_i \in \mathrm{dom}(L_i^*)\}$.

Proof: *Consider the Hilbert space* $\mathcal{Z}$ *given as the direct sum of* $\mathcal{Z}_i$, $i = 1, \ldots, m$, *i.e.* $\mathcal{Z} = \mathcal{Z}_1 \oplus \cdots \oplus \mathcal{Z}_m$. *The inner product on* $\mathcal{Z}$ *is given by* $\langle (z_1, \ldots, z_m), (y_1, \ldots, y_m)\rangle_{\mathcal{Z}} = \sum_{i=1}^m \langle z_i, y_i\rangle_{\mathcal{Z}_i}$. *Consider then the linear operator* $L : \cap_{i=1}^m \mathrm{dom}(L_i) \to \mathcal{Z}$ *given as* $Lf = (L_1 f, \ldots, L_m f)$*. By assumption,* $\cap_{i=1}^m \mathrm{dom}(L_i)$ *is a dense subset of* $\mathcal{H}$ *and thus* L *is a densely defined operator. Further* $\mathrm{graph}(L) = \{(x, Lx) : x \in \mathrm{dom}(L)\}$ *is a closed subset since for every converging sequence* $x_n \in \mathrm{dom}(L)$, $Lx_n = (L_1 x_n, \ldots, L_m x_n)$ *converges to a point* $(L_1 x, \ldots, L_m x)$ *with* $(x, L_i x) \in \mathrm{graph}(L_i)$ *(since* L_i *is a closed operator). Thus* L *is a closed, densely defined operator with* $\mathrm{dom}(L) = \cap_{i=1}^m \mathrm{dom}(L_i)$.

Clearly $\mathcal{N}_L = \mathcal{N}_{L_1,\ldots,L_m} = \cap_{i=1}^m \mathcal{N}_{L_i}$. The adjoint domain $\operatorname{dom}(L^*) = \{(z_1,\ldots,z_m) \in \mathcal{Z} : f(h) = \sum_{i=1}^m \langle L_i f, z_i\rangle_{\mathcal{Z}_i}$ *is bounded*$\} = \operatorname{dom}(L_1^*) \times \cdots \times \operatorname{dom}(L_m^*)$. *The adjoint $L^* : \operatorname{dom}(L^*) \to \mathcal{H}$, is such that, for all $f \in \operatorname{dom}(L)$ and $z = (z_1,\ldots,z_m) \in \operatorname{dom}(L^*)$, $\langle L^*z, f\rangle_{\mathcal{H}} = \langle z, Lf\rangle_{\mathcal{Z}} = \sum_{i=1}^m \langle z_i, L_i f\rangle_{\mathcal{Z}_i} = \langle \sum_{i=1}^m L_i^* z_i, f\rangle_{\mathcal{H}}$. Thus $L^*z = \sum_{i=1}^m L_i^* z_i$. Then by Lemma 2.1, $\mathcal{N}_L^\perp = \operatorname{range}(L^*) = \{\sum_{i=1}^m L_i^* z_i : z_i \in \operatorname{dom}(L_i)\}$.* □

When rewriting functionals of the form $C(L_1f,\ldots,L_mf)$ as $C(Lf)$, Corollary 2.1 gives the required characterization of the orthogonal null space. Thus the adjoint operator plays a key role in characterizing the null space of an operator $\mathcal{N}_L$ and its orthogonal complementary space $\mathcal{N}_L^\perp$.

Closed range characterization for bounded linear operators in terms of the operator spectrum are given in [21, Theorem 2.5] or equivalently by [17, Lemma 5.6.13]. Characterization of closed, densely defined operators is given by [25, Theorem 3.3].

By [17, Proposition 5.6.13], a bounded adjoint $T^* : \mathcal{Z} \to \mathcal{H}$ is a closed range if and only if

$$\inf\{||L^*z||_{\mathcal{H}} : ||z||_{\mathcal{Z}} = 1\} > 0 \tag{2.11}$$

or equivalently

$$\inf\{\langle z, LL^*z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0 \tag{2.12}$$

By [25, Theorem 3.3] a densely defined operator is closed if and only if, there exists a $\gamma > 0$ such that the spectrum $\sigma(L^*L) \subseteq \{0\} \cup [\gamma, \infty)$.

Adjoint for operators of common interest

Below we show a few examples of adjoint operator for densely defined, closed linear operators commonly seen in learning and control algorithms.

Example 2.1. *(Evaluation Operators)*
Let $\mathcal{Z}$ be a separable Hilbert space and $\mathcal{C}_b(\mathcal{X},\mathcal{Z})$ be the separable Banach space of $\mathcal{Z}$-valued continuous and bounded functions with a domain set $\mathcal{X}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ induced by a Gaussian measure on $\mathcal{C}_b(\mathcal{X},\mathcal{Z})$. A parametric linear evaluation operator $L_x : \mathcal{H} \to \mathcal{Z}$, given by $L_x(f) = f(x)$ for some fixed parameter $x \in \mathcal{X}$ is then a bounded linear operator (see Theorem C.9) and $\mathcal{H}$ is a dense subset in $\mathcal{C}_b(\mathcal{X},\mathcal{Z})$ [26, Theorem 3.9.5]. The operator commonly occurs in machine learning and data fitting problems where x is the training input data and $f(x)$ gives a predicted value for the output in $\mathcal{Z}$. The adjoint $L_x^* : \mathcal{Z} \to \mathcal{H}$ can be found as follows.

Note that by definition of L_x and its adjoint L_x^*, $\forall g \in \mathcal{H}, z \in \mathcal{Z}$, $\langle L_x^* z, g\rangle_{\mathcal{H}} = \langle L_x g, z\rangle_{\mathcal{Z}}$, i.e., $\langle L_x^* z, g\rangle_{\mathcal{H}} = \langle g(x), z\rangle_{\mathcal{Z}}$. When $\mathcal{H}$ is a reproducing kernel Hilbert space with kernel K, L_x^* is well defined and coincides with the definition of the RKHS kernel (see [18, Definition 2.1] or Theorem C.9). Thus RKHS spaces provide a case where the adjoint operator for evaluation operators is well defined and $L_x^* = K(\cdot, x)$, i.e. we have $\operatorname{dom}(L_x) = \mathcal{H}$ *and* $\operatorname{dom}(L_x^*) = \mathcal{Z}$.

The closed range property for L_x^* thus corresponds to the closed range property of the kernel. Using (2.12), this corresponds to checking $\inf\{\langle z, K(x,x)z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0$. *For a positive definite kernel K, this is automatically satisfied and the adjoint is a closed range, bounded linear operator.*

Example 2.2. *(Linear Transformations of an explicit basis ϕ)*
Let $\mathcal{H}, \mathcal{Y}, \mathcal{Z}$ be arbitrary Hilbert spaces. Let $\mathcal{L}_{\mathcal{Y},\mathcal{Z}}$ be the space of bounded, closed range operators from $\mathcal{Y}$ to $\mathcal{Z}$. Let $\phi : \mathcal{X} \to \mathcal{Y}$ be some given $\mathcal{Y}$-valued function and $x \in \mathcal{X}$ be an evaluation point such that $||\phi(x)||_{\mathcal{Y}} < \infty$. Let $\ell : \mathcal{H} \to \mathcal{L}_{\mathcal{Y},\mathcal{Z}}$ be a bounded linear map from $\mathcal{H}$ to $\mathcal{L}_{\mathcal{Y},\mathcal{Z}}$ such that there exists a $\kappa_\ell \in [0, \infty)$ satisfying, for all $W \in \mathcal{H}$, $||\ell(W)||_{\mathcal{L}_{\mathcal{Y},\mathcal{Z}}} \leq \kappa_\ell ||W||_{\mathcal{H}}$. Then we can define a bounded, closed range linear operator $L_{x,\phi} : \mathcal{H} \to \mathcal{Z}$ given as $L_{x,\phi}(W) := \ell(W)\phi(x)$ for any $W \in \mathcal{H}$. The boundedness for the operator follows from the fact that $||L_{x,\phi}(W)||_{\mathcal{Z}} = ||\ell(W)\phi(x)||_{\mathcal{Z}} \leq ||\ell(W)||_{\mathcal{L}_{\mathcal{Y},\mathcal{Z}}} ||\phi(x)||_{\mathcal{Y}} \leq \kappa_\ell ||\phi(x)||_{\mathcal{Y}} ||W||_{\mathcal{H}}$. The adjoint operator satisfies $\langle L^*_{x,\phi} z, W\rangle_{\mathcal{H}} = \langle \ell(W)\phi(x), z\rangle_{\mathcal{Z}}$ and its form depends on further specification of ℓ.

The operator $L_{x,\phi}$ is closed range, if $\inf\{\langle L_{x,\phi} L^*_{x,\phi} z, z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0$, i.e., $\inf\{\langle \ell(L^*_{x,\phi} z)\phi(x), z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0$.

We look at two examples below giving ℓ explicitly and making the adjoint and closed range characterization for the given cases.

Example 2.2(a). Finite dimensional $\mathcal{Z}$ example
Let $\mathcal{Y} = \mathbb{R}^n$, $\mathcal{Z} = \mathbb{R}^k$, $\mathcal{H} = \mathbb{R}^{n \times k}$ and $\phi : \mathcal{X} \to \mathcal{Y}$ is a given basis function and $x \in \mathcal{X}$ with $||\phi(x)||_{\mathcal{Y}} < \infty$. Let $\ell(W) := W^T$ be the bounded operator from $\mathcal{H}$ to $\mathcal{L}_{\mathcal{Y},\mathcal{Z}}$. Then for any $W \in \mathcal{H}$, $L_{x,\phi}(W) = W^T\phi(x)$ and $L_{x,\phi}$ is a bounded operator. Such an operator is common when W represent weights or coefficients to be learned and ϕ is a given vector of basis functions.

Let the inner product on $\mathcal{H}$ be the Frobenius inner product of matrices, i.e, $\langle w_1, w_2\rangle_{\mathcal{H}} = \mathrm{trace}(w_1^T w_2)$. Let inner product on $\mathcal{Z}$ be $\langle z_1, z_2\rangle_{\mathcal{Z}} = z_1^T z_2$. Then for the adjoint operator $\langle L^*_{x,\phi} z, W\rangle_{\mathcal{H}} = \langle W^T\phi(x), z\rangle_{\mathcal{Z}}$, $\forall z \in \mathcal{Z}$ implying $\mathrm{trace}(W^T L^*_{x,\phi} z) = \phi(x)^T W z$. Noting then that $\phi(x)^T W z = \mathrm{trace}(\phi(x)^T W z) = \mathrm{trace}(z^T W^T \phi(x)) = \mathrm{trace}(W^T\phi(x) z^T)$, we can define $L^*_{x,\phi} z := \phi(x) z^T$ with $\mathrm{dom}(L_{x,\phi}) = \mathcal{H}$ and $\mathrm{dom}(L^\star_{x,\phi}) = \mathcal{Z}$.

The operator $L_{x,\phi}$ is closed range if $\inf\{\phi(x)^T\phi(x) z^T z : ||z||_{\mathcal{Z}} = 1\} = \phi(x)^T\phi(x) > 0$.

Example 2.2(b). Infinite dimensional $\mathcal{Z}$ example
Let $\mathcal{X} = \mathbb{R}^n$, $\mathcal{U} = \mathbb{R}^m$ and $\mathcal{H} = \mathbb{R}^{m \times N}$. Let $\{\mathcal{Y}_i : i = 1, \ldots, N\}$ be a collection of RKHS spaces of functions $f : \mathcal{X} \to \mathcal{U}$ with kernels $K_1, \ldots, K_N$. Let $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$ and $\phi(x) = \begin{pmatrix} K_1(\cdot, x) \\ K_2(\cdot, x) \\ \vdots \\ K_N(\cdot, x) \end{pmatrix}$ and let $\mathcal{Z}$ be some infinite dimensional Hilbert space of functions $g : \mathcal{X} \to \mathcal{U}$ with inner product $\langle g_1, g_2\rangle_{\mathcal{Z}} = \int_{\mathcal{X}} \langle g_1(x), g_2(x)\rangle_{\mathcal{U}} dx$. Then we can define a continuous linear operator $L_{x,\phi} : \mathcal{H} \to \mathcal{Z}$ for any $W \in \mathcal{H}$ as $L_{x,\phi}(W) := \sum_{i=1}^N K_i(\cdot, x) W_i$, with W_i denoting the i^{th} column of W. Using the Frobenius inner product on $\mathcal{H}$, $\langle L^*_{x,\phi} g, W\rangle_{\mathcal{H}} = \langle L_{x,\phi}(W), g\rangle_{\mathcal{Z}} = \sum_{i=1}^N \int_{\mathcal{X}} \langle K_i(y, x) W_i, g(y)\rangle_{\mathcal{U}} dy = \sum_{i=1}^N \int_{\mathcal{X}} W_i^T K_i(x, y) g(y) dy$. Also note that $\langle L^*_{x,\phi} g, W\rangle_{\mathcal{H}} = \mathrm{trace}(W^T L^*_{x,\phi} g) = \sum_{i=1}^N W_i^T [L^*_{x,\phi} g]_i$. Thus $[L^*_{x,\phi} g]_i = \int_{\mathcal{X}} K_i(x, y) g(y) dy$ gives the adjoint.

The operator $L_{x,\phi}$ is closed range, if $\inf\{\langle L^*_{x,\phi}(L_{x,\phi} W), W\rangle_{\mathcal{H}} : ||W||_{\mathcal{H}} = 1\} = \inf\{\langle L^*_{x,\phi} W, L^*_{x,\phi} W\rangle_{\mathcal{Z}} : ||W||_{\mathcal{H}} = 1\} = \inf\{\sum_{i=1}^N \sum_{j=1}^N (\int_{\mathcal{X}} W_i^T K_i(x, y)^T K_j(x, y) W_j dy) : ||W||_{\mathcal{H}} = 1\} > 0$.

Such an operator can be used to pose an optimization for learning with weighted kernels.

Example 2.3. *(Derivative operator in Sobolov Hilbert spaces)*
Let $\Omega \subset \mathbb{R}^n$ be a open subset of $\mathbb{R}^n$ with a smooth boundary $\partial\Omega$. Let $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ be a multi-index and $\partial^\alpha f = \partial_{x_1}^{\alpha_1}, \ldots, \partial_{x_n}^{\alpha_n} f$. Let $L^2(\Omega, \mathbb{R}, \mu)$ be the space of $\mathbb{R}$-valued functions, square integrable on Ω with respect to a non-negative measure μ and $H^k(\Omega, \mathbb{R}, \mu)$ be the Sobolov Hilbert space such that $\partial^\alpha f \in L^2(\Omega, \mu)$ for all multi-index $\alpha \in \mathbb{N}^n$ such that $|\alpha| \leq k$. The inner product on $H^k(\Omega, \mathbb{R}, \mu)$ is given by $\langle f, g\rangle_{H^k(\mu)} = \sum_{i=1}^k \sum_{\alpha:|\alpha|\leq k} \langle \partial^\alpha f, \partial^\alpha g\rangle_{L^2(\Omega,\mu)}$. It is also known that $H^k(\Omega, \mathbb{R}, \mu) \subset L^2(\Omega, \mathbb{R}, \mu)$ is a dense subset of $L^2(\Omega, \mathbb{R}, \mu)$ [27, Prop. 3.10]. Thus any differential operator $D : H^k(\Omega, \mathbb{R}, \mu) \to L^2(\Omega, \mathbb{R}, \mu)$ defined on $H^k(\Omega, \mathbb{R}, \mu)$ is densely defined on $L^2(\Omega, \mathbb{R}, \mu)$ with $\mathrm{dom}(D) = H^k(\Omega, \mathbb{R}, \mu)$. Consider then a differential operator $Df = \phi(\cdot)^T \nabla f + \Delta f$ for a given smooth function $\phi \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$, ∇ and Δ are the gradient and Laplace operators respectively. Such an operator D is closable [24, Page 78]. Thus we have D as a closable, densely defined operator on $L^2(\Omega, \mathbb{R}, \mu)$, implying the adjoint D^* is closed and densely defined on $L^2(\Omega, \mathbb{R}, \mu)$. For any g with differentiability upto order two and $f \in \mathrm{dom}(D)$ we have, using integration by parts,

$$\begin{aligned}
\langle Df, g\rangle_{L^2(\Omega,\mathbb{R},\mu)} &= \int_\Omega (\phi(x)^T \nabla f(x) + \Delta f(x)) g(x) d\mu(x) && (2.13)\\
&= \int_\Omega f(x)(-\nabla \cdot (g\phi)(x) + \Delta g(x)) d\mu(x) && (2.14)\\
&\quad + \int_{\partial\Omega} (\phi f g + g\nabla f - f \nabla g) \cdot dS && (2.15)
\end{aligned}$$

Then for the boundary conditions $g(x) = 0$ and $\nabla g(x) = 0$ for all $x \in \partial\Omega$, and defining

$$D^* g = -\nabla \cdot (g\phi)(x) + \Delta g(x)$$

we have the integral terms over the boundary going to zero and,

$$\langle Df, g\rangle_{L^2(\Omega,\mathbb{R},\mu)} = \langle f, D^* g\rangle_{L^2(\Omega,\mathbb{R},\mu)} \tag{2.16}$$

for all $f \in H^k(\Omega, \mathbb{R}, \mu)$ and $g \in \mathrm{dom}(D^*) = \{g \in H^2(\mathbb{R}^n, \mathbb{R}, \mu) : \forall x \in \partial\Omega,\ g(x) = 0, \nabla g(x) = 0\}$.

Example 2.3 shows an example of an unbounded operator where the domain $\mathrm{dom}(L^*)$ is a strict subset of $\mathcal{Z}$ unlike in the case of bounded, closed operators in Examples 2.1 and 2.2. Derivative operators with boundary conditions are common in numerical methods for control, signal processing and partial differential equation applications. Similarly the use of integral operators for learning has been considered for learning in [28].

2.2.2 Subspace Valued Maps

The notion of subspace valued maps expands the class of regularizers that a generalized representer theorem can explain and was introduced in [9]. Let $\mathcal{H}$ be a separable Hilbert space, $2^{\mathcal{H}}$ be the power set on $\mathcal{H}$ and $\mathcal{V}(\mathcal{H})$ be a set of all closed vector subspaces of $\mathcal{H}$. Also for any subsets $A, B \subseteq \mathcal{H}$, let $A+B$ denote the set $\{a+b : a \in A, b \in B\}$. A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is then called a subspace valued map. For evaluation on any set $A \subseteq \mathcal{H}$, we denote $S(A)$ to mean, $S(A) = \cup_{x\in A} S(x)$. The union operation, thus, extends the map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, in general, to a set valued map $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$

(as the union of vector spaces is not necessarily a vector space). Below we present a few definitions of terms we will use in the context of subspace valued maps and show conditions under which the union leads to closed vector spaces.

Definition 2.3. *(Subspace valued map)*
Let $\mathcal{H}$ be a separable Hilbert space and $\mathcal{V}(\mathcal{H})$ be a set of all closed vector subspaces of $\mathcal{H}$. A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***subspace valued****.*

Definition 2.4. *(Union extension)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a subspace valued map. Then the extension of $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ given by an union operation $S(A) = \cup_{x \in A} S(x)$ is called the union extension of S.

Definition 2.5. *(Inclusive map)*
A subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***inclusive****, if, for all $x \in \mathcal{H}$, $x \in S(x)$*

Definition 2.6. *(Super additive map)*
A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***super additive*** *if its union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ is super-additive, i.e. for all vector subspaces $A, B \in \mathcal{V}(\mathcal{H})$,*

$$S(A) + S(B) \subseteq S(A + B)$$

Note the the above name is a misnomer since we do not require $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ to be super-additive, but only its union extension to be super-additive. The misnomer is used for the purposes of brevity.

Definition 2.7. *(Closed map)*
A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called closed if its union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ maps closed subspaces from $\mathcal{V}(\mathcal{H})$ to closed subsets in $2^{\mathcal{H}}$.

Definition 2.8. *(Orthogonal subspace)*
For any $A \subseteq \mathcal{H}$, we define $S(A)^{\perp} := \{b \in \mathcal{H} : \forall a \in S(A), \langle a, b \rangle_{\mathcal{H}} = 0\}$

The following shows a few examples of inclusive and super-additive subspace valued maps that are used for application examples in Section 2.4,

Example 2.4. *Subspace valued maps*

1. *$S_{\mathbb{R}}(a) := \{\lambda a : \lambda \in \mathbb{R}\}$ is a closed, inclusive, super additive subspace valued map. Inclusivity of $S_{\mathbb{R}}$ is straightforward to see since $a = 1 \cdot a \in S(a) = \{\lambda \cdot a : \lambda \in \mathbb{R}\}$. Further for any $A, B \in \mathcal{V}(\mathcal{H})$, $S(A) + S(B) = \{\lambda_1 a + \lambda_2 b : \lambda_1, \lambda_2 \in \mathbb{R}, a \in A, b \in B\} = \{\lambda a : \lambda \in \mathbb{R}, a \in A + B\} = S(A + B)$. Also for any closed subspace $A \in \mathcal{V}(\mathcal{H})$, the union extension is such that $S_{\mathbb{R}}(A) = \cup_{a \in A} S_{\mathbb{R}}(a) = A$ and thus maps closed subspaces to closed subspaces.*

2. *Let $K = \{L_i : \mathcal{H} \to \mathcal{H} : i = 1, \ldots, n\}$ be a finite set of linearly independent, closed and bounded linear operators with the identity operator $I \in \mathrm{span}(K)$. Then $S_{\mathcal{L}}(a) := \{\sum_{i=1}^{n} \lambda_i L_i a : \lambda_i \in \mathbb{R}\}$ is a closed, inclusive and super additive subspace valued map. The fact that $S_{\mathcal{L}}$ is closed, can be seen by noting that for any closed subspace A, we have*

$\mathcal{S}_{\mathcal{L}}(A) = \sum_{i=1}^{n} L_iA$. Since L_i are closed linear operators, the sets L_iA are closed and the sum of finitely many closed sets remains closed. $S_{\mathcal{L}}$ being inclusive follows from the fact that the identity operator $Ia = a$ belongs to span(K)*, and thus $a \in S_{\mathcal{L}}(a)$, implying $S_{\mathcal{L}}$ is inclusive. Also for any closed vector subspaces $A, B \in \mathcal{V}(\mathcal{H})$, $S(A) + S(B) = \{\sum_{i=1}^{\infty} \lambda_i L_i a + \lambda'_i L_i b : \lambda_i, \lambda'_i \in \mathbb{R}, a \in A, b \in B\} = \{\sum_{i=1}^{\infty} L_i(\lambda_i a + \lambda'_i b) : \lambda_i, \lambda'_i \in \mathbb{R}, a \in A, b \in B\} = \{\sum_{i=1}^{\infty} L_i a : a \in A + B\} = S(A + B)$.*

3. *A special case of the above example is the case when $\mathcal{H} = \mathbb{R}^n$ and $E = \{e_1, \ldots, e_n\}$ is the standard orthonormal basis for $\mathbb{R}^n$. Then $S_{proj}(a) := \{\sum_{i=1}^{n} \lambda_i \langle a, e_i\rangle_{\mathcal{H}} e_i : e_i \in E, \lambda_i \in \mathbb{R}\}$ is an inclusive, super additive subspace valued map. The S_{proj} corresponds to $S_{\mathcal{L}}$ from the previous example, with $L_i : \mathcal{H} \to \mathcal{H}$, being a set of projections onto the orthonormal basis, given as $L_i a = \langle a, e_i\rangle_{\mathcal{H}} e_i$*

4. *A countable counterpart of the example above can be presented for the space of square summable sequences, $\mathcal{H} = \ell^2(\mathbb{N}, \mathbb{R})$, taking values in $\mathbb{R}$ and indexed by natural numbers $\mathbb{N}$. Let $\{\delta_i \in \ell^2(\mathbb{N}, \mathbb{R}) : i \in \mathbb{N}\}$ with $\delta_i(j) = 1$ if $i = j$ and 0 otherwise, be the orthonormal basis for $\ell^2(\mathbb{N}, \mathbb{R})$. Let $f(i)$ denote the i^{th} member of a sequence and let $\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f(i)g(i)$. Then $S_{proj}(f) = \left\{\sum_{i=1}^{\infty} \lambda(i) \frac{\langle f, \delta_i\rangle_{\mathcal{H}} \delta_i}{||f||_{\mathcal{H}}} : \lambda \in \ell^2(\mathbb{N}, \mathbb{R})\right\}$ for $||f||_{\mathcal{H}} \neq 0$ and $S_{proj}(f) = \{0\}$ if $||f||_{\mathcal{H}} = 0$, is an inclusive, closed and super additive subspace valued map. The S_{proj} defined can be seen to be inclusive as for any $f \in \ell^2(\mathbb{N}, \mathbb{R})$, there exists a representation for f in terms of the orthonormal basis $f = \sum_{i=1}^{\infty} a(i)\delta_i$ for some coefficients sequence $a \in \ell^2$. $S_{proj}(f) = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0 \text{ if } a(i) = 0\}$ and thus $f = \sum_{i=1}^{n} a(i)\delta_i$ belongs to $S_{proj}(f)$. Similarly for any $f = \sum_{i=1} a(i)\delta_i$ and $g = \sum_{i=1} b(i)\delta_i$ with $a, b \in \ell^2(\mathbb{N}, \mathbb{R})$, we have $S_{proj}(f) + S_{proj}(g) = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0 \text{ if } a(i) = 0\} + \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0 \text{ if } b(i) = 0\} = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0 \text{ if } b(i) = a(i) = 0\} = S_{proj}(f + g)$. Thus $S_{proj}(A) + S_{proj}(B) = S_{proj}(A + B)$ for all $A, B \in \mathcal{V}(\mathcal{H})$ and thus it is trivially super-additive. Also S_{proj} is closed as it maps any $A \in \mathcal{V}(\mathcal{H})$, to $S_{proj}(A) = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}) \text{ and } \lambda(i) = 0 \text{ if } a(i) = 0 \text{ for all } a \in A\}$ which is a closed vector subspace of $\ell^2(\mathbb{N}, \mathbb{R})$*

Examples 2.4-3 and 2.4-4 are used to construct representers for regularizers given by ℓ_1 norm.

Noting that the union extension of a subspace valued map $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$, in general, is not subspace valued, the following Lemma shows that a subspace valued union extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ to a subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ exists, if and only if, the union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ is super-additive.

Lemma 2.2. *(Extending $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ to $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a subspace valued map and its union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ be given by $S(A) = \cup_{x \in A} S(x)$. Then the extension maps into $\mathcal{V}(\mathcal{H})$, if and only if, S is super-additive and closed.

Proof: *We first prove that if S is super-additive and closed then for any $A \in \mathcal{V}(\mathcal{H})$, $S(A) \in \mathcal{V}(\mathcal{H})$ and thus the extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ has range in $\mathcal{V}(\mathcal{H})$.*

To show $S(A) \in \mathcal{V}(\mathcal{H})$, we need to show that for any $a, b \in S(A)$, $\lambda a + \mu b \in S(A)$ for all $\lambda, \mu \in \mathbb{R}$ and that any converging sequence $\{a_n \in S(A)\}$ converges within $S(A)$.

First, we show that $S(A)$ is a vector space if S is super-additive.

Since $S(A) = \cup_{x \in A} S(x)$, for any $a \in S(A)$, there exists a $x_a \in A$ such that $a \in S(x_a)$. Further, $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ maps $x_a \in \mathcal{H}$ to a closed vector space $S(x_a) \in \mathcal{V}(\mathcal{H})$. Thus $a \in S(x_a)$ implies $\lambda a \in S(x_a)$ for all $\lambda \in \mathbb{R}$, also implying $\lambda a \in S(A)$. By the same arguments, for all $\mu \in \mathbb{R}$, $b \in S(A)$, implies $\mu b \in S(A)$. Thus the one dimensional closed vector spaces $K_a = \{\lambda a : \lambda \in \mathbb{R}\}$ and $K_b = \{\mu b : \mu \in \mathbb{R}\}$ are subspaces in $S(A)$ i.e., $K_a \subseteq S(A)$ and $K_b \subseteq S(A)$. Thus $K_a + K_b \subseteq S(A) + S(A)$. By super-additive property of S, $S(A) + S(A) \subseteq S(A + A) = S(A)$ (because for vector space A, $A + A = A$). Also, $\lambda a + \mu b \in K_a + K_b \subseteq S(A)$, implying for all $a, b \in S(A)$, $\lambda, \mu \in \mathbb{R}$, $\lambda a + \mu b \in S(A)$.

$S(A)$ is also closed, as S is taken to be a closed subspace valued map. Thus we have shown that S being super-additive and closed implies for all $A \in \mathcal{V}(\mathcal{H})$, $S(A) \in \mathcal{V}(\mathcal{H})$. Thus the union extension can be written as $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$.

Next we show the reverse statement that a union extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ implies S is super-additive and closed.

For all $A, B \in \mathcal{V}(\mathcal{H})$, we have $A + B \in \mathcal{V}(\mathcal{H})$, as the sum of two closed vector spaces is a closed vector space. Also $A \subseteq A + B$ and $B \subseteq A + B$. Thus $S(A) = \cup_{x \in A} S(x) \subseteq \cup_{x \in A+B} S(x) = S(A+B)$. Similarly, $S(B) \subseteq S(A + B)$. Given S maps $\mathcal{V}(\mathcal{H})$ into $\mathcal{V}(\mathcal{H})$, we have for $A, B, A + B \in \mathcal{V}(\mathcal{H})$, $S(A), S(B), S(A + B) \in \mathcal{V}(\mathcal{H})$. Since $S(A) \subseteq S(A + B)$ and $S(B) \subseteq S(A + B)$, $S(A) + S(B) \subseteq S(A + B)$ implying S is super-additive. S being closed follows from the assumption that $S(A)$ was in $\mathcal{V}(\mathcal{H})$ which is a space of closed vector spaces. □

The notions of quasilinear and idempotent maps from [9] are related to the notion of super additivity by noting that for any quasilinear, idempotent S, $S_{sup}(A) := \sum_{w \in A} S(w)$ can be defined as the corresponding super additive map. Also the representers from [9] are of the form $\sum_{i=1}^{m} S(w_i)$ and thus equivalently can be written as $S_{sup}(\text{span}(\{w_1, \ldots, w_m\}))$. Thus considering a super-additive subspace valued map does not lead to any loss of generality. Furthermore [9] assumed the maps to be idempotent, i.e., $S(S(x)) = S(x)$, which implicitly assumes that S has a subspace valued union extension and thus all idempotent subspace valued maps are implicitly required to be super-additive.

Another property that is of interest for us is the preservation of $\mathcal{N}_L^{\perp} = \text{range}(L^*)$ for a given operator $L : \mathcal{H} \to \mathcal{Z}$ under a subspace valued map, i.e., we want $\text{range}(L^*) \subseteq S(\text{range}(L^*))$. S being inclusive is a sufficient condition for such a range preserving property. Formally we define this property as follows,

Definition 2.9. *(Range preserving map)*
*Let $L : \mathcal{H} \to \mathcal{Z}$ be a closable, densely defined operator as considered in Section 2.2.1 and let $\mathcal{N}_L^{\perp} = \text{range}\,(L^*)$ be the null space orthogonal of L. Then a subspace valued map $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ is called **range preserving** with respect to L if*

$$\mathcal{N}_L^{\perp} \subseteq S(\mathcal{N}_L^{\perp})$$

or equivalently, $S(\mathcal{N}_L^{\perp})^{\perp} \subseteq \mathcal{N}_L$.

Given that $\mathcal{N}_L^{\perp}$ and $\mathcal{N}_L$ are closed, orthogonal complementary subspaces in $\mathcal{H}$, the subspace valued extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ implies $S(\mathcal{N}_L^{\perp})$ and $S(\mathcal{N}_L^{\perp})^{\perp}$ are also closed, orthogonal complementary spaces in $\mathcal{H}$.

The range preserving property $S(\mathcal{N}_L^{\perp})^{\perp} \subseteq \mathcal{N}_L$ implies that any $g \in S(\mathcal{N}_L^{\perp})^{\perp}$, $g \in \mathcal{N}_L$, i.e., $Lg = 0$. This property will be useful later when proving the generalized theorem.

Lemma 2.3. *(Inclusive implies range preserving)*
If $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ is inclusive then it is range preserving with respect to any closable, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$.

Proof: *If S is inclusive, then for all $A \in \mathcal{V}(\mathcal{H})$, $A \subseteq S(A)$. For a closable, densely defined operator, the orthogonal to the null space $\mathcal{N}_L^{\perp} = \operatorname{range}(L^*)$ is a closed vector subspace in $\mathcal{V}(\mathcal{H})$ and thus inclusivity implies $\mathcal{N}_L^{\perp} = \operatorname{range}(L^*) \subseteq S(\mathcal{N}_L^{\perp})$.* □

The range preserving property and orthogonal complementary nature of $S(\mathcal{N}_L^{\perp})$ and $S(\mathcal{N}_L^{\perp})^{\perp}$ will be key in characterizing the conditions for the existence of a representer theorem.

2.2.3 Orthomonotone Functionals

[8, 9] introduced orthomonotone functionals as a way to expand the class of regularizers. The following reiterates the notions introduced there in the context of subspace valued maps of the form $S : \mathcal{V}(\mathcal{Z}) \to \mathcal{V}(\mathcal{Z})$ and separates out the notions of orthomonotonicity with respect to a single closed subspace (which gives a sufficient condition for the existence of a representer) and orthomonotonicity with respect to a subspace valued map, which gives as a necessary and sufficient condition when considering existence of representers for a family of minimization problems.

Definition 2.10. *(Orthomonotonicity with respect to a subspace)*
*Let $\mathcal{Z}$ be a Hilbert space and $\mathcal{K} \subseteq \mathcal{Z}$ be a closed subspace of $\mathcal{Z}$. Let $\mathcal{K}^{\perp}$ denote the orthogonal complementary space to $\mathcal{K}$. A functional $\Omega : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is called **orthomonotone** with respect to the subspace $\mathcal{K}$ if*

$$\forall f \in \mathcal{K}, g \in \mathcal{K}^{\perp}, \qquad \Omega(f+g) \geq \Omega(f)$$

Definition 2.11. *(Orthomonotonicity with respect to a subspace valued map)*
*Let $\mathcal{Z}$ be a Hilbert space. A functional $\Omega : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is called **orthomonotone** with respect to a subspace valued map $S : \mathcal{V}(\mathcal{Z}) \to \mathcal{V}(\mathcal{Z})$ if*

$$\forall A \in \mathcal{V}(\mathcal{Z}), f \in S(A), g \in S(A)^{\perp}, \qquad \Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$$

Consider the subspace valued map $S_{\mathbb{R}}$ from Example 2.4. [8, Theorem 1] showed that a functional Ω is orthomonotone with respect to $S_{\mathbb{R}}$ if and only if there exists a monotonically increasing functional $h : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $\Omega(z) = h(||z||), \forall z \in \mathcal{Z}$. Note that while the above characterization with a monotonically increasing functional restricts its analysis to inner product induced norms, other kinds of orthomonotone functionals can be constructed as well, and orthomonotonicity with respect to a subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ was introduced in [9] as a means to expands the class of regularizers to non inner product terms. Example 2.5 belows shows a few examples of orthomonotone regularizers.

Example 2.5. *Orthomonotone functionals*

1. *$\Omega(z) = ||z||_{\mathcal{Z}}^p$, for any $p > 0$ is orthomonotone w.r.t. $S_{\mathbb{R}}$*

2. *Let $\mathcal{Z} = \mathbb{R}^n$ and $||\cdot||_1$ denote the ℓ_1 norm. Then, $\Omega(z) = ||z||_1$ is orthomonotone w.r.t. S_{proj} (S_{proj} as defined in Example 2.4-3).*

The proof for the first statement follows directly from [8, Theorem 1] since $\Omega(z) = ||z||_{\mathcal{Z}}^p$, for any $p > 0$ is a monotonically increasing function of the inner product induced norm. The proof for the second statement follows from Theorem 2.1.

The second statement in the example above shows how sparse regularization problems involving the ℓ_1 norm are also covered by the notion of orthomonotone functionals.

The orthomonotonicity of ℓ_1 regularizers is formalized with the following theorem,

Theorem 2.1. *Orthomonotonicity of ℓ_1 regularizers*
Let $\mathcal{Z} = \mathbb{R}^n$, S_{proj} be the subspace valued map defined in Example 2.4 and let $h : [0, \infty] \to \mathbb{R} \cup \{+\infty\}$ be a monotonic increasing function. Then $\Omega(z) = h(||z||_1)$ is orthomonotone with respect to S_{proj}.

Proof: *We first show $\Omega(z) = ||z||_1$ is orthomonotone w.r.t. S_{proj}. The result for monotonic increasing h follows from there.*

Let $E = \{e_1, \ldots, e_n\}$ be the standard basis for $\mathbb{R}^n$. Note that for any $z \in \mathbb{R}^n$, $S_{proj}(z) = \{\sum_{i=1}^n \lambda_i \langle z, e_i\rangle_{\mathbb{R}^n} e_i : e_i \in E, \lambda_i \in \mathbb{R}\}$ and $(S_{proj}(z))^\perp = \{\sum_j \lambda_j e_j : \langle z, e_j\rangle_{\mathbb{R}^n} = 0, e_j \in E, \lambda_j \in \mathbb{R}\}$. Similarly for a set $A \subset \mathbb{R}^n$, $S_{proj}(A) = \{\sum_{i=1}^n \lambda_i \langle z, e_i\rangle_{\mathbb{R}^n} e_i : e_i \in E, \lambda_i \in \mathbb{R}, z \in A\}$ and $(S_{proj}(A))^\perp = \{\sum_j \lambda_j e_j : e_j \in E, \lambda_j \in \mathbb{R}, \forall z \in A, \langle z, e_j\rangle_{\mathbb{R}^n} = 0\}$. Now for any $z \in S_{proj}(A)$ and $c \in S_{proj}(A)^\perp$, $||z + c||_1 = \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} \neq 0\}} |z_i| + \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} = 0\}} |c_i|$ with $z_i = \langle z, e_i\rangle_{\mathbb{R}^n}$ and $c_i = \langle c, e_i\rangle_{\mathbb{R}^n}$. Also $||z||_1 = \sum_{i=1}^n |z_i| = \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} \neq 0\}} |z_i|$ and $||c||_1 = \sum_{i=1}^n |c_i| = \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} = 0\}} |c_i|$. Thus we see $||z + c||_1 = ||z||_1 + ||c||_1 \geq \max\{||z||_1, ||c||_1\} \implies \Omega(z) = ||z||_1$ is orthomonotone with respect to S_{proj}.

For any monotonically increasing function h, for any $a, b \in [0, \infty)$, $a > b$ implies $h(a) > h(b)$. Thus $||z+c||_1 \geq \max\{||z||_1, ||c||_1\}$ implies $h(||z+c||_1) \geq \max\{h(||z||_1), h(||c||_1)\}$. And thus $\Omega(z) = h(||z||_1)$ is orthomonotone with respect to S_{proj} for any monotonically increasing function h. □

The theorem can also be extended to a countable space of sequences as follows,

Theorem 2.2. *(Orthomonotonicity of ℓ_1 regularizers in countable spaces)*
Let $\mathcal{Z} = \ell^2(\mathbb{N})$ be the Hilbert space of $\mathbb{R}-$valued square summable sequences on $\mathbb{N}$. Let $||f||_1 = \begin{cases} \sum_{i=1}^\infty |f_i| & \text{if summation is bounded} \\ +\infty & \text{otherwise} \end{cases}$. Let S_{proj} be the subspace valued map considered in Example 2.4-4 and $h : [0, \infty] \to \mathbb{R} \cup \{\infty\}$ be a monotonic increasing function. Then $\Omega(f) = h(||f||_1)$ is orthomonotone with respect to S_{proj}.

Proof: *For any $A \in \mathcal{V}(\mathcal{Z})$, $f \in S_{proj}(A)$, $g \in S_{proj}(A)^\perp$, we have $f = \sum_{i \in K_A} \lambda_i \delta_i$ and $g = \sum_{j \in \mathbb{N}\setminus K_A} \lambda_j \delta_j$, for δ_i being the orthonormal basis of $\ell^2(\mathbb{N})$ considered in Example 2.4-4 and K_A being some subset of indices in $\mathbb{N}$ for which A has a non-zero projection on δ_i, written as $K_A = \{i \in \mathbb{N} :$ there exists some $a \in A$ such that $\langle a, \delta_i\rangle_{\ell^2} \neq 0\}$. Thus we have $||f + g||_1 = ||f||_1 + ||g||_1$ (including the case when any of them takes the value of ∞) as both f and g have disjoint supports.*

Thus we have $||f+g||_1 \geq \max\{||f||_1, ||g||_1\}$ *for all* $A \in \mathcal{V}(\mathcal{Z})$ *and* $f \in S_{proj}(A)$, $g \in S_{proj}(A)^{\perp}$. *Then for any monotonically increasing function* h, *we have* $h(||f+g||_1) \geq \max\{h(||f||_1), h(||g||_1)\}$ *and thus* Ω *is orthomonotone with respect to* S_{proj}. □

For more properties of orthomonotone functional regarding compositions and sums we refer the reader to [9]. With the notions of linear and adjoint operators combined with subspace valued maps and orthomonotone functionals, we are now ready to present the main result for the generalized representer theorem.

2.3 Generalized representer theorem

Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces. Let $L : \mathcal{H} \to \mathcal{Z}$ be closed, densely defined operators on $\mathcal{H}$. Let $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be some lower semi-continuous functionals.

Functionals of the form $C' : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_m \to \mathbb{R} \cup \{\infty\} := C'(L_1 f, \ldots, L_m f)$ are written without loss of generality in terms of a Hilbert space $\mathcal{Z}$ considered above, as follows. For any $m \in \mathbb{N}$ and $i \in \{1, \ldots, m\}$, let $L_i : \mathcal{H} \to \mathcal{Z}_i$ be closed, densely defined linear operators from $\mathcal{H}$ to separable Hilbert spaces $\mathcal{Z}_i$. Let $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_m$ and let $L : \mathcal{H} \to \mathcal{Z}$ be given by $Lf = (L_1 f, \ldots, L_m f)$, thus equivalently writing C' as a functional $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$.

Now, consider the optimization problem,

$$f_{opt} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \quad C(Lf) + \Omega(f) \tag{2.17}$$

The inclusion of $\{+\infty\}$ in the range of lower semi-continuous C and Ω allows one to consider constrained optimization problems. A few examples of learning problems written in this form are shown below,

Example 2.6. *(Learning and control problems)*

1. *Let* $\mathcal{H}$ *be an RKHS space of functions taking values in* $\mathcal{Z}_i = \mathbb{R}^n$. *Consider the evaluation operator from Example 2.1 such that* $L_x : \mathcal{H} \to \mathcal{Z}_i$ *is given by* $L_x f := f(x)$. *Let* $\{(x_i, y_i) : i = 1, \ldots, m\}$ *be a training data set. Let* $L_1, \ldots, L_m$ *be given by* $L_{x_1}, \ldots, L_{x_m}$ *and* $L' : \mathcal{H} \to \mathcal{H}$ *be the identity operator. Let* $C(L_1 f, \ldots, L_m f) := \sum_{i=1}^m ||y_i - \sigma(L_{x_i} f)||_{\mathcal{Z}}^2$ *for some activation function* $\sigma : \mathbb{R}^n \to \mathbb{R}^n$. *Let* $\Omega(L'f) := ||f||_{\mathcal{H}}^2$. *Then for* $J(f) = \sum_{i=1}^m ||y_i - \sigma(L_{x_i} f)||_{\mathcal{Z}_i}^2 + ||f||_{\mathcal{H}}^2$ *we get a regularized least squares problem in the RKHS space if* σ *is linear and an RKHS based neural network layer for some nonlinear* σ.

2. *Let* $\Omega(f) = ||f||_1^2$ *in the above example and we get a* ℓ_1 *regularized problem.*

3. *Let* $\mathcal{Z}_i = \mathbb{R}$, $y_i \in \{+1, -1\}$, $C(L_1 f, \ldots, L_m f) := \begin{cases} 0 & \forall i \in \{1, \ldots, m\}; \quad y_i L_i f > 0 \\ +\infty & \textit{otherwise} \end{cases}$ *and* $\Omega(f) = ||f||^2$. *Then* $J(f) = C(L_1 f, \ldots, L_m f) + \Omega(f)$ *gives the hard margin support vector machine objective for binary classification.*

4. *Let* μ *be a positive measure on the measurable space* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *Let* f, u *be functions in* $L^2([0,\infty), \mu; \mathbb{R}^n)$ *and* $L^2([0,\infty), \mu; \mathbb{R}^m)$ *respectively. Consider the regularizer* $\Omega(f, u) =$

$||f||^2_{L^2} + ||u||^2_{L^2}$ and $C(L(f,u)) = \begin{cases} 0 & \text{if}^{\,\partial_t f(t_i) - \phi(f(t_i),u(t_i))=0 \text{ for all } i=1,\ldots,m}_{\,f(0)=x_0,u(0)=u_0} \\ +\infty & \text{otherwise} \end{cases}$ for some known nonlinear function $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, finite set of points $\{t_i \in [0,\infty) : i = 1,\ldots,m\}$ and $(x_0,u_0) \in \mathbb{R}^n \times \mathbb{R}^m$. Then $J((f,u)) = C(L(f,u)) + \Omega(f,u)$ gives the objective function for solving a collocation based approximation to a continuous time nonlinear optimal control problem, where ϕ is a known function for the dynamics of the system, f denotes the continuous time trajectory and u denotes the continuous time control signal. The measure μ is used as a weighting measure to determine the growth rate of the functions considered in the hypothesis space for the solutions. Note also that the derivative operator ∂_t is only a closed, densely defined operator and not a bounded one.

Given a learning problem in the form of (2.17), let Ω be orthomonotone with respect to an inclusive, super-additive subspace valued map $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. The generalized representer theorem states that a minimizer for (2.17) exists in the subspace given by $S(\mathcal{N}_L^\perp)$ and the problem (2.17) is said to be linearly representable.

The notion of linear representability is significant as it often allows one to reformulate infinite dimensional optimization problems in $\mathcal{H}$ into equivalent finite dimensional optimization in $\mathcal{Z}$ given as

$$f_{opt} = \underset{f \in S(\text{range}(L^*))}{\arg\min} \quad C(Lf) + \Omega(f) \tag{2.18}$$

(2.18) gives a finite dimensional optimization if $\mathcal{Z}$ is finite dimensional and $S(\text{range}(L^*))$ is a finite dimensional subspace.

Below we state and prove, first the sufficient condition for linear representability of a functional $J(f) = C(Lf) + \Omega(f)$ and then the complete statement of necessary and sufficient condition for linear representability over a given family of functionals.

2.3.1 Sufficient conditions for linear representability

Theorem 2.3. *Generalized Representer Theorem (Sufficient condition)*
Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces and $L : \mathcal{H} \to \mathcal{Z}$ be a closed, densely defined linear operator with the null space orthogonal $\mathcal{N}_L^\perp = \text{range}(L^*)$. Let $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ be a closed and super additive subspace valued map, range preserving with respect to L. Let $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous functionals, with Ω orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$. Then for the problem,

$$f_{opt} = \underset{f \in \mathcal{H}}{\text{argmin}} \quad C(Lf) + \Omega(f) \tag{2.19}$$

if the minimizers are attainable, atleast one minimizer is linearly representable with respect to S, such that $f_{opt} \in S(\mathcal{N}_L^\perp)$.

Proof: *Since Ω is orthomonotone with respect to the closed subspace $S(\mathcal{N}_L^\perp)$, $\forall f \in S(\mathcal{N}_L^\perp), g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \Omega(f)$. Also by Definition 2.9, S is range preserving with respect to L, implies $S(\mathcal{N}_L^\perp)^\perp \subseteq \mathcal{N}_L$. Thus for all $g \in S(\mathcal{N}_L^\perp)^\perp$, $Lg = 0$.*

For a closed, densely defined operator L, $\mathcal{N}_L^\perp = \text{range}(L^*)$ is a closed vector subspace and thus by definition is mapped to a closed subspace $S(\mathcal{N}_L^\perp)$ by the subspace valued map. Thus $S(\mathcal{N}_L^\perp)$

and $S(\mathcal{N}_L^\perp)^\perp$ form an orthogonal complementary pair for $\mathcal{H}$ and for any $F \in \mathcal{H}$ we can find a decomposition $F = f + g$, with $f \in S(\mathcal{N}_L^\perp)$, $g \in S(\mathcal{N}_L^\perp)^\perp$. Then

$$\begin{aligned} J(F) &= C(L(f+g)) + \Omega(f+g) && (2.20)\\ &= C(Lf) + \Omega(f+g) && (2.21)\\ &\geq C(Lf) + \Omega(f) && (2.22)\end{aligned}$$

Thus $\forall F \in \mathcal{H}$, $\exists f \in S(\mathcal{N}_L^\perp)$ such that $J(f) \leq J(F)$. Thus if J admits a minimizer in $\mathcal{H}$, a minimizer must exists in $S(\mathcal{N}_L^\perp)$, implying J is linearly representable w.r.t. S. □

2.3.2 Necessary and sufficient conditions for linear representability

The generalized representer theorem we present here differs from its previous counterpart [9, Theorem 3.1] in three significant ways. Firstly, there is no assumption for a finite dimensional r-regularity property on the subspace valued map and secondly, the loss functional C can be defined on arbitrary infinite dimensional Hilbert spaces $\mathcal{Z}$. These two changes become significant since when dealing with stochastic regression problems the output space $\mathcal{Z}$ is an infinite dimensional Hilbert space of random variables (or measurable functions) and when dealing with ℓ_1 regularization problems in function spaces, the corresponding subspace valued map S_{proj} is not r-regular for any finite r. We will expand upon these differences in Section 2.4 with corresponding application examples. Lastly, we consider closed and densely defined operators in the loss function which allows for unbounded, derivative like operators in learning and control problems.

Now note that problems of the form (2.17) are typically considered over families of linear operators $L : \mathcal{H} \to \mathcal{Z}$ where L depends on training data for the learning problem and scaled regularizers $\{\gamma\Omega : \gamma \in (0,\infty)\}$, and if (2.17) is linearly representable for some choice of L and γ, it is natural to expect the problem to be linearly representable for all possible problems in this family. In fact if Ω is orthomonotone with respect to a closed, inclusive and super-additive subspace valued map S, this follows from Theorem 2.3 for all closed, densely defined linear operators (since an inclusive S is null space preserving for any operator L, by Lemma 2.3). The necessary condition in the representer theorem considers the reverse proposition, that is, if (2.17) is linearly representable with respect to a closed, inclusive and super-additive subspace valued map S for all closed, densely defined operators L and all $\gamma \in (0,\infty)$, then under certain additional assumptions on C and Ω it can be concluded that Ω must be orthomonotone with respect to S.

Thus, consider the family of functionals, given a closed, inclusive and super-additive subspace valued map S,

$$\mathcal{J}_S = \{C \circ L + \gamma\Omega \mid \gamma \in (0,\infty), L : \mathcal{H} \to \mathcal{Z} \text{ is closed, densely defined}\} \qquad (2.23)$$

and for fixed functionals $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ such that C admits a unique non-zero minimizer $z^\star$ in $\mathcal{Z}\backslash\{0\}$ with compact sub-level sets in its neighborhood and Ω admits a minimizer at 0. Note that the assumption on Ω is not a new one. Any Ω orthomonotone with respect to a subspace valued map must admit a minimizer at 0 and thus was not explicitly stated in Theorem 2.3. For the reverse proposition of the representer theorem, however Ω is not assumed to be orthomonotone and thus for it to be orthomonotone by the reverse proposition, a minimizer at 0 must be assumed (the minimizer at 0 need not be a unique minimizer).

The necessary and sufficient conditions for the generalized representer theorem can then be stated as follows,

Theorem 2.4. *Generalized Representer Theorem (Necessary and Sufficient Conditions)* *Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces. Let $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ be a closed, inclusive and super additive subspace valued map. Let $\Omega : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ and $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous functionals, such that Ω admits a minimizer at 0 and C admits a unique minimizer $z^\star$ in $\mathcal{Z} \backslash \{0\}$ with sequentially compact sub-level sets around $z^\star$. Let $\mathcal{J}_S = \{J_{L,\gamma} = C \circ L + \gamma\Omega \mid \gamma \in (0, \infty), L : \mathcal{H} \to \mathcal{Z}$ is a closed, densely defined linear operator$\}$ be the family of functionals corresponding to all closed, densely defined linear operators $L : \mathcal{H} \to \mathcal{Z}$ and constants $\gamma \in (0, \infty)$. For each functional in $J_{L,\gamma} \in \mathcal{J}_S$ consider the problem,*

$$f_{opt} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \quad J_{L,\gamma}(f) \tag{2.24}$$

Then, each problem in the family $\{\min_{f\in\mathcal{H}} J_{L,\gamma}(f) : J_{L,\gamma} \in \mathcal{J}_S\}$ is linearly representable with respect to S if and only if, Ω is orthomonotone with respect to S

Proof: *The proof for sufficiency (i.e. orthomonotone $\Omega \implies$ existence of linear representer) follows from Theorem 2.3 and Lemma 2.3.*

To prove necessity of orthomonotone Ω, assume that all functionals $J_{L,\gamma} \in \mathcal{J}_S$ corresponding to a linear operator L and constant γ are linear representable w.r.t. to S, i.e., for all functionals $J_{L,\gamma} = C \circ L + \gamma\Omega \in \mathcal{J}_S$ a minimizer exists in $S(\mathcal{N}_L^\perp)$. Note that a minimizer $J_{L,\gamma}$ exists because both C and Ω admit minimizers in $\mathcal{Z}\backslash\{0\}$ and $\mathcal{H}$ respectively and range(L) *is a closed subset in $\mathcal{Z}$.*

We first show that for all closed densely defined operators $L : \mathcal{H} \to \mathcal{Z}$ we must have $\Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$ for a family of functionals $\{J_{L,\gamma} \in \mathcal{J}_S : \gamma \in (0, \infty)\}$ to be linearly representable with respect to S. We show this in two parts, first we show $\Omega(f + g) \geq \Omega(f)$ and then $\Omega(f + g) \geq \Omega(g)$ for $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$.

Finally we show that there exists a one to one correspondence between the space of all closed vector subspaces $A \in \mathcal{V}(\mathcal{H})$ and a set of closed and bounded linear operators (which is a subset of closed, densely defined linear operators) and thus for all $A \in \mathcal{V}(\mathcal{H})$, we must have a closed, bounded operator $L : \mathcal{H} \to \mathcal{Z}$ such that $A = \mathcal{N}_L^\perp$. Thus for all $A \in \mathcal{V}(\mathcal{H})$, $f \in S(A)$ and $g \in S(A)^\perp$ we have $\Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$. Thus we show that if the family of functionals $\mathcal{J}_S = \{J_{L,\gamma} = C \circ L + \gamma\Omega \mid \gamma \in (0, \infty), L : \mathcal{H} \to \mathcal{Z}$ is a closed, densely defined linear operator$\}$ for a given tuple of functionals and subspace valued map (C, Ω, S) are all linearly representable with respect to S then Ω must be orthomonotone with respect to S.

We start by proving the result that $\Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$.

(i) Consider first the corner case for $f \in S(\mathcal{N}_L^\perp)$ such that $f = 0$, and $g \in S(\mathcal{N}_L^\perp)^\perp$. Then, we have $\Omega(f + g) = \Omega(g) \geq \Omega(g)$ (trivially true) and $\Omega(f + g) = \Omega(g) \geq \Omega(0) = \Omega(f)$ (since Ω admits a minimizer at 0 and $f = 0$). Thus for the case of $f \in S(\mathcal{N}_L^\perp)$, $f = 0$, we have shown that $\Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $g \in S(\mathcal{N}_L^\perp)^\perp$.

(ii) Next, consider the corner case, where the operator $L : \mathcal{H} \to \mathcal{Z}$ is such that $S(\mathcal{N}_L^\perp) = \{0\}$, i.e. there exists no $f \neq 0$ in $S(\mathcal{N}_L^\perp)$. If $S(\mathcal{N}_L^\perp) = \{0\}$, then result (i) implies that for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$.

(iii) Now for the general case where $S(\mathcal{N}_L^\perp) \neq \{0\}$, there exist a $f \neq 0$ in $S(\mathcal{N}_L^\perp)$. Let $z^\star \neq 0 \in \mathcal{Z}$ denote the unique minimizer for functional C. From result (i) we already have the result that for $f = 0$, $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $g \in S(\mathcal{N}_L^\perp)^\perp$. Thus, consider the case for $f \neq 0$. By Proposition 2.1 below, we have shown that for any closed, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$ for which $S(\mathcal{N}_L^\perp) \neq \{0\}$, given a $f \neq 0 \in S(\mathcal{N}_L^\perp)$, we have a closed and bounded linear operator $L'_f : \mathcal{H} \to \mathcal{Z}$, such that $L'_f f = z^\star$ and for any $g \in S(\mathcal{N}_L^\perp)^\perp$, $L'_f g = 0$. Since L' is closed and bounded we have the functional $J_{L'_f,\gamma} = C \circ L'_f + \gamma\Omega$ in $\mathcal{J}_S$. Let $h^\star_{f,\gamma} \in S(\mathcal{N}_L^\perp)$ be a minimizer for $J_{L'_f,\gamma}$.

Next, note that since $z^\star$ is a minimizer for C, we have $C(z^\star) \leq C(L'_f h^\star_{f,\gamma})$ and thus

$$C(z^\star) + \gamma\Omega(h^\star_{f,\gamma}) \leq C(L'_f h^\star_{f,\gamma}) + \gamma\Omega(h^\star_{f,\gamma}) = J_{L'_f,\gamma}(h^\star_{f,\gamma}) \tag{2.25}$$

Also $h^\star_{f,\gamma}$ is the minimizer for $J_{L'_f,\gamma}$ and thus

$$J_{L'_f,\gamma}(h^\star_{f,\gamma}) = C(L'_f h^\star_{f,\gamma}) + \gamma\Omega(h^\star_{f,\gamma}) \leq C(L'_f(f+g)) + \gamma\Omega(f+g)$$

for all $g \in S(\mathcal{N}_L)^\perp$.

But from Proposition 2.1, we have $L'_f(g) = 0$ and $L'_f f = z^\star$, implying $C(L'_f(f+g)) = C(z^\star)$, giving

$$C(L'_f h^\star_{f,\gamma}) + \gamma\Omega(h^\star_{f,\gamma}) \leq C(z^\star) + \gamma\Omega(f+g) \tag{2.26}$$

Thus we have the inequality

$$C(z^\star) + \gamma\Omega(h^\star_{f,\gamma}) \leq J_{L'_f,\gamma}(h^\star_{f,\gamma}) \leq C(z^\star) + \gamma\Omega(f+g) \tag{2.27}$$

for all $\gamma \in (0, \infty)$ or equivalently,

$$\Omega(h^\star_{f,\gamma}) \leq \Omega(f+g) \tag{2.28}$$

for all $\gamma \in (0, \infty)$, $g \in S(\mathcal{N}_L^\perp)^\perp$ and all minimizers $h^\star_{f,\gamma}$.

Also, from (2.25), we have the inequality $C(L'_f h^\star_{f,\gamma}) - C(z^\star) \geq 0$ and from (2.26) we have $C(L'_f h^\star_{f,\gamma}) - C(z^\star) \leq \gamma(\Omega(f+g) - \Omega(h^\star_{f,\gamma}))$. Thus we have an inequality

$$0 \leq C(L'_f h^\star_{f,\gamma}) - C(z^\star) \leq \gamma(\Omega(f+g) - \Omega(h^\star_{f,\gamma})) \tag{2.29}$$

for all $\gamma \in (0, \infty)$, $g \in S(\mathcal{N}_L^\perp)^\perp$ and minimizers $h^\star_{f,\gamma}$. For the case where $\Omega(f+g) = \infty$, $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ is trivially satisfied. When $\Omega(f+g) < \infty$, so is $\Omega(h^\star_{f,\gamma})$ (by (2.29)). Thus for the case of $\Omega(f+g) < \infty$, we have $\Omega(f+g) - \Omega(h^\star_{f,\gamma}) < \infty$ and thus by (2.29),

$$\gamma \to 0 \implies C(L'_f h^\star_{f,\gamma}) \to C(z^\star)$$

Since the sub-level sets around $C(z^\star)$, $V_\epsilon = \{z \in \mathcal{Z} : C(z) \leq C(z^\star) + \epsilon\}$ are given to be sequentially compact, and $z^\star$ is the unique minimizer, this implies $L'_f h^\star_{f,\gamma} \to z^\star$ as $\gamma \to 0$. But $f, h^\star_{f,\gamma} \in S(\mathcal{N}_L^\perp)$

and thus by Proposition 2.1, we have strong convergence $h^\star_{f,\gamma} \to f$.

Thus from (2.28), under the limit $\gamma \to 0$, we have $\Omega(f) \leq \Omega(f+g)$ for all $g \in S(\mathcal{N}_L^\perp)^\perp$. Since the above argument holds for all $f \neq 0$, $f \in S(\mathcal{N}_L^\perp)$ and we have the result from (i) for $f = 0$, we have for all $f \in S(\mathcal{N}_L^\perp)$ and all $g \in S(\mathcal{N}_L^\perp)^\perp$, the result that

$$\Omega(f+g) \geq \Omega(f)$$

(iv) To show the remaining inequality $\Omega(f+g) \geq \Omega(g)$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, similar arguments to result (i), (ii) and (iii) are required and are presented in the following.

(iv-i) Firstly note that for $g = 0$, $g \in S(\mathcal{N}_L^\perp)^\perp$ and for any $f \in S(\mathcal{N}_L^\perp)$, we have $\Omega(f+g) = \Omega(f) \geq \Omega(f)$ (trivially true) and $\Omega(f+g) \geq \Omega(0) = \Omega(g)$ (because Ω admits a minimizer at 0 and $g = 0$). Thus for $g = 0$, we have $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $f \in S(\mathcal{N}_L^\perp)$.

(iv-ii) Now consider the corner case, where $S(\mathcal{N}_L^\perp)^\perp = \{0\}$. In such a case, using result (iv-i), we have for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \Omega(g)$

(iv-iii) For the general case when $S(\mathcal{N}_L^\perp)^\perp \neq \{0\}$, there exists a $g \in S(\mathcal{N}_L^\perp)^\perp$ such that $g \neq 0$. For $g = 0$, we already have the required inequality from (iv-i). Thus we consider the case for $g \in S(\mathcal{N}_L^\perp)^\perp$ and $g \neq 0$. From Proposition 2.1, using the $A = S(\mathcal{N}_L^\perp)^\perp$, we have a closed and bounded operator $L'_g : \mathcal{H} \to \mathcal{Z}$ such that $L'_g g = z^\star$ and for all $f \in S(\mathcal{N}_L^\perp)$, $L'_g f = 0$. Thus we have the functional $J_{L'_g,\gamma} = C \circ L'_g + \gamma\Omega$ in $\mathcal{J}_S$. Let $h^\star_{g,\gamma}$ be a minimizer for $J_{L'_g,\gamma}$. Then following the same arguments as before from (iii), we have the analogous inequality

$$\Omega(h^\star_{g,\gamma}) \leq \Omega(f+g) \tag{2.30}$$

for all $f \in S(\mathcal{N}_L^\perp)$, $\gamma \in (0,\infty)$ and minimizers $h^\star_{g,\gamma}$.

As $\gamma \to 0$, we have as before, a sequence of minimizers $h^\star_{g,\gamma} \to g$ and thus in the limit, we have $\Omega(f+g) \geq \Omega(g)$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $g \neq 0$. Combining with the result from (iv-i) for $g = 0$, we have for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$,

$$\Omega(f+g) \geq \Omega(g)$$

(v) Thus from (iii) and (iv), we have shown that for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, we have,

$$\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$$

for all closed, densely defined operators $L : \mathcal{H} \to \mathcal{Z}$.

(vi) Finally, we show that there is a one to one correspondence between the set of closed vector spaces $A \in \mathcal{V}(\mathcal{H})$ and a set of closed, bounded linear operators $L : \mathcal{H} \to \mathcal{Z}$, such that for any $A \in \mathcal{V}(\mathcal{H})$ there exists a closed bounded operator L satisfying $A = \mathcal{N}_L^\perp$. Using this correspondence and the result from (v), we have the final result stating that for all closed vector subspaces $A \in \mathcal{V}(\mathcal{H})$, and for all $f \in S(A)$ and $g \in S(A)^\perp$,

$$\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$$

implying Ω is orthomonotone with respect to S.

To show the correspondence between A and L consider the following.

For any closed vector subspace $A \in \mathcal{V}(\mathcal{H})$, let $P_A : \mathcal{H} \to \mathcal{H}$ denote the orthogonal projection onto the closed vector subspace A. Since A is a closed vector subspace of $\mathcal{H}$, A and $A^\perp$ form an orthogonal complementary pair of subspaces such that $\text{range}(P_A) = A$ *and null space of P_A is $A^\perp$, and thus by the closed graph theorem, it follows that P_A is a closed operator. Since for any $F \in \mathcal{H}$, there exists an unique decomposition $F = f + g$ such that $f \in A$ and $g \in A^\perp$ and $P_A F = P_A(f + g) = f$, it also follows that $||P_A F||_\mathcal{H} = ||f||_\mathcal{H} \leq ||F||_\mathcal{H}$ and P_A is thus a closed, bounded linear operator. It is also easy to see that $\langle P_A F_1, F_2\rangle_\mathcal{H} = \langle F_1, P_A F_2\rangle_\mathcal{H}$ for all $F_1, F_2 \in \mathcal{H}$ and thus P_A is a self-adjoint, closed, bounded operator.*

Let $L : \mathcal{H} \to \mathcal{Z}$ be any closed, bounded linear operator with null space $\mathcal{N}_L = \{0\}$. Then the composition $L'_A = L \circ P_A$ is also closed and bounded, and $\mathcal{N}^\perp_{L'_A} = \text{range}((L'_A)^*) = \text{range}(P_A L^*) = A$. *Thus for every $A \in \mathcal{V}(\mathcal{H})$, we have a closed and bounded operator given by L'_A such that $\mathcal{N}^\perp_{L'_A} = A$. The result for orthomonotonicity of Ω then follows, as stated above.*

□

To prove the necessary part of theorem 2.4, the following proposition is considered.

Proposition 2.1. *Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces. Let there exist a minimizer $z^\star \neq 0 \in \mathcal{Z}$ for the lower semicontinuous functional $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$. Let $A \in \mathcal{V}(\mathcal{H})$ be a closed vector subspace of $\mathcal{H}$. Let there exist a $f \in A$ such that $f \neq 0$. Let $\mathcal{H}$ be spanned by a orthonormal basis $\{f/||f||, \phi_1, \phi_2, \ldots\}$, let $\mathbb{N}_A$ be a subset of $\mathbb{N}$ such that A is spanned by $\{f/||f||\} \cup \{\phi_k : k \in \mathbb{N}_A\}$ and $A^\perp$ is spanned by $\{\phi'_k : k' \in \mathbb{N}\backslash\mathbb{N}_A\}$. Then there exists a closed and bounded linear operator $L'_f : \mathcal{H} \to \mathcal{Z}$ given by*

$$L'_f h = z^\star \left\langle \sum_{k\in\mathbb{N}_A} \frac{\phi_k}{k^2} + \frac{f}{||f||^2}, h \right\rangle_\mathcal{H}$$

such that

1. $L'_f g = 0$ *for all* $g \in A^\perp$
2. $L'_f f = z^\star$
3. $h \in S(A)$ *and* $L'_f h = z^\star$, *implies* $h = f$

Proof: *Firstly, note that the existence of a countable basis $\{f/||f||, \phi_1, \phi_2, \cdots\}$ is guaranteed by E. Schmidt's orthogonalization [24, Chapter III-5] for a separable Hilbert space. Since A and $A^\perp$ are orthogonal complementary subspaces and $f \in A$, they split the orthonormal basis into two disjoint countable subset as mentioned in the statement of the proposition given by the index set $\mathbb{N}_A$.*

To see that L'_f is bounded, note that for any $h \in \mathcal{H}$, $||L'_f h||_\mathcal{Z} = ||z^\star||_\mathcal{Z} |\langle \sum_{k\in\mathbb{N}_A} \phi_k/k^2 + f/||f||^2, h\rangle_\mathcal{H}|$. Then note that $|\langle \sum_{k\in\mathbb{N}_A} \phi_k/k^2 + f/||f||^2, h\rangle_\mathcal{H}| \leq \sum_{k\in\mathbb{N}_A} |\langle \phi_k/k^2, h\rangle_\mathcal{H}| + |\langle f/||f||^2, h\rangle_\mathcal{H}| \leq (\sum_{k\in\mathbb{N}_A} 1/k^2 + 1/||f||)||h||_\mathcal{H}$. Now since $\sum_{k\in\mathbb{N}_A} 1/k^2 \leq \sum_{k\in\mathbb{N}} 1/k^2 = \pi^2/6 < \infty$ (since summation of a series $1/k^2$ over $\mathbb{N}$ is known to be bounded), $0 < ||f||_\mathcal{H} < \infty$ and $||z^\star||_\mathcal{Z} < \infty$, we have $||L'_f h||_\mathcal{Z} \leq M||h||_\mathcal{H}$ for some bounded constant $M = ||z^\star||(\sum_{k\in\mathbb{N}_A} 1/k^2 + 1/||f||) < \infty$.

Also since the null space of L'_f denoted $\ker(L'_f)$ is $A^\perp$, $\inf\{||L'_f h||_\mathcal{Z} : h \in \ker(L'_f)^\perp, ||h||_\mathcal{H} = 1\} > 0$ and thus L'_f is closed by [17, Proposition 6.5.5].

Then for any $g \in S(A)^{\perp}$, we have $L'_f g = z^{\star}\langle \sum_{k\in\mathbb{N}_A} \phi_k/k^2 f/||f||^2, g\rangle_{\mathcal{H}} = 0$, showing the first property stated for L'_f.

The second statement $L'_f f = z^{\star}$, follows by substituting f into the definition for $L'_f f$. Since $\{f/||f||, \phi_1, \phi_2 \ldots\}$ are orthonormal basis, f is orthogonal to all ϕ_k and thus $\langle \phi_k, f\rangle_{\mathcal{H}} = 0$, which leaves the term $z^{\star}\langle f/||f||^2, f\rangle_{\mathcal{H}} = z^{\star}$.

The last statement can be seen from the fact that $L'_f h = z^{\star}$ implies $L'_f h = L'_f f$ or $L'_f(h-f) = 0$, i.e., $\langle \sum_{k\in\mathbb{N}_A} \phi_k + f, f - h\rangle_{\mathcal{H}} = 0$ implying $f - h \in S(A)^{\perp}$ (since ϕ_k and f span $S(A)$). But both f and h are given to be in $S(A)$ and thus they must be in $S(A) \cap S(A)^{\perp} = \{0\}$. Thus $h = f$.

□

2.3.3 Related work

We presented here a generalized version of representer theorems for problems of the form

$$f_{opt} = \underset{f\in\mathcal{H}}{\arg\min}\ C(Lf) + \Omega(f) \tag{2.31}$$

for a loss function $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ on a separable Hilbert space $\mathcal{Z}$ and closed, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$ and Ω orthomonotone with respect to a subspace valued map S. The assumption of "r-regularity" on subspace valued maps from previous counterparts of the theorem was dropped to allow for more general regularization like the ℓ_1 norm on function spaces, $\mathcal{Z}$ was considered as separable Hilbert spaces to allow for loss functional on infinite dimensional Hilbert space, as occurring in examples from learning in Hilbert spaces of stochastic processes and the linear operators were considered to be closed and densely defined to allow for unbounded operators like the derivative operators that occur commonly in optimal control problems.

Special cases of the theorem addressing learning with bounded functionals like the least squares regularization for vector valued functions in Reproducing Kernel Hilbert Space (RKHS) framework can be found in [18, Theorems 3.1, 4.1]. Special cases of the theorem for ℓ_1 regularization can be found in [5]. A generalized version of the representer theorems for general loss functions but still restricted to Hilbert spaces of real valued functions and bounded functionals can be found in [8, 6]. The far more general framework of subspace valued maps was introduced in [9, Theorem 3.1] and a variant of the presented theorem with an assumption of $r-$regularity, for bounded linear functionals and with the loss functional C on $\mathcal{Z} = \mathbb{R}^m$ can be found there. Representer theorems for problems with general constraints involving differential operators in the functional C and squared norm regularizers in Ω are presented in [29].

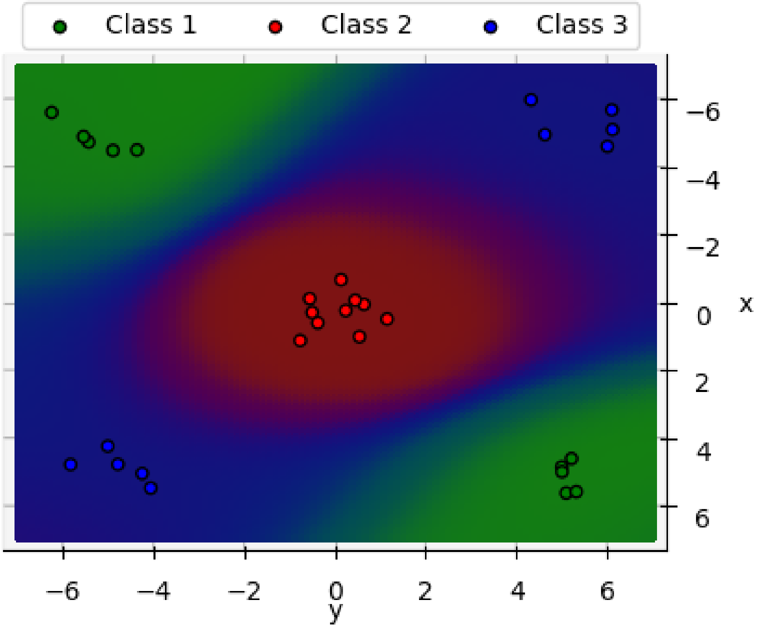


Figure 2.1: Multi class classification with a 3 layer, squared exponential kernel based neural network. Class probabilities shaded as red, blue, green values. Training data shown as point clusters.

2.4 Application examples

2.4.1 Deep neural networks

Motivation

Consider, first, a single layer perceptron with a nonlinear activation function σ, with input x and output y. Given m training samples $\{(x_i, y_i) : i \in \mathbb{N}_m\}$ consider the variational learning problem

$$\min_{f\in\mathcal{H}} \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i} f)||^2_{\mathcal{Z}} + \lambda||f||^2_{\mathcal{H}} \tag{2.32}$$

Let $\mathcal{Z} = \mathbb{R}^n$, $\mathcal{H}$ be an RKHS space with kernel K and $L_{x_i} : \mathcal{H} \to \mathcal{Z}$ be a closed bounded linear evaluation operator $L_{x_i} f = f(x_i)$ on the RKHS space. This minimization problem fits exactly the form of (2.17) by taking $C(L_{x_1} f, \ldots, L_{x_m} f) = \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i}(\cdot))||^2$ and Ω to be $||f||^2_{\mathcal{H}}$. Since Ω is orthomonotone with respect to $S_{\mathbb{R}}$, we know a minimizer of the form $\sum_{i=1}^{m} L^*_{x_i} z_i$ must exist. Substituting this form into the minimization above we can get a finite dimensional minimization problem.

$$\min_{z_j\in\mathcal{Z}} \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i} \sum_{j=1}^{m} L^*_{x_j} z_j)||^2_{\mathcal{Z}} + \lambda||\sum_{j=1}^{m} L^*_{x_j} z_j||^2_{\mathcal{H}} \tag{2.33}$$

On the RKHS $\mathcal{H}$, the adjoint L^*_x is known to be the kernel section $K(\cdot, x)$ (see Example 2.1) and $L_{x_i} L^*_{x_j} = K(x_i, x_j)$. Thus we have a nonlinear program to solve for a kernel based single layer perceptron with $z_i \in \mathcal{Z}$ being the new decision variables. Note that the program becomes nonlinear due to a nonlinear activation function σ and only thus differs from a generalized least squares setting.

So far we see nothing new as the problem is simply a least squares like problems in the RKHS space with finite dimensional outputs. Such problems can easily be covered by representer theorems from [9].

Now consider a N-layer concatenation of such perceptrons. Let the inputs for the first layer be denoted as $y^{(0)} = (y_1^{(0)}, \dots, y_m^{(0)}) \in \mathbb{R}^{n_0 \times m}$ taking values $y_i^{(0)} = X_i$ from a training data set $\mathcal{D} = \{(X_i, Y_i) \in \mathbb{R}^{n_0} \times \mathbb{R}^{n_N} : i = 1, \dots, m\}$. Let the function $f^{(1)}$ for the first layer be learned from an RKHS space $\mathcal{H}^{(1)}$ of $\mathbb{R}^{n_1}$-valued functions and let the output for the first layer be the unknown latent variables $y^{(1)} = (y_1^{(1)}, \dots, y_m^{(1)}) \in \mathbb{R}^{n_1 \times m}$. Let $\mathcal{Z}^{(1)}$ denote the separable Hilbert space $\mathbb{R}^{n_1 \times m} \times \mathcal{H}^{(1)}$ for notational convenience. The learning of the function $f^{(1)}$ can thus be considered as the variational problem,

$$y_{opt}^{(1)}, f_{opt}^{(1)} = \underset{(y^{(1)}, f^{(1)}) \in \mathcal{Z}^{(1)}}{\arg\min} \; C_1(L^{(1)}(y^{(1)}, f^{(1)})) + \Omega_1((y^{(1)}, f^{(1)})) \tag{2.34}$$

with $L^{(1)}(y^{(1)}, f^{(1)}) = (y_1^{(1)} - L_{y_1^{(0)}} f^{(1)}, \dots, y_m^{(1)} - L_{y_m^{(0)}} f^{(1)})$ being the bounded linear operator $L^{(1)} : \mathcal{Z}^{(1)} \to \mathbb{R}^{n_1 \times m}$, $C_1 : \mathbb{R}^{n_1 \times m} \to \mathbb{R} \cup \{\infty\}$ being the loss functional such that $C_1(L^{(1)}(y^{(1)}, f^{(1)})) = \begin{cases} 0 & \text{, if } y^{(1)} = L_{y^{(0)}} f^{(1)} \\ \infty & \text{, otherwise} \end{cases}$ and $\Omega_1((y^{(1)}, f^{(1)})) = ||(f^{(1)})||^2_{\mathcal{H}^{(1)}}$ being the regularizer. Again, nothing new so far, we have a Hilbert search space $\mathcal{Z}^{(1)}$ and a finite dimensional domain for the loss functional, $\mathbb{R}^{n_1 \times m}$. Also, note that this variational problem is ill posed since only the input data is fixed and the output data is left free and thus the minimizer for the above problem is at $y_{opt}^{(1)} = 0$ and $f_{opt}^{(1)} = 0$. We ignore the ill-posed nature of the optimization for now, as additional concatenated layers connecting to the final output data will force the minimizer to become non trivial.

Consider next the second layer for the network. Let $y^{(2)} \in \mathbb{R}^{n_2 \times m}$ be the latent variables, $\mathcal{H}^{(2)}$ be a an RKHS space of $\mathbb{R}^{n_2}$-valued functions and $f^{(2)} \in \mathcal{H}^{(2)}$ be the learned function for this layer. Let $\mathcal{Z}^{(2)}$ denote the Hilbert space $\mathbb{R}^{n_2 \times m} \times \mathcal{H}^{(2)}$. The learning problem for the second layer can then be posed as,

$$y_{opt}^{(1)}, y_{opt}^{(2)}, f_{opt}^{(2)} = \underset{y^{(1)} \in \mathbb{R}^{n_1 \times m}, (y^{(2)}, f^{(2)}) \in \mathcal{Z}^{(2)}}{\arg\min} \; C_2((y^{(1)}, y^{(2)}, f^{(2)})) + \Omega_2((y^{(2)}, f^{(2)})) \tag{2.35}$$

with $C_2((y^{(1)}, y^{(2)}, f^{(2)})) = \begin{cases} 0 & \text{, if } y^{(2)} = L_{y^{(1)}} f^{(2)} \\ \infty & \text{, otherwise} \end{cases}$ and $\Omega_2((y^{(2)}, f^{(2)})) = ||f^{(2)}||^2_{\mathcal{H}^{(2)}}$.

This is where we see a significant difference from the standard least squares like problem for the first time. Here $y^{(1)}$ being an unknown latent variable, is considered as a decision variable for the problem and thus $L_{y^{(1)}}$ is not a linear operator on the search space $\mathbb{R}^{n_1 \times m} \times \mathcal{Z}^{(2)}$. Thus unlike the first layer we cannot write the loss functional for the second layer as $C_2(L(y^{(1)}, y^{(2)}, f^{(2)}))$ for some linear operator $L : \mathbb{R}^{n_1 \times m} \times \mathcal{Z}^{(2)} \to \mathbb{R}^{n_2 \times m}$. The operator $L_{y^{(1)}}$ makes the operator $L(y^{(1)}, y^{(2)}, f^{(2)}) = y^{(2)} - L_{y^{(1)}} f^{(2)}$ a non-linear operator. Instead we consider a non-linear loss functional $C : \mathbb{R}^{n_1\times m} \times \mathbb{R}^{n_2\times m} \times \mathcal{H}^{(2)} \to \mathbb{R} \cup \{\infty\}$ as given in (2.35).

The problem for learning the first and second layer together can then be written as

$$\begin{aligned} y^{(1)}_{opt}, f^{(1)}_{opt}, y^{(2)}_{opt}, f^{(2)}_{opt} = & \underset{(y^{(1)},f^{(1)})\in\mathcal{Z}^{(1)},(y^{(2)},f^{(2)})\in\mathcal{Z}^{(2)}}{\arg\min} C_1(L^{(1)}(y^{(1)}, f^{(1)})) \\ & + C_2(y^{(1)}, y^{(2)}, f^{(2)}) + \Omega_1((y^{(1)}, f^{(1)})) + \Omega_2((y^{(2)}, f^{(2)})) \end{aligned} \tag{2.36}$$

Also note that we did not use any activation functions σ in the construction above. This was done to show clearly that the nonlinearity of the operation $L_{y^{(1)}} f^{(2)}$ present in C_2 has nothing to do with the activation function. Even with a simple interpolation or least squares like loss function we have to treat C_2 as a nonlinear functional on the Hilbert space $\mathbb{R}^{n_1\times m} \times \mathcal{Z}^{(2)}$. Having shown that C_2 is a nonlinear functional on $\mathbb{R}^{n_1\times m} \times \mathcal{Z}^{(2)}$ in any case, we can reintroduce the activation function and write $C^{(2)} : \mathbb{R}^{n_1\times m} \times \mathcal{Z}^{(2)} \to \mathbb{R} \cup \{\infty\}$ as the functional

$$C_2((y^{(1)}, y^{(2)}, f^{(2)})) = \begin{cases} 0 & , \text{ if } y^{(2)} = \sigma(L_{y^{(1)}} f^{(2)}) \\ \infty & , \text{otherwise} \end{cases} \tag{2.37}$$

For the functional C_1, reintroducing σ makes the operator $L^{(1)} : \mathbb{R}^{n_1\times m} \times \mathcal{H}^{(1)} \to \mathbb{R}^{n_1\times m}$ defined above, nonlinear. We can instead view the operator $L^{(1)}$ as the linear operator $L^{(1)} : \mathbb{R}^{n_1\times m} \times \mathcal{H}^{(1)} \to \mathbb{R}^{n_1\times m} \times \mathbb{R}^{n_1\times m}$ given as the mapping

$$L^{(1)}(y^{(1)}, f^{(1)}) = (y^{(1)}, L_{y^{(0)}} f^{(1)})$$

and C_1 as a corresponding nonlinear functional on $\mathbb{R}^{n_1\times m} \times \mathbb{R}^{n_1\times m}$. Thus we can view C_1 as the functional $C_1 : \mathbb{R}^{n_1\times m} \times \mathbb{R}^{n_1\times m} \to \mathbb{R} \cup \{\infty\}$, given as,

$$C_1(L^{(1)}(y^{(1)}, f^{(1)})) = \begin{cases} 0 & , \text{ if } y^{(1)} = \sigma(L_{y^{(0)}} f^{(1)}) \\ \infty & , \text{otherwise} \end{cases} \tag{2.38}$$

A similar construction can be done for each layer upto the $(N-1)^{th}$ layer. Note also that, while we introduced $\mathcal{H}^{(l)}$ as a RKHS space and $L_{y^{(l-1)}}$ as the linear evaluation operator evaluating functions at the point $y^{(l-1)}$, the same construction remains valid for any separable Hilbert space $\mathcal{H}^{(l)}$ and any closed, densely defined linear operator $L_{y^{(l-1)}}$, where the subscript $y^{(l-1)}$ denotes that the operators action depends on the output of the previous layer. The following describes the construction of the full N-layer neural network.

Formal construction

Let $y^{(l)} \in \mathbb{R}^{n_l\times m}$ be the latent output variable for each layer $l = 1, \dots, N-1$. Let $y^{(0)} = (X_1, \dots, X_m)$ and $y^{(N)} = (Y_1, \dots, Y_m)$ be the known input and output data respectively, used for training the network. Let $f^{(l)}$ denote the function learned for the l^{th} layer from a separable Hilbert space $\mathcal{H}^{(l)}$ of $\mathbb{R}^{n_l}-$valued functions. Let $\mathcal{O}_{\mathcal{H}^{(l)},\mathbb{R}^{n_l\times m}}$ be a set of closed, densely defined operators from $\mathcal{H}^{(l)}$ to $\mathbb{R}^{n_l\times m}$. Let $\phi_l : \mathbb{R}^{n_{l-1}\times m} \to \mathcal{O}_{\mathcal{H}^{(l)},\mathbb{R}^{n_l\times m}}$ be known functions mapping the output, $y^{(l-1)}$, of the $(l-1)^{th}$ layer to some closed, densely defined operator, $L_{y^{(l-1)}} \in \mathcal{O}_{\mathcal{H}^{(l)},\mathbb{R}^{n_l\times m}}$, denoted as $L_{y^{(l-1)}} = \phi_l(y^{(l-1)})$. Let $L^*_{y^{(l-1)}} : \mathbb{R}^{n_l \times m} \to \mathcal{H}^{(l)}$ denote the adjoint to $L_{y^{(l-1)}}$ and ϕ^*_l denote the map $\phi^*_l(y^{(l-1)}) = L^*_{y^{(l-1)}}$. An example for $\mathcal{O}_{\mathcal{H}^{(l)},\mathbb{R}^{n_l \times m}}$ is the set of all evaluation operators on an RKHS space and the function ϕ maps $y^{(l-1)}$ to the linear operator evaluating a function in the RKHS space at $y^{(l-1)}$. Another example for $\mathcal{O}_{\mathcal{H}^{(l)},\mathbb{R}^{n_l \times m}}$ is the set of gradient operators ∇_x computing the gradient of a function in $\mathcal{H}^{(l)}$ at a point $x \in \mathbb{R}^{n_{l-1} \times m}$ with $\phi(y^{(l-1)}) = \nabla_{y^{(l-1)}}$.

For notational convenience, let $z^{(l)} = (y^{(l)}, f^{(l)})$ and $\mathcal{Z}^{(l)} = \mathbb{R}^{n_l \times m} \times \mathcal{H}^{(l)}$.

Let

$$C_l(y^{(l-1)}, z^{(l)}) = \begin{cases} 0 & y^{(l)} = \sigma_l(\phi_l(y^{(l-1)})f^{(l)}) \\ \infty & \text{otherwise} \end{cases} \quad \text{for } l = 1, \dots, N-1 \tag{2.39}$$

be the lower semi-continuous functional $C_l : \mathbb{R}^{n_l \times m} \times \mathcal{Z}^{(l)} \to \mathbb{R} \cup \{\infty\}$, with $\sigma_l : \mathbb{R} \to \mathbb{R}$ being a lower semi-continuous function, interpreted as acting on each component for a matrix in $\mathbb{R}^{n_l \times m}$. Let,

$$\Omega_l(z^{(l)}) = ||f^{(l)}||^2_{\mathcal{H}^{(l)}} \text{ for } l = 1, \dots, N$$

be the regularizer $\Omega_l : \mathcal{Z}^{(l)} \to \mathbb{R} \cup \{\infty\}$.

For the final N^{th} layer, let $y^{(N)} = (Y_1, \dots, Y_m) \in \mathbb{R}^{n_N \times m}$ be a known output vector. Let the loss functional $C_N : \mathbb{R}^{n_{N-1} \times m} \times \mathcal{H}^{(N)} \to \mathbb{R} \cup \{\infty\}$ be given as

$$C_N(y^{(N-1)}, f^{(N)}) = ||y^{(N)} - \sigma_N(L_{y^{(N-1)}} f^{(N)})||^2_{\mathbb{R}^{n_N \times m}}$$

Given a training data set $\mathcal{D} = \{(X_i, Y_i) : i = 1, \dots, m\}$ of input-output pairs, we can write the full N-layer neural network learning problem as

$$\begin{aligned} z^{(1)}_{opt}, \dots, z^{(N-1)}_{opt}, f^{(N)}_{opt} = \underset{\substack{z^{(1)},\dots,z^{(N-1)},f^{(N)} \\ \in \mathcal{Z}^{(1)} \times \dots \times \mathcal{Z}^{(N-1)} \times \mathcal{H}^{(N)}}}{\arg\min} \quad & C_N(y^{(N-1)}, f^{(N)}) + \sum_{l=1}^{N-1} C_l(y^{(l-1)}, z^{(l)}) \\ & + \sum_{l=1}^{N} \Omega_l(z^{(l)}) \end{aligned} \tag{2.40}$$

Applying the generalized representer theorem to the neural network

(2.40) written in the standard form for the representer theorem,

$$F_{opt} = \underset{F \in \mathcal{H}}{\arg\min} \quad C(LF) + \Omega(F) \tag{2.41}$$

is a problem considered on the Hilbert space $\mathcal{H} = \mathcal{Z}^{(1)} \times \dots \times \mathcal{Z}^{(N-1)} \times \mathcal{H}^{(N)}$. Let $F \in \mathcal{H}$, be the concatenated vector $F = (z^{(1)}, \dots, z^{(N-1)}, f^{(N)})$. The operator $L : \mathcal{H} \to \mathbb{R}^{n_1 \times m} \times \mathbb{R}^{n_1 \times m}$ be a closed, densely defined operator, given by the oblique projection $L(F) = (y^{(1)}, L_{y_0} f^{(1)})$. Given the adjoint operator $L^*_{y_0} : \mathbb{R}^{n_1 \times m} \to \mathcal{H}^{(1)}$, we can write the adjoint $L^* : \mathbb{R}^{n_1 \times m} \times \mathbb{R}^{n_1 \times m} \to \mathcal{H}$ as $L^*(y, c) = ((y, L^*_{y_0} c), 0, 0, \dots, 0)$. Thus L is an operator $L : \mathcal{H} \to (\mathbb{R}^{n_1 \times m} \times \mathcal{H}^{(1)})$ with the null space orthogonal

$$\mathcal{N}^{\perp}_L = \mathbb{R}^{n_l \times m} \times \text{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\} \tag{2.42}$$

with the $\{0\}$ sets corresponding to $\mathcal{Z}^{(2)} \times \mathcal{Z}^{(3)} \times \ldots \mathcal{Z}^{(N-1)} \times \mathcal{H}^{(N)}$. The functional $C(LF) = C_1(y^{(0)}, LF)$ with C_1 as defined by (2.39) and

$$\Omega(F) = \Omega_1(z^{(1)}) + \sum_{l=2}^{N} (C_l(y^{(l-1)}, z^{(l)}) + \Omega_l(z^{(l)})) \tag{2.43}$$

Let $S_{\mathbb{R}}$ be the inclusive, closed, super-additive subspace valued map $S_{\mathbb{R}}(a) = \{\lambda a : a \in \mathbb{R}\}$, considered in Example 1.

Then, consider the subspace valued maps,

$$S_1(z^{(1)}) = S_{\mathbb{R}}(y^{(1)}) \times S_{\mathbb{R}}(\text{range}(L^*_{y^{(0)}})) \tag{2.44}$$

For $l = 1, \ldots, N-1$, let $\mathcal{Y}_l \subseteq \mathbb{R}^{n_l \times m}$ be a Borel measurable subset of $\mathbb{R}^{n_l \times m}$ given by the range of the function σ_l, i.e., $\mathcal{Y}_l = \{\sigma_l(y) : y \in \mathbb{R}^{n_l \times m}\} \subseteq \mathbb{R}^{n \times m}$. Let $\mathcal{B}(\mathcal{Y}_l)$ be the Borel $\sigma-$algebra on $\mathcal{Y}_l$ (inherited from the Borel $\sigma-$algebra on $\mathbb{R}^{n_l \times m}$).

For $l = 1, \ldots, N-1$, recall that $C_l(z^{(l)})$, forces $y^{(l)} = \sigma(L_{y^{(l-1)}} f^{(l)})$ for a non-infinite cost. Then, the range of values for $y^{(l)}$, $\mathcal{Y}_l$ restricts the possible input values for ϕ_{l+1} and shrinks the solution space in which a minimizer may lie. For measurable, bounded variation functions ϕ^*_l, we can exploit this fact by considering the following subspace valued map over the $\mathcal{Z}^{(l)}$, for $l = 2, \ldots, N-1$,

$$S_l(z^{(l)}) = S_{\mathbb{R}}(y^{(l)}) \times \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi^*_l(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right) \tag{2.45}$$

where $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ is the Banach space of signed, $\mathbb{R}^{n_l \times m}-$valued Borel measures with bounded total variation (see [30]) on the measurable space $(\mathcal{Y}_l, \mathcal{B}(\mathcal{Y}_l))$.

Lemma 2.4 shows that $S_l : \mathcal{Z}^{(l)} \to \mathcal{V}(\mathcal{Z}^{(l)})$ defined in (2.45) under a certain regularity assumptions for the map ϕ_l and ϕ^*_l over the domain $\mathcal{Y}_{l-1}$, is a closed and super-additive subspace valued map.

Lemma 2.4. *(S_l is closed and super-additive)*
Let $\mathcal{Y}_{l-1}$ be a Borel measurable subset of $\mathbb{R}^{n_{l-1} \times m}$. Let $||T||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} = \inf\{c \geq 0 : ||Tv||_{\mathbb{R}^{n_l \times m}} \leq c||v||_{\mathbb{R}^{n_l \times m}}$ for all $v \in \mathbb{R}^{n_l \times m}\}$ be the standard operator norm for bounded operators mapping $\mathbb{R}^{n_l \times m}$ into itself. Let ϕ_l, ϕ^*_l be measurable functions such that, ϕ^*_l is a function of bounded variation and the self-adjoint operator given by $\phi(y)\phi^*(y) = L_y L^*_y$ is a closed and bounded linear operator, for all $y \in \mathcal{Y}_{l-1}$ and there exists a constant $M < \infty$ satisfying the bound $\sup\{||\phi_l(y)\phi^*_l(y)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} : y \in \mathcal{Y}_{l-1}\} = M$ for all $y \in \mathcal{Y}$. Let $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ be the Banach space of signed, $\mathbb{R}^{n_l \times m}-$valued Borel measures with finite total variation. Then $S_l : \mathcal{Z}^{(l)} \to \mathcal{V}(\mathcal{Z}^{(l)})$ as defined by (2.45) is a closed and super-additive subspace valued map.

Proof: *The map S_l is a product of $S_{\mathbb{R}}$ with the set $K =$ closure$\left(\left\{\int_{\mathcal{Y}_l} \phi^*_l(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right)$. $S_{\mathbb{R}}$ is already known to be closed and super-additive (from Example 1). Thus it only remains to be shown that the set K is closed, super-additive and actually subspace valued i.e. $K \subseteq \mathcal{H}^{(l)}$ and $K \in \mathcal{V}(\mathcal{H}^{(l)})$.*

If ϕ^*_l is assumed to be integrable with respect to every $c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$, it is easy to see that K is a vector space since for any $f_1, f_2 \in K$, there exist $c^1_l, c^2_l \in$

$M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ such that $f_i = \int_{\mathcal{Y}_l} \phi_l^*(y) dc_l^i(y)$ for $i = 1, 2$. Thus by linearity for the integral we have for any $\alpha, \beta \in \mathbb{R}$, $\alpha f_1 + \beta f_2 = \int_{\mathcal{Y}_{l-1}} \phi_l^*(y) d(\alpha c_l^1(y) + \beta c_l^2(y))$. Since $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ is a closed vector space (actually a Banach space), we have a $c_l' = \alpha c_l^1 + \beta c_l^2 \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ and thus $\alpha f_1 + \beta f_2$ belongs to K. Next we show that under the conditions mentioned for ϕ_l, ϕ_l^*, ϕ_l^* is actually integrable with respect to every $c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ and that $K \subseteq \mathcal{H}^{(l)}$ is actually a closed subspace, i.e. $K \in \mathcal{V}(\mathcal{H}^{(l)})$.

By [31, Definition 1], the measurable function $\phi_l^* : \mathcal{Y}_{l-1} \to \mathcal{L}_{\mathbb{R}^{n_l \times m}, \mathcal{H}^{(l)}}$ is integrable with respect to a $\mathbb{R}^{n_l \times m}-$valued measure $c_l : \mathcal{B}(\mathcal{Y}_{l-1}) \to \mathbb{R}^{n_l \times m}$ if there exists a $h \in \mathcal{H}^{(l)}$ such that for every $\epsilon > 0$, and every countable partition $\{E_i\}$ of $\mathcal{Y}_{l-1}$ with maximum volume ϵ for any cell in the partition, and any selection of points $y_i \in E_i$, we have $||h - \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \epsilon$. For this to be true we need that $|| \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \infty$ for any partition $\{E_i\}$ and selection $y_i \in E_i$, and then convergence of $\sum_{E_i} \phi_l^*(y_i) c_l(E_i)$ to a unique $h \in \mathcal{H}^{(l)}$.

First we show that $|| \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \infty$. Note that $|| \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} \leq \sum_{E_i} ||\phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} = \sum_{E_i} (\langle \phi_l^*(y_i) c_l(E_i), \phi_l^*(y_i) c_l(E_i) \rangle_{\mathcal{H}^{(l)}})^{1/2} = \sum_{E_i} (\langle c_l(E_i), \phi_l(y_i) \phi_l^*(y_i) c_l(E_i) \rangle_{\mathbb{R}^{n_l \times m}})^{1/2} \leq \sum_{E_i} (||c_l(E_i)||_{\mathbb{R}^{n_l \times m}} ||\phi_l(y_i) \phi_l^*(y_i) c_l(E_i)||_{\mathbb{R}^{n_l \times m}})^{1/2} \leq \sum_{E_i} (||c_l(E_i)||^2_{\mathbb{R}^{n_l \times m}} ||\phi_l(y_i) \phi_l^*(y_i)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}})^{1/2} = \sum_{E_i} ||c_l(E_i)||_{\mathbb{R}^{n_l \times m}} ||\phi_l(y_i) \phi_l^*(y_i)||^{1/2}_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}}$. But note from the assumptions on $\phi_l(y) \phi_l^*(y)$, that $||\phi_l(y_i) \phi_l(y_i)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} < M$. Thus $|| \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} \leq M \sum_{E_i} ||c_l(E_i)||_{\mathbb{R}^{n_l \times m}} \leq M ||c_l||_{M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}), \mathbb{R}^{n_l \times m})}$ (the inequality follows from the definition of the norm on the space of vector measures [32], $||c_l||_{M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}), \mathbb{R}^{n_l \times m})} = \sup \sum_{E_i} ||c_l(E_i)||_{\mathbb{R}^{n_l \times m}}$, over all partitions $\{E_i\}$ of $\mathcal{Y}_{l-1}$). Thus for any $c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}), \mathbb{R}^{n_l \times m})$ and any partition $\{E_i\}$ of $\mathcal{Y}_{l-1}$, $|| \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \infty$.

Now for the uniqueness in convergence, note that for any refinement $\{E_i'\}$ of a partition $\{E_i\}$ such that $E_i = E_{2i}' \cup E_{2i+1}'$ for all $i = 0, \ldots, \infty$, and selection of points without loss of generality, as $y_i = y_{2i}'$, we have $|| \sum_{E_i} \phi_l^*(y_i) c_l(E_i) - \sum_{E_i'} \phi_l^*(y_i') c_l(E_i')|| = || \sum_{E_{2i}'} \phi_l^*(y_{2i}') c_l(E_{2i}') + \sum_{E_{2i+1}'} \phi_l^*(y_{2i}') c_l(E_{2i+1}') - \sum_{E_{2i}'} \phi_l^*(y_{2i}') c_l(E_{2i}') - \sum_{E_{2i+1}'} \phi_l^*(y_{2i+1}') c_l(E_{2i+1}')|| = || \sum_{E_{2i+1}'} (\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1})) c_l(E_{2i+1}')|| \leq \sum_{E_{2i+1}'} ||(\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1}'))|| ||c_l(E_{2i+1}')||$. Since ϕ_l^* has bounded total variation for any partition E_i', we have $\sum_{E_{2i+1}'} ||(\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1}'))|| \leq \sup \sum_{E_i'} ||(\phi_l^*(y_i') - \phi_l^*(y_{i+1}'))|| < \infty$. Then as the partition refinements converge $E_i' \to E_i$, $c_l(E_{2i+1}')$ tends to 0 and thus we have $\sum_{E_{2i+1}'} ||(\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1}'))|| ||c_l(E_{2i+1}')||$ converging to 0.

Thus we have shown that the conditions of ϕ_l^* and $\phi_l(y) \phi_l^*(y)$ ensure that ϕ_l^* is integrable with respect to all c_l and thus the integral $h = \int_{\mathcal{Y}_l} \phi_l^*(y) dc_l(y)$ belongs to $\mathcal{H}^{(l)}$ for all c_l, implying K is a vector subspace of $\mathcal{H}^{(l)}$.

Finally taking the closure of K, makes the subspace a closed subspace of $\mathcal{H}^{(l)}$ (since $\mathcal{H}^{(l)}$ is closed).

Now, since $S_l(y^{(l)}, f) = S_{\mathbb{R}}(y^{(l)}) \times K$ for any $f \in \mathcal{H}^{(l)}$ (K does not depend on f), we have for any closed subspaces $A, B \in \mathcal{V}(\mathcal{Z}^{(l)})$, we have $S_l(A) = A_y \times K$ and $S_l(B) = B_y \times K$, where A_y, B_y are the vector subspaces in $\mathcal{V}(\mathbb{R}^{n_{l-1} \times m} \times \mathbb{R}^{n_l \times m})$ corresponding to the additive subspace valued map $S_{\mathbb{R}}$. Thus we have $S_l(A) + S_l(B) = A_y \times K + B_y \times K = (A_y + B_y) \times K = S_l(A + B)$ (since we have $S_{\mathbb{R}}(A_y) = A_y, S_{\mathbb{R}}(A_y) = B_y$ and $S_{\mathbb{R}}(A_y + B_y) = A_y + B_y$), implying S_l is closed and super-additive (trivially, since its additive). □

For the last layer, consider,

$$S_N(f^{(N)}) = \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi_l^*(y)dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{N-1}, \mathcal{B}(\mathcal{Y}_{N-1}); \mathbb{R}^{n_N \times m})\right\}\right) \tag{2.46}$$

which is again a closed, super-additive subspace valued map by the arguments in Lemma 2.4.

Now consider the subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ for the complete network given by

$$S(F) = S_1(z^{(1)}) \times S_2(z^{(2)}) \times \cdots \times S_{N-1}(z^{(N-1)}) \times S_N(f^{(N)}) \tag{2.47}$$

Theorems 2.5 below, shows that the functional Ω in (2.43) is orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$ for $\mathcal{N}_L^\perp$ as defined in (2.42).

Theorem 2.5. *Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be the closed, super-additive subspace valued map defined in (2.46). Let $\mathcal{N}_L^\perp$ be the closed subspace as defined in (2.42) and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be the functional from (2.43). Then Ω is orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$, i.e. for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \Omega(f)$.*

Proof: *For the vector subspace $\mathcal{N}_L^\perp = \mathbb{R}^{n_l \times m} \times \text{range}(L_{y_0}^*) \times \{0\} \times \{0\} \cdots \times \{0\}$ in $\mathcal{V}(\mathcal{H})$, we have $S(\mathcal{N}_L^\perp) = \mathbb{R}^{n_1 \times m} \times \text{range}(L_{y^{(0)}}^*) \times (\Pi_{l=2}^{N-1} \mathbb{R}^{n_l \times m} \times K_l) \times K_N$, for the subspace $K_l = \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi_l^*(y)dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right)$ for each $l = 2, \ldots, N$. Also $S(\mathcal{N}_L^\perp)^\perp = \{0\} \times \text{range}(L_{y_0}^*)^\perp \times (\Pi_{l=2}^{N-1}\{0\} \times K_l^\perp) \times K_N^\perp$. Also recall that $\Omega(F) = \Omega_1(z^{(1)}) + \sum_{l=2}^N (C_l(y^{(l-1)}, z^{(l)}) + \Omega_l(z^{(l)}))$*

Thus for $F = (z^{(1)}, \ldots, z^{(N-1)}, f^{(N)}) \in S(\mathcal{N}_L^\perp)$, we have $z^{(1)} = (y^{(1)}, f^{(1)}) \in \mathbb{R}^{n_1 \times m} \times \text{range}(L_{y_0}^*)$, $z^{(l)} = (y^{(l)}, f^{(l)}) \in \mathbb{R}^{n_l \times m} \times K_l$ for $l = 2, \ldots, N-1$ and $z^{(N)} \in K_N$.

Similarly for $G = (x^{(1)}, \ldots, x^{(N-1)}, g^{(N)}) \in S(\mathcal{N}_L^\perp)^\perp$, we have $x^{(l)} = (y'^{(l)}, g^{(l)})$ for $y'^{(l)} = 0$ and $g^{(l)} \in K_l^\perp$, for each $l = 2, \ldots, N-1$, $g^{(N)} \in K_N^\perp$. And $x^{(1)} = (y'^{(1)}, g^{(1)})$ for $y'^{(1)} = 0$ and $g^{(1)} \in \text{range}(L_{y_0}^*)^\perp$.

Now for $l = 1$, note that the only term in Ω depending on $z^{(1)}$ and $x^{(1)}$ is Ω_1 defined as $\Omega(f) = ||f||^2_{\mathcal{H}^{(1)}}$. The squared norm functional is orthomonotone for any pair of orthogonal subspaces (from Example 2.5). Thus for any orthogonal $z^{(1)}$ and $x^{(1)}$ as defined above we have $\Omega_1(z^{(1)} + x^{(1)}) = ||z^{(1)} + x^{(1)}||^2 = ||z^{(1)}||^2 + ||x^{(1)}||^2 \geq ||z^{(1)}||^2 = \Omega_1(z^{(1)})$ (the equality of square of sum, to sum of squares, follows from orthogonality of the two vectors).

Similarly for each $l = 2, \ldots, N-1$, for the orthogonal vectors $z^{(l)}$ and $x^{(l)}$, we have $\Omega_l(z^{(l)} + x^{(l)}) \geq \Omega_l(z^{(l)})$ and for $f^{(N)}, g^{(N)}$, $\Omega(f^{(N)} + g^{(N)}) \geq \Omega(f^{(N)})$.

The terms remaining to be shown orthomonotone are the functional C_l. Note that for all $l = 2, \ldots, N$, we have $y^{(l-1)} \in \mathbb{R}^{n_{l-1} \times m}$, $y^{(l)} \in \mathbb{R}^{n_l \times m}$ and $f^{(l)} \in K_l$, and we have $y'^{(l-1)} = 0$, $y'^{(l)} = 0$ and $g^{(l)} \in K_l^\perp$.

Then, $C_l(y^{(l-1)} + y'^{(l-1)}, y^{(l)} + y'^{(l)}, f^{(l)} + g^{(l)}) = C_l(y^{(l-1)}, y^{(l)}, f^{(l)} + g^{(l)})$ (since $y'^{(l-1)} = 0$ and $y'^{(l-1)} = 0$).

Now note that for any $l = 1, \ldots, N-1$, if $y^{(l)} \notin \mathcal{Y}_l$, then $C_l(y^{(l-1)}, y^{(l)}, f) = \infty$ for any $f \in \mathcal{H}^{(l)}$. Then we trivially have $\Omega(F+G) = \Omega(F) = \infty$ (thus satisfying the orthomonotone inequality trivially).

For all $y^{(l)} \in \mathcal{Y}_l$, for $f^{(l)} \in K_l$, $g^{(l)} \in K_l^{\perp}$, we have $C_l(y^{(l-1)}, y^l, f^{(l)} + g^{(l)}) = C_l(y^{(l)} - \sigma_l(\phi_l(y^{(l-1)})f^{(l)} + \phi_l(y^{(l-1)})g^{(l)}))$ for some $c_l \in M_\sigma(\mathcal{Y}_{l-1})$.

Now since for all $z \in \mathbb{R}^{n_l \times m}$, $c_l = z\delta_{y^{(l-1)}}$ (where $\delta_{y^{(l-1)}}$ is the dirac measure centered on $y^{(l-1)}$) belongs to $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$, $\mathcal{N}^{\perp}_{\phi_l(y^{(l-1)})} = \mathrm{range}(\phi_l^*(y^{(l-1)})) \subseteq K_l$, implying $K_l^{\perp} \subseteq \mathcal{N}_{\phi_l(y^{(l-1)})}$ for all $y^{(l-1)} \in \mathcal{Y}_{l-1}$. Thus for any $g \in K_l^{\perp}$, $\phi_l(y^{(l-1)})g = 0$. Thus $C_l(y^{(l-1)}, y^l, f^{(l)} + g^{(l)}) = C_l(y^{(l-1)}, y^l, f^{(l)})$, for all $y^{(l)} \in \mathcal{Y}_l$, $y^{(l-1)} \in \mathcal{Y}_{l-1}$, $f^{(l)} \in K_l$, $g^{(l)} \in K_l^{\perp}$.

Thus for all $l = 2, \dots, N$, we have shown $C_l(y^{(l-1)} + y'^{(l-1)}, y^{(l)} + y'^{(l)}, f^{(l)} + g^{(l)}) + \Omega_l(y^{(l)} + y'^{(l)}, f^{(l)} + g^{(l)}) \geq C_l(y^{(l-1)}, y^{(l)}, f^{(l)}) + \Omega_l(y^{(l)}, f^{(l)})$ and $\Omega_1(z^{(l)} + x^{(l)}) \geq \Omega_1(z^{(l)})$.

Thus $\Omega(F+G) \geq \Omega(F)$ for all $F \in S(\mathcal{N}_L^{\perp})$, $G \in S(\mathcal{N}_L^{\perp})^{\perp}$, i.e. Ω is orthomonotone with respect to $S(\mathcal{N}_L^{\perp})$. □

Corollary 2.2. *(S is range preserving with respect to L)*
For $L : \mathcal{H} \to \mathbb{R}^{n_1 \times m} \times \mathbb{R}^{n_1 \times m}$ defined as $LF = (y^{(1)}, L_{y_0} f^{(1)})$ for $F = (z^{(1)}, \dots, z^{(N-1)}, f^{(N)}) \in \mathcal{H}$, $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ as defined in (2.46) and $\mathcal{N}_L^{\perp}$ as defined in (2.42), we have $\mathcal{N}_L^{\perp} \subseteq S(\mathcal{N}_L^{\perp})$.

Proof:

$$\mathcal{N}_L^{\perp} = \mathbb{R}^{n_l \times m} \times \mathrm{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\}$$

and

$$S(\mathcal{N}_L^{\perp}) = \mathbb{R}^{n_1 \times m} \times \mathrm{range}(L^*_{y^{(0)}}) \times (\Pi_{l=2}^{N-1} \mathbb{R}^{n_l \times m} \times K_l) \times K_N$$

for the subspace $K_l = \mathrm{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi_l^*(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right)$ for each $l = 2, \dots, N$. From the above expressions, it is visible that $\mathcal{N}_L^{\perp} \subseteq S(\mathcal{N}_L^{\perp})$ □

$\mathcal{N}_L^{\perp} = \mathbb{R}^{n_l \times m} \times \mathrm{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\}$ in $\mathcal{V}(\mathcal{H})$, we have

Corollary 2.3. *(Linear representer for the neural network exists in $S(\mathcal{N}_L^{\perp})$)*
There exists an optimal set of representers $c_{1,opt} \in \mathbb{R}^{n_1 \times m}$, $y^{(l)}_{opt} \in \mathcal{Y}_l$ for $l = 1, \dots, N-1$, and $c_{l,opt} \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ for $l = 2, \dots, N$ such that a minimizer for (2.41) of the form

$$F_{opt} = \left(y^{(1)}_{opt}, \dots, y^{(N-1)}_{opt}, L^*_{y_0} c_{1,opt}, \int_{\mathcal{Y}_1} \phi_2^* c_{2,opt}, \dots, \int_{\mathcal{Y}_{N-1}} \phi_N^* c_{N,opt}\right) \tag{2.48}$$

exists.

Proof: *Theorem 2.5 showed that Ω is orthomonotone with respect to the subspace $S(\mathcal{N}_L^{\perp})$ and Corollary 2.2 showed that S is range preserving with respect to L. Thus by the sufficient condition for existence of representers (Theorem 2.3) a linear representer exists in the subspace $S(\mathcal{N}_L^{\perp})$, written as (2.48).* □

Now since we know that, given the optimal solution $y^{(l)}_{opt}, y^{(l-1)}_{opt}$, for the $(l-1)^{th}$ and $(l)^{th}$ layer, we have,

$$f^{(l)}_{opt} = \underset{h^{(l)} \in \mathcal{H}^{(l)}}{\arg\min}\, C_l(y^{(l)}_{opt} - \sigma_l(L_{y^{(l-1)}_{opt}} f^{(l)})) + ||f^{(l)}||^2_{\mathcal{H}^{(l)}} \tag{2.49}$$

for all $l = 2, \dots, N$, we have $f^{(l)}_{opt} = L^*_{y^{(l-1)}_{opt}} p_{l,opt}$ for some $p_{l,opt} \in \mathbb{R}^{n_l \times m}$.

This implies that an optimal solution of the form

$$F_{opt} = \left(y^{(1)}_{opt}, \dots, y^{(N-1)}_{opt}, L^*_{y_0} c_{1,opt}, \int_{\mathcal{Y}_1} \phi^*_2 p_{2,opt} d\delta_{y^{(1)}_{opt}}, \dots, \int_{\mathcal{Y}_{N-1}} \phi^*_N p_{N,opt} d\delta_{y^{(N-1)}_{opt}} \right) \tag{2.50}$$

exists for some $p_{l,opt} \in \mathbb{R}^{n_l \times m}$, i.e. we know that there exist dirac measures $\delta_{y^{(l)}_{opt}}$ corresponding to the optimal measures $c_{l,opt}$ from (2.48). Such a representer in terms of diracs is not directly useful, since the points at which the optimal diracs are centered $y^{(l)}_{opt}$ are unknown apriori. We can however use this knowledge to guide our search for measures converging towards diracs.

We can thus design a scheme to iteratively optimize over the space of measures $c^{(l)}$ and outputs $y^{(l)}$ such that the measures converge to dirac's centered at the predicted output. In particular if the maps ϕ_l and ϕ^*_l are differentiable (in addition to the regularity conditions of Theorem 2.4), we can design a scheme to optimize directly over the centers of dirac measures. We show in the next subsection a numerical example for such a scheme for a squared exponential kernel, satisfying the regularity and differentiability conditions.

Numerical example

Consider a N-layer network with each layer given by an RKHS space $\mathcal{H}^{(l)}$ with a matrix valued square exponential kernel,

$$K_l(x, y) = \begin{pmatrix} e^{-a^{(l)}_{11}||x-y||^2} & \dots & e^{-a^{(l)}_{1n_l}||x-y||^2} \\ \vdots & \vdots & \vdots \\ e^{-a^{(l)}_{n_l 1}||x-y||^2} & \dots & e^{-a^{(l)}_{n_l n_l}||x-y||^2} \end{pmatrix} \tag{2.51}$$

for some known constants $a^{(l)}_{11}, \dots, a^{(l)}_{n_l n_l}$, mapping $x, y \in \mathbb{R}^{n_l}$ to a matrix in $\mathbb{R}^{n_l \times n_l}$.

Given m training samples, we denote the output of a layer as $y^{(l)} = (y^{(l)}_1, \dots, y^{(l)}_m) \in \mathbb{R}^{n_l \times m}$. The kernel function is extended to inputs from $\mathbb{R}^{n_l \times m}$ by computing the matrix,

$$K_l(x, y) = \begin{pmatrix} K_l(x_1, y_1) & \dots & K_l(x_1, y_m) \\ \vdots & \vdots & \vdots \\ K_l(x_m, y_1) & \dots & K_l(x_m, y_m) \end{pmatrix} \tag{2.52}$$

where x_i, y_i denotes the i^{th} column of x and y respectively.

Let $E_y : \mathcal{H}^{(l)} \to \mathbb{R}^{n_l \times m}$ denote the evaluation operator such that $E_y f = (f(y_1), \dots, f(y_m))$. The adjoint to the evaluation operator on the RKHS space is given by the kernel function and thus we have the adjoint $E^*_y = K_l(\cdot, y)$.

Let $\mathcal{O}_l = \{E_y : y \in \mathbb{R}^{n_l \times m}\}$ be the set of all the linear evaluation operators on $\mathcal{H}^{(l)}$. Similarly, let $\mathcal{O}^*_l = \{E^*_y : y \in \mathbb{R}^{n_l \times m}\}$ denote the set of all adjoints to the linear evaluation operators on $\mathcal{H}^{(l)}$.

Thus we have the function $\phi^* : \mathbb{R}^{n_l \times m} \to \mathcal{O}^*_l$ given by $\phi^*(y) = K_l(\cdot, y)$ and a function $\phi : \mathbb{R}^{n_l \times m} \to \mathcal{O}_l$ given by $\phi(y) = E_y$.

Let $\sigma_l : \mathbb{R} \to \mathbb{R}$ be the hyperbolic tangent function $\sigma(x) = \tanh(x)$, extended to inputs from

$\mathbb{R}^{n_l \times m}$, as

$$\sigma(X) = \begin{pmatrix} \tanh(X_{11}) & \cdots & \tanh(X_{1m}) \\ \vdots & \vdots & \vdots \\ \tanh(X_{m1}) & \cdots & \tanh(X_{mm}) \end{pmatrix}$$

where X_{ij} denotes the $(i,j)^{th}$ component of the matrix X.
Thus the activation function restricts the output of the $l^{th}-$layer to the set

$$\mathcal{Y}_l = \{X \in \mathbb{R}^{n_l \times m} : X_{ij} \in (-1,1) \text{ for all } i,j\}$$

for all $l = 1, \ldots, N$

Since the function $\phi_l^* : \mathcal{Y}_{l-1} \to \mathcal{O}_l^*$, given by $\phi^*(y) = K_l(\cdot, y)$ has a bounded domain $\mathcal{Y}_{l-1}$ and $K_l(\cdot, y)$ is a smooth bounded function in y, ϕ_l^* is a function of bounded variation on $\mathcal{Y}_{l-1}$ and thus satisfies the regularity condition for Lemma 2.4. Similarly $||\phi_l(y)\phi_l^*(y)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} = ||K_l(y,y)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} = ||K_l(0,0)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} < \infty$ for all $y \in \mathcal{Y}_{l-1}$, we have the regularity condition for $\phi_l \phi_l^*$ satisfied as well.

Now since $\phi^*(y)$ is a smooth function in y and we know that a optimal solution to the linear representer of the form (2.50) exists on the smooth manifold such that $(y_{opt}^{(l-1)}, f_{opt}^{(l)}) \in \{(y, \phi_l^*(y)p_l) : y \in \mathcal{Y}_{l-1}, p_l \in \mathbb{R}^{n_l \times m}\}$, we can instead solve the smooth finite dimensional optimization problem

$$\begin{matrix} p_{1,opt}, \ldots, p_{N,opt}, \\ y_{1,opt}, \ldots, y_{N-1,opt} \end{matrix} = \underset{p_l \in \mathbb{R}^{n_l \times m}, y_l \in \mathcal{Y}_l \subseteq \mathbb{R}^{n_l \times m}}{\arg\min} \sum_{l=1}^{N} C_l(y^{(l)} - \sigma_l(L_{y^{(l-1)}} L^*_{y^{(l-1)}} p_l)) + \Omega_l(f^{(l)}) \tag{2.53}$$

Figure 2.1 shows the output of a three layer neural network trained in such a way for 3 way classification of a given set of points in $\mathbb{R}^2$. The output of the network is in $\mathbb{R}^3$, with the training data given such that the i^{th} component is set to 1 if a point is in the i^{th} class and the other components are set to 0. The trained network provides an output in $\mathbb{R}^3$ and the output is passed through a soft-max function (to rescale values in each component to [0,1]) and interpreted as class probability for points in $\mathbb{R}^2$ shaded with corresponding RGB color values (a small problem in $\mathbb{R}^2$ is chosen to allow for easy visualization of the results). Also note that the optimization scheme in (2.53) can only guarantee convergence to a local minimizer, but this is most often the case in neural networks due to the non-convex nature of the problem.

2.4.2 Multi-output stochastic regression with uncertain observations

Let $\mathcal{Z} = \mathcal{C}_b(\mathcal{X})$ be the Banach space of continuous and bounded $\mathbb{R}^n-$valued functions on some domain set $\mathcal{X}$. Let $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z}$ and $\mu : \mathcal{B}(\mathcal{Z}) \to [0,1]$ be a Gaussian measure on $\mathcal{Z}$. Let $\mathcal{Z}_\mu$ be the Banach space of all affine measurable functions $X : \mathcal{Z} \to \mathcal{Z}$ with $\mathcal{B}(\mathcal{Z}_\mu)$ being the Borel σ-algebra on $\mathcal{Z}_\mu$. $\mathcal{Z}_\mu$ defines a space of Gaussian processes on the probability measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ (see Section D.1). Let $\nu : \mathcal{B}(\mathcal{Z}_\mu) \to [0,1]$ be a Gaussian measure on $\mathcal{Z}_\mu$ and let $\mathcal{H}_{\mu,\nu}$ be the RKHS space of Gaussian processes induced by the measure ν on $\mathcal{Z}_\mu$ as defined in Section D.2. Let $\mathcal{Y} = \mathbb{R}^n$ and $\mathcal{B}(\mathcal{Y})$ be the Borel σ-algebra on $\mathbb{R}^n$. Let $L_x : \mathcal{Z} \to \mathcal{Y}$ be the closed, bounded linear evaluation operator $L_x f = f(x)$. The linear operator L_x induces induces a push forward Gaussian measure $\mu \circ L_x^{-1}$ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ (L_x^{-1} denoting the preimage operation, not the linear inverse). Let $\mathcal{Y}_{x,\mu}$ denote the Hilbert space of affine measurable functions $y : \mathcal{Y} \to \mathcal{Y}$ induced by the push forward measure $\mu \circ L_x^{-1}$ with the inner product $\langle y_1, y_2 \rangle_{\mathcal{Y}_{x,\mu}} = \int_{\mathcal{Y}} y_1(\omega) y_2(\omega) d(\mu \circ L_x^{-1})(\omega)$. $\mathcal{Y}_{x,\mu}$ is thus denotes a Hilbert space of $\mathbb{R}^n-$valued Gaussian random vectors. The extension of L_x to $\mathcal{H}_{\mu,\nu}$, for any affine function $X : \mathcal{Z} \to \mathcal{Z}$ in $\mathcal{Z}_\mu$, is given as $L_x X(f) = X(L_x f) = X(f(x))$, and defines a linear operator $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{x,\mu}$, to space of Gaussian random vectors in $\mathcal{Y}_{x,\mu}$. The extension $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{x,\mu}$ also preserves the closed and bounded property of $L_x : \mathcal{Z} \to \mathcal{Y}$ (by Lemma D.5).

Assuming that the $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{x,\mu}$ induces equivalent Gaussian measures on $\mathcal{Y}$ for all $x \in \mathcal{X}$, we can write the map as $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{c,\mu}$, mapping into a common probability measure space on $\mathcal{Y}$. The adjoint L_x^* can then be specified by a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}^+_{\mathcal{Y}_{c,\mu},\mathcal{Y}_{c,\mu}}$ for the RKHS space $\mathcal{H}_{\mu,\nu}$ such that for all $y \in \mathcal{Y}_{c,\mu}$ and $f \in \mathcal{H}_{\mu,\nu}$, $\langle L_x^* y, f \rangle_{\mathcal{H}_{\mu,\nu}} = \langle K(\cdot, x) y, f \rangle_{\mathcal{H}_{\mu,\nu}} = \langle y, L_x f \rangle_{\mathcal{Y}_{c,\mu}}$. Note that $\mathcal{L}^+_{\mathcal{Y}_{c,\mu},\mathcal{Y}_{c,\mu}}$ denotes the space of closed, bounded symmetric positive definite linear operators from $\mathcal{Y}_{c,\mu}$ into itself. Since $\mathcal{Y}_{c,\mu}$ is a Banach space of Gaussian random vectors given by all affine transformations of $\mathcal{Y}$, we must have the kernel as a deterministic function taking values in $\mathcal{L}^+_{\mathcal{Y},\mathcal{Y}}$ (else the Gaussianity will be lost), i.e., $K(x_1, x_2)(\omega) = K'(x_1, x_2)$ for all $\omega \in \mathcal{Y}$ and $K' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n \times n}$ being a deterministic kernel of the kind usually used in non stochastic variants of kernel regression (see for example the squared exponential kernel used in Section 2.4.1). The form of the kernel is determined by the choice of the Gaussian measure ν and vice versa (in general the kernel function is chosen and the measure ν is as a result determined implicitly as there is a one to one correspondence between Gaussian measures on separable Banach spaces and the induced RKHS spaces).

Now with the spaces and adjoint defined we can consider a regression problem on the RKHS space of Gaussian processes $\mathcal{H}_{\mu,\nu}$. Let $\mathcal{H}_{\mu,\nu}$ be the RKHS space of Gaussian process with a kernel K. Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}_{c,\mu} : i = 1, \dots, m\}$ be a given training data set with observations $y_i \in \mathcal{Y}_{c,\mu}$ given as $\mathbb{R}^n-$valued Gaussian random vectors. Then consider the regression problem,

$$f_{opt} = \underset{f \in \mathcal{H}_{\mu,\nu}}{\operatorname{argmin}} \quad \sum_{i=1}^{m} ||y_i - L_{x_i} f||^2_{\mathcal{Y}_{c,\mu}} + \lambda ||f||^2_{\mathcal{H}_{\mu,\nu}} \tag{2.54}$$

Note that the observations $y_i \in \mathcal{Y}_{\mu,\nu}$ are now $\mathbb{R}^n-$valued Gaussian random vectors and not points in $\mathbb{R}^n$, making the loss functional $C_i : \mathcal{Y}_{c,\mu} \to \mathbb{R} \cup \{\infty\}$, given as $||y_i - L_{x_i} f||^2_{\mathcal{Y}_{c,\mu}}$, an example of a loss functional defined on a separable Hilbert space different from $\mathbb{R}^n$. Also note that even though the functional $C_i(y_i - L_{x_i}(f))$ can be written as in terms of the mean and covariance of a $\mathbb{R}^n-$valued Gaussian random vector, i.e. a functional of the form $C_i' : \mathbb{R}^n \times \mathbb{R}^{n \times n} \to \mathbb{R} \cup \{\infty\}$, as we will see below, we cannot write this reformulated objective as an equivalent functional $C_i' \circ L_{x_i}' : \mathcal{H}_{\mu,\nu} \to \mathbb{R} \cup \{\infty\}$ for a linear operator $L_{x_i}' : \mathcal{H}_{\mu,\nu} \to \mathbb{R}^n \times \mathbb{R}^{n \times n}$, mapping the stochastic process f to its mean and covariance at x_i (since the mapping from f to its covariance will be a nonlinear operator). Thus (2.54) presents an example of a regression problem where the functional $C_i : \mathcal{Y}_{c,\mu} \to \mathbb{R} \cup \{\infty\}$ must be considered on the infinite dimensional Hilbert space of measurable affine maps given by $\mathcal{Y}_{c,\mu}$ in order to establish a representer in terms of the adjoint $L_{x_i}^*$.

Since, $\Omega(f) = ||f||^2_{\mathcal{H}_{\mu,\nu}}$ is known to be orthomonotone with respect to the the subspace valued map $S_{\mathbb{R}}$, we can write the representer for (2.54) as

$$S_{\mathbb{R}}\left(\sum_{i=1}^{m}\operatorname{range}(L^*_{x_i})\right)=\left\{\sum_{i=1}^{m}K(\cdot,x_i)z_i : z_i\in\mathcal{Y}_{c,\mu}\right\} \tag{2.55}$$

Substituting a representer into (2.54), we can write the equivalent optimization problem,

$$\begin{aligned} f_{opt} &= \sum_{i=1}^{m} K(\cdot,x_i)z_i^{opt} \\ z_1^{opt},\dots,z_m^{opt} &= \arg\min_{z_i\in\mathcal{Y}_{c,\mu}}\sum_{i=1}^{m}||y_i-\sum_{j=1}^{m}K(x_i,x_j)z_j||^2_{\mathcal{Y}_{c,\mu}}+\sum_{i=1}^{m}\sum_{j=1}^{m}\langle z_i,K(x_i,x_j)z_j\rangle_{\mathcal{Y}_{c,\mu}} \\ &= \arg\min_{z_i\in\mathcal{Y}_{c,\mu}}\sum_{i=1}^{m}\mathbb{E}_\mu\left[||y_i-\sum_{j=1}^{m}K(x_i,x_j)z_j||^2_{\mathbb{R}^n}+\sum_{j=1}^{m}z_i^TK(x_i,x_j)z_j\right] \end{aligned} \tag{2.56}$$

where $\mathbb{E}_\mu$ is the expectation with respect to the μ. We can expand the expectation from (2.56) as

$$\begin{aligned} &\mathbb{E}_\mu[||y_i||^2_{\mathbb{R}^n}]+\mathbb{E}_\mu\left[||\sum_{j=1}^{m}K(x_i,x_j)z_j||^2_{\mathbb{R}^n}\right] \\ -2\mathbb{E}_\mu&\left[y_i^T\sum_{j=1}^{m}K(x_i,x_j)z_j\right]+\mathbb{E}_\mu\left[\sum_{j=1}^{m}z_i^TK(x_i,x_j)z_j\right] \end{aligned} \tag{2.57}$$

For the terms involving the decision variables $z_i\in\mathcal{Y}_{c,\mu}$, let $K^{xx}\in\mathbb{R}^{nm\times nm}$ denote the symmetric positive definite kernel matrix such that its block $K^{xx}_{i,j}$ is the kernel evaluation $K(x_i,x_j)\in\mathbb{R}^{n\times n}$ and let Z and y be the concatenation of all z_i and y_i respectively in to the vectors $Z=(z_1,\dots,z_m)$ and $y=(y_1,\dots,y_m)$, i.e., we write,

$$K^{xx}=\begin{pmatrix}K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_m)\\ \vdots & \vdots & \vdots & \vdots \\ K(x_m,x_1) & K(x_m,x_2) & \cdots & K(x_m,x_m)\end{pmatrix},\qquad Z=\begin{pmatrix}z_1\\ z_2\\ \vdots\\ z_m\end{pmatrix}\quad y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{pmatrix} \tag{2.58}$$

$$\tag{2.59}$$

And let the mean and covariance be denoted as

$$\mathbb{E}_\mu[Z]=\mu_Z,\qquad \mathbb{E}_\mu[y]=\mu_y \tag{2.60}$$

$$\mathbb{E}_\mu[(y-\mu_y)(y-\mu_y)^T]=\Sigma_y\qquad \mathbb{E}_\mu[(Z-\mu_Z)(Z-\mu_Z)^T]=LL^T \tag{2.61}$$

for some lower triangular matrix $L\in\mathbb{R}^{nm\times nm}$ and the given covariance matrix Σ_y for the observations. We can then, rewrite the terms involving the decision variables z_i, as

$$\begin{aligned} -2\mu_y^T K^{xx}\mu_Z + \mathbb{E}_\mu[(Z(K^{xx})^{1/2})^T(K^{xx})^{1/2}Z] + \mathbb{E}_\mu[(K^{xx}Z)^T K^{xx} Z] && (2.62)\\ -2\mathbb{E}_\mu[(Y-\mu_y)^T K^{xx}(Z-\mu_Z)] && (2.63)\end{aligned}$$

and using the properties of Gaussian random vectors under affine transformations, we have,

$$\begin{aligned} \mathbb{E}_\mu[(K^{xx}Z)^T K^{xx}Z] &= \mu_Z^T K^{xx}K^{xx}\mu_Z + \mathrm{trace}(K^{xx}LL^T K^{xx}) && (2.64)\\ \mathbb{E}_\mu[(Z(K^{xx})^{1/2})^T(K^{xx})^{1/2}Z] &= \mu_Z^T K^{xx}\mu_Z + \mathrm{trace}(K^{xx}LL^T) && (2.65)\\ \mathbb{E}_\mu[(Y-\mu_y)^T K^{xx}(Z-\mu_Z)] &= \mathrm{trace}(K^{xx}L(\Sigma_y^{1/2})^T) && (2.66)\end{aligned}$$

(2.66) follows from the fact that y and Z are jointly Gaussian under a common measure $\mu : \mathcal{B}(\mathcal{Y}) \to [0,1]$ and are thus related to each other through the affine transformation

$$\begin{pmatrix} y(\zeta)\\ Z(\zeta)\end{pmatrix} = \begin{pmatrix} \Sigma_y^{1/2}\zeta\\ L\zeta\end{pmatrix} + \begin{pmatrix}\mu_y\\ \mu_Z\end{pmatrix}$$

(without loss of generality, taking μ to be the Gaussian measure for the standard normal distribution $\mathcal{N}(0,I)$ on $\mathbb{R}^n$)

Thus for the new decision variables $\mu_Z \in \mathbb{R}^{nm}$ and $L \in (\mathbb{R}^{nm\times nm})_{lt}$ (lower triangular matrix denoted with subscript lt), we can write the equivalent finite dimensional problem to (2.56) as,

$$\begin{aligned} {\mu_z}^{opt}, L^{opt} = \operatorname*{arg\,min}_{\mu_z\in\mathbb{R}^{nm}, L\in(\mathbb{R}^{nm\times nm})_{lt}} & (\mu_y - K^{xx}\mu_Z)^T(\mu_y - K^{xx}\mu_Z) + \mu_Z^T(K^{xx}K^{xx})\mu_Z \\ & + \mathrm{trace}(L^T K^{xx}L + (K^{xx}L - (\Sigma_y^{1/2}))^T(K^{xx}L - (\Sigma_y^{1/2}))) \end{aligned} \quad (2.67)$$

From (2.67) it is easy to see the problem is an unconstrained quadratic program in μ_Z and L and thus has an unique minimizer.

The final function form of f_{opt} from (2.56), is then given by the affine transformation of the random vector Z^{opt} having mean μ_Z^{opt} and covariance $L^{opt}(L^{opt})^T$.

$$f^{opt}(\cdot) = K(\cdot, X)Z^{opt} \quad (2.68)$$

where $K(\cdot,X)$ is the matrix $\begin{pmatrix} K(\cdot,x_1) & K(\cdot,x_2) & \cdots & K(\cdot,x_m)\end{pmatrix}$.
Thus we have

$$\mathbb{E}_\mu[f_{opt}(\cdot)] = K(\cdot,X)\mu_Z^{opt}$$

and covariance,

$$\mathrm{Covar}_\mu[f_{opt}(\cdot)] = K(\cdot,X)L^{opt}(L^{opt})^T K(\cdot,X)^T$$

Note that the mean coincides, as expected with the Bayesian posterior mean, however the covariance is quite different. Instead of acquiring certainty at points of observations, the regression model tries to fit the Gaussian process to the specified covariances of the observations.

Figure 2.2 shows an example for such a regression with a squared exponential kernel mapping with the output $y_i \in \mathcal{Z}$ being a two dimensional Gaussian random vector and $x_i \in \mathbb{R}$.

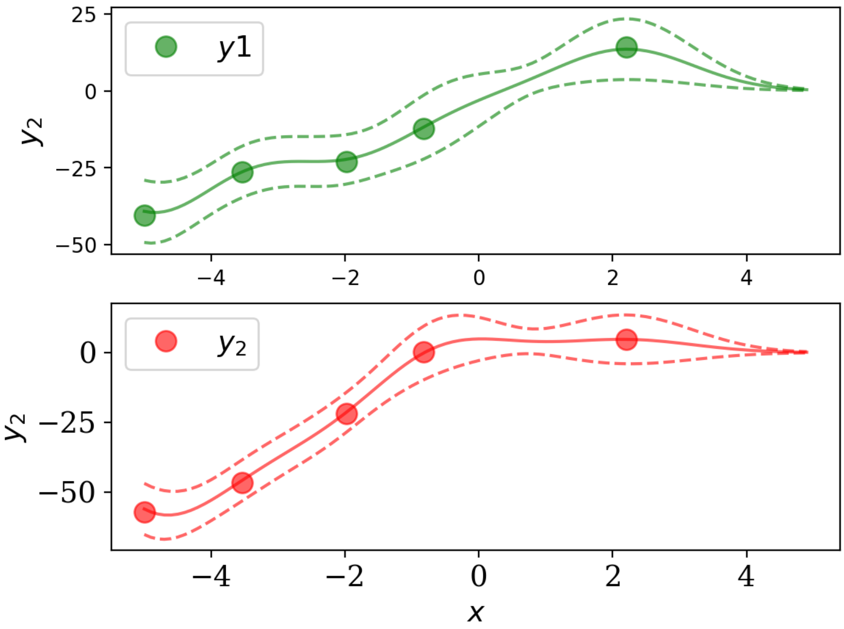


Figure 2.2: Learning a $\mathbb{R}^2$-valued Gaussian process in an RKHS of Gaussian processes

Note that while we restricted our Banach space $\mathcal{Z}_\mu$ in the beginning to a space of Gaussian processes, there is no restriction from the point of view of the representer theorem, requiring Gaussianity. The above process can in principle be repeated for any given Banach space of stochastic processes (including non-Gaussian ones) and appropriate linear operator (as the evaluation operator may not be linear for non Gaussian cases). We limit ourselves to Gaussian processes in this case as it leads to simple analytically tractable computations. Also while we restricted ourselves to a simple regression problem, note that by virtue of the generalized representer theorem we can apply the above process to many other loss functionals and regularizers to create stochastic variants of any kernel based learning algorithms like the SVM, or the neural network example from Section 2.4.1, where the RKHS space of Gaussian processes alongside a moment matching constraint between the layers can be considered, to create a Gaussian process variant for the neural network example.

The example is left limited to this simple case, as it demonstrates the key issue being considered, which is the utility of extending the loss functional to $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ for arbitrary separable Hilbert spaces $\mathcal{Z}$, like the Hilbert space of measurable functions $\mathcal{Y}_\mu$ considered above.

2.4.3 ℓ_1-Regularization

Motivating finite dimensional example

Consider first an example of the ℓ_1-regularization problem in a finite dimensional decision space. Let $\mathcal{X} = \mathbb{R}^l$, $\mathcal{Y} = \mathbb{R}^{n\times k}$, $\mathcal{Z} = \mathbb{R}^k$ and $\mathcal{H} = \mathbb{R}^n$. Let $\phi : \mathcal{X} \to \mathcal{Y}$ be a given collection of features and let $\{e_1, \ldots, e_n\}$ be the standard basis for $\mathbb{R}^n$. Consider the continuous linear operator $L_{x,\phi} : \mathcal{H} \to \mathcal{Z}$ from Example 2.2(a), where $L_{x,\phi}(w) = \phi(x)^T w$. Then consider the ℓ_1-regularization problem for feature selection given a set of observations $\mathcal{D} = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{Z}, i = 1, \ldots, m\}$ given by,

$$\min_{w \in \mathcal{H}} \quad \sum_{i=1}^{m} ||y_i - L_{x_i,\phi} w||^2_{\mathcal{Z}} + \lambda ||w||_1^2 \tag{2.69}$$

where the $||w||_1 = (\sum_{i=1}^{n} |w_i|)$ is the standard ℓ_1-norm on $\mathbb{R}^n$. Let w_i denote the i^{th} component of a vector $w \in \mathbb{R}^n$. From Theorem 2.2 we know that the ℓ_1 norm is orthomonotone with respect to a subspace valued map $S_{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, given by

$$S_{proj}(w) = \left\{ \sum_{i=1}^{n} \lambda_i \langle w, e_i \rangle_{\mathcal{H}} e_i : \lambda_i \in \mathbb{R} \right\} = \left\{ \sum_{\{i : w^T e_i \neq 0\}} \lambda_i e_i : \lambda_i \in \mathbb{R} \right\} \tag{2.70}$$

Example 3 showed that S_{proj} is an inclusive, quasilinear subspace valued map with the union extension $S_{proj} : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. Then from the representer theorem (Theorem 2.3) we know that a minimizer for (2.69) must exist in $S_{proj}(\sum_{i=1}^{m} \text{range}(L^*_{x_i,\phi})) = S_{proj}(\{\sum_{i=1}^{m} L^*_{x_i,\phi} z_i : z_i \in \mathbb{R}^k\})$.

From Example 2.2(a), we also know that $L^*_{x_i,\phi} z_i = \phi(x_i) z_i$. Thus we have

$$\begin{aligned} S_{proj}\left(\sum_{i=1}^{m} \text{range}(L^*_{x_i,\phi}) \right) &= \sum_{i=1}^{m} S_{proj}(\text{range}(L^*_{x_i,\phi})) = \sum_{i=1}^{m} S_{proj}(\{\phi(x_i) z_i : z_i \in \mathbb{R}^k\}) && (2.71) \\ &= \sum_{i=1}^{m} \left\{ \sum_{\{j : \phi(x_i)^T e_j \neq 0\}} \lambda_j e_j : \lambda_j \in \mathbb{R} \right\} && (2.72) \\ &= \left\{ \sum_{\{j:\ \phi(x_i)^T e_j \neq 0\ \forall i = 1, \ldots, m\}} \lambda_j e_j : \lambda_j \in \mathbb{R} \right\} && (2.73) \end{aligned}$$

Substituting this form of the minimizer into (2.69), we can then find the optimal λ_js. The above problem is often used as a means for sparse feature selection in learning problems.

The subspace valued map S_{proj} defined above is a $n-$regular subspace valued map as it is quasilinear, idempotent, inclusive and $S_{proj}(w)$ for any $w \in \mathcal{H}$ has dimension at most n. However if we let $n \to \infty$, S_{proj} will lose the $r-$regularity property. This does not however mean that the representer for the case of $n \to \infty$ will be infinite dimensional. In fact since the dimension of $\sum_{i=1}^{m} \text{range}(L^*_{x_i,\phi})$ is at most m, the dimension for the representer is at most $\max\{n, m\}$, even when S_{proj} is not $r-$regular for any finite r, i.e., for any $n > m$, the representer dimension is limited to m.

We show below an example of ℓ_1 regularization in an infinite dimensional space ($n = \infty$) and show an example of applying the representer theorem to a problem with a non $r-$regular subspace valued map.

A non $r-$regular example

To show an application of a non $r-$regular subspace valued map, consider an analogue of the finite dimensional example presented above over an infinite dimensional Hilbert space. For this purpose, let $\mathcal{X} = \mathbb{N}$ be the set of natural numbers and $\mathcal{Z} = \mathbb{R}$. Let $\mathcal{H} = \ell^2(\mathbb{N}, \mathbb{R})$ be the space of square summable sequences taking values in $\mathbb{R}$. For any sequence $f \in \mathcal{H}$, let $f(i)$ denote the i^{th} member of the sequence f and let $||f||_2 = (\sum_{i\in\mathbb{N}} |f(i)|^2)^{1/2} < \infty$ be the ℓ_2 norm. Let $\langle f, g\rangle_{\mathcal{H}} = \sum_{i\in\mathbb{N}} f(i)g(i)$ be the inner product on $\mathcal{H}$. Let $\langle z_1, z_2\rangle_{\mathcal{Z}} = z_1 z_2$ be the scalar product on $\mathcal{Z} = \mathbb{R}$.

As an analogue to the orthonormal basis in $\mathbb{R}^n$, consider a set of orthonormal basis for $\mathcal{H}$ given by $\{\delta_i \in \mathcal{H} : i \in \mathbb{N}\}$ with δ_i defined as $\delta_i(j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$. The above space of ℓ_2 functions forms a separable Hilbert space as shown by [33, Riesz-Fischer Theorem].

For all $f \in \mathcal{H}$, let $||f||_1 = \sum_{i\in\mathbb{N}} |f(i)|$ denote the ℓ_1 norm for the sequence. If a sequence $f \in \mathcal{H}$ is not absolutely summable, i.e. $\sum_{i\in\mathbb{N}} |f(i)|$ is not bounded, then we set $||f||_1 = \infty$.

Further note that the evaluation operator $L_x : \mathcal{H} \to \mathcal{Z}$ defined as $L_x f = f(x)$ for any $x \in \mathbb{N}$ is a bounded linear operator on $\ell_2(\mathbb{N}, \mathbb{R})$ with the adjoint L_x^* given by $\delta_x(\cdot)$, since for all $z \in \mathbb{R}$, $\langle z, L_x f\rangle_{\mathcal{Z}} = z f(x) = \langle z\delta_x, f\rangle_{\mathcal{H}} = \langle L_x^* z, f\rangle_{\mathcal{H}}$.

Then for the problem,

$$\min_{f\in\mathcal{H}} \sum_{i=1}^{m} ||y_i - L_{x_i} f||_{\mathcal{Z}}^2 + \lambda ||f||_1^2 \tag{2.74}$$

we have $\Omega : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ given by $\Omega(f) = ||f||_1^2$. The functional Ω is orthomonotone with respect to the subspace valued map

$$S_{proj}(f) = \left\{ \sum_{i=1}^{\infty} \lambda(i) \frac{\langle f, \delta_i\rangle_{\mathcal{H}} \delta_i}{||f||_{\mathcal{H}}} : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}) \right\}$$

Example 4 shows that $S_{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ defined above is an inclusive, quasilinear and superadditive subspace valued map with a union extension $S_{proj} : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. Theorem 2.2 shows that $\Omega(f) = \begin{cases} ||f||_1^2 & \sum_{i=1}^{\infty} |f(i)| < \infty \\ +\infty & \text{otherwise} \end{cases}$ is orthomonotone with respect to the S_{proj} defined above.

Note also that $S_{proj}(f)$ in general can be infinite dimensional and thus is not r-regular for any finite r. However by Theorem 2.4 we know the minimizer for (2.74) must be of the form

$$S_{proj}\left(\left\{\sum_{i=1}^{m} L_{x_i}^* z_i : z_i \in \mathbb{R}\right\}\right) = S_{proj}\left(\left\{\sum_{i=1}^{m} \delta_{x_i}(\cdot) z_i : z_i \in \mathbb{R}\right\}\right) = \left\{\sum_{i=1}^{m} \delta_{x_i}(\cdot) z_i : z_i \in \mathbb{R}\right\}$$

The above representer can then be substituted for f in (2.74) and the optimization can be posed as a finite dimension optimization over $\{z_1, \ldots, z_m\}$. Thus (2.74) provides an example of problems where a non r-regular subspace valued map is required and thus was not be covered by previous counterparts of the generalized representer theorem. Also note that the non $r-$regularity of S_{proj} does not lead to an infinite dimensional representer as the dimension of the space is limited by the range of the adjoint.

2.5 Conclusion

We presented here an extension to existing work on generalized representer theorems by extending the result to apply to learning arbitrary Hilbert space-valued function spaces with loss functionals composed with closed, densely defined operators on separable Hilbert spaces. Subspace valued maps with a super additive property were introduced and the property was shown to be necessary and sufficient for preserving a vector space structure for the union extension of a subspace valued map. The assumption of "r-regularity" was removed from the generalized theorem in order to allow more general subspace valued maps and its implications were shown for the ℓ_1 regularization problem in function spaces. The formalism of linear operators and adjoints was introduced into the generalized representer theorem and infinite dimensional representer spaces were treated as part of the result. The ℓ_1 norm was shown to be orthomonotone with respect to a projection based subspace valued map that shows the sparsity inducing nature of the ℓ_1 norm regularizers. An example from regression in a space of stochastic processes was shown to demonstrate the utility of the theorem when dealing with loss functionals on infinite dimensional Hilbert spaces and linear operators from one infinite dimensional Hilbert space to another. Finally, an example from kernel based neural networks was presented to show an approximation scheme based on the representer theorem to a kernel based neural network.

2.6 Appendix

2.6.1 Subspace Valued Maps

Definition 2.12. *(Quasilinear map)*
A subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***quasilinear*** *if*

$$\forall x, y \in \mathcal{H}, \lambda_1, \lambda_2 \in \mathbb{R}, \qquad S(\lambda_1 x + \lambda_2 y) \subseteq S(x) + S(y)$$

For any $A \in \mathcal{V}(\mathcal{H})$, let $S(A) = \cup_{x \in A} S(x)$. Then idempotence can be defined as,

Definition 2.13. *(Idempotent map)*
A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***idempotent*** *if*

$$\forall x \in \mathcal{H}, \qquad S(S(x)) = S(x)$$

Definition 2.14. *($r-$regular maps)*
For some $r \in \mathbb{N}$, we call a map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, ***$r$-regular*** *if*

1. *it is quasilinear and idempotent*
2. *for all $a \in U$, dimension of $S(a)$ is at most r*
3. *$\forall x \in \mathcal{H}$, $x \in S(x)$*

Lemma 2.5. *(Summation of subspace valued maps are super-additive)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a subspace valued map. Let $S_{sup}(A) = \sum_{x \in A} S(x)$. Then for any $A, B \in \mathcal{V}(\mathcal{H})$, we have $S_{sup}(A) + S_{sup}(B) \subseteq S_{sup}(A + B)$.

Proof: *The proof follows directly from the definition of S_{sup}, $S_{sup}(A) + S_{sup}(B) = \sum_{x \in A} S(x) + \sum_{y \in B} S(y) = \sum_{x \in A \cup B} S(x)$. For vector spaces $A, B \in \mathcal{V}(\mathcal{H})$, we must have $A \cup B \subseteq A + B$. Thus $S_{sup}(A) + S_{sup}(B) = \sum_{x \in A \cup B} S(x) \subseteq \sum_{x \in A+B} S(x) = S_{sup}(A + B)$.* □

The following example shows how addition of sets works in practice and shows a non $r-$regular example of a super-additive subspace valued map.

Example 2.7. *(Summation of subspace valued maps are super-additive)*
Let $\mathcal{H} = \ell^2(\mathbb{N})$ be the space of square summable sequences and let $||a||_\ell^2 = (\sum_{i=1}^{\infty} a_i^2)^{1/2}$. Consider the non r-regular subspace valued map $S^{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ given by

$$S^{proj}(a) = \begin{cases} \{\sum_{i=1}^{\infty} \lambda_i \frac{\langle a, \delta_i \rangle_{\mathcal{H}}}{||a||_{\ell^2}} \delta_i : \{\lambda_i\} \in \ell^2(\mathbb{N})\} & , ||a||_{\ell^2} \neq 0, \\ \{0\} & , otherwise \end{cases}$$

where $\delta_i(j) = 1$ for $j = i$ and 0 elsewhere. Let $S^{proj}_{sup}(A) = \sum_{a \in A} S^{proj}(a)$. Consider the subspaces $A = \{\sum_{i=1}^{\infty} \lambda_{2i}\delta_{2i} : \lambda_i \in \ell^2(\mathbb{N})\}$ and $B = \{\sum_{i=1}^{\infty} \lambda_{3i}\delta_{3i} : \lambda_i \in \ell^2(\mathbb{N})\}$. We have $S^{proj}_{sup}(A) = \{\sum_{i=1}^{\infty} \lambda_{2i}\delta_{2i} : \lambda_i \in \ell^2(\mathbb{N})\} = A$ and likewise $S^{proj}_{sup}(B) = B$ and $S^{proj}_{sup}(A+B) = A+B$. $S^{proj}_{sup}(A) + S^{proj}_{sup}(B) = A + B = \{\sum_{i=1}^{\infty} \lambda_{2i}\delta_{2i} + \lambda'_{3i}\delta_{3i} : \lambda_i, \lambda'_i \in \ell^2(\mathbb{N})\} = A + B = S^{proj}_{sup}(A+B)$ (equality trivially implies the inclusion required for super-additivity).

Note that the representers in [9] are given as $\sum_{i=1}^{m} S(w_i)$ for some r-regular subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. The consideration of super-additive subspace valued maps does not lead to any loss of generality as we can consider the map $S_{sup}(A) = \sum_{x \in A} S(x)$ as given by the above lemma as our super-additive subspace valued map and then the representer is equivalently written as $S_{sup}(\text{span}\{w_1, \ldots, w_m\}) = \sum_{i=1}^{m} S(w_i)$. Note also that the super-additivity of S_{sup} does not contradict the sub-additive property of S required by quasi linearity, as we are considering S_{sup} as a new subspace valued map, entirely different from S, thus while S may be sub-additive, its summation S_{sup} is super-additive (in fact additive, in such a case, as shown below).

Lemma 2.6. *(Summation of quasilinear maps is additive)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a quasilinear subspace valued map. Let $S_{sup}(A) = \sum_{x \in A} S(x)$ be the corresponding summation map defined as $S_{sup} : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. Then S_{sup} is additive, i.e., for any $A, B \in \mathcal{V}(\mathcal{H})$, $S(A) + S(B) = S(A + B)$.

Proof: *$S_{sup}(A) + S_{sup}(B) = \sum_{x \in A} S(x) + \sum_{y \in B} S(y) = \sum_{x,y \in A \cup B} S(x) + S(y) \supseteq \sum_{x,y \in A \cup B} S(x + y) = S_{sup}(A + B)$. Thus using a quasilinear S we get $S_{sup}(A) + S_{sup}(B) \supseteq S_{sup}(A + B)$. From Lemma 2.5, we already have $S_{sup}(A) + S_{sup}(B) \subseteq S_{sup}(A + B)$. Thus combining the two results, we have additivity, $S_{sup}(A) + S_{sup}(B) = S_{sup}(A + B)$.* □

Example 2.8. *(A non-idempotent, non-r-regular, subspace valued map)*
Let $E_m = \{e_1, \ldots, e_m\}$ be the standard orthonormal basis for $\mathbb{R}^m$ and $E_{mn} = \{e_{11}, \ldots, e_{mn}\}$ be the standard orthonormal basis for $\mathbb{R}^{m \times n}$. Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}^m$-valued smooth, square integrable polynomial functions supported on $[-1, 1]^n \subseteq \mathbb{R}^n$ with the Legendre polynomials, given as

$$\{p_{ij}e_i \in \mathcal{H} : p_{ij}(x) = c_j \partial^j_{x_i}[(x_i^2 - 1)^j], c_j = (j + 0.5)^{\frac{1}{2}}(2^j j!)^{-1}, j \in \mathbb{N}, e_i \in E_m, x_i = \langle x, e_i \rangle_{\mathbb{R}^n}\}$$

as the orthonormal basis for $\mathcal{H}$, where p_{ij} is a polynomial of order j. Let $\mathcal{Y}$ be the space of $\mathbb{R}^{m\times n}$-valued functions and $\nabla : \mathcal{H} \to \mathcal{Y}$ be the Jacobian operator, computing the Jacobian for a $\mathbb{R}^m$-valued function. Let $\ell^2(\{1,\dots,m\}\times\mathbb{N})$ be the space of dual indexed sequences $\{\lambda_i j : i \in \{1,\dots,m\}, j \in \mathbb{N}\}$ that are square summable. Consider the subspace valued map $S_{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ given by,

$$S_{proj}(a) = \begin{cases} \left\{\sum_{j=0}^{\infty}\sum_{i=1}^{m} \lambda_{ij} \frac{\langle ae_i, p_{ij}e_i\rangle_{\mathcal{H}} e_i}{||a||_{\mathcal{H}}||p_{ij}||_{\mathcal{H}}} : \lambda_{ij} \in \ell^2(\{1,\dots,m\}\times\mathbb{N})\right\} & , ||a||_{\mathcal{H}} \neq 0 \\ \{0\} & , otherwise \end{cases}$$

Let for a matrix valued function $f \in \mathcal{Y}$, let f_i denote the i^{th} row of the matrix. Let $\nabla\cdot$ be the divergence operator and $\nabla^* : \mathcal{Y} \to \mathcal{H}$ be the adjoint operator to ∇, given as $\nabla^* f = -(\nabla\cdot f_1, \dots, \nabla\cdot f_m)$. A subspace valued map $S' : \mathcal{Y} \to \mathcal{V}(\mathcal{Y})$ is then induced by the jacobian operator ∇ given as,

$$S'(f) = \nabla(S_{proj}(\nabla^* f))$$

Let $f_n \in \mathcal{Y}$ denote a polynomial of maximum order n. Then note that $S'(f_n)$ contains polynomials of order at most $n-2$. Thus clearly $f_n \notin S'(f_n)$. Thus S' is not inclusive.

Also $S'(S'(f_n))$ contains polynomials of order at most $n-4$ and thus $S'(S'(f_n)) \neq S'(f_n)$, implying S' is not idempotent. Also in general $f \in \mathcal{Y}$ can be an infinite order polynomial and thus the dimension of $S'(f)$ can be infinity.

Note that $S' : \mathcal{Y} \to \mathcal{V}(\mathcal{Y})$ is however a quasilinear, super-additive subspace valued map and can still be used to establish a representer theorem, provided it is range preserving with respect to the L being used with the loss functional. An example of such an operator would be the smooth kernel of an RKHS defined on $\mathcal{H}$. Since range(L^*) *then contains polynomials of order upto infinity, $\mathcal{N}_L^{\perp} \subseteq S'(\mathcal{N}_L^{\perp})$.*

Part II

Variational problems in model predictive control

Chapter 3

Manifolds Learning for Path Following, Obstacle Avoidance MPC

A novel manifold learning approach is presented to incorporate computationally efficient obstacle avoidance constraints in optimal control algorithms. The method presented provides a significant computational benefit by reducing the number of constraints required to avoid N obstacles from linear complexity $O(N)$ in traditional obstacle avoidance methods to a constant complexity $O(1)$. The application to autonomous driving problems is demonstrated by incorporation of the manifold constraints into optimal trajectory planning and tracking model predictive control algorithms in the presence of static and dynamic obstacles.

3.1 Introduction

Autonomous driving is an important application domain where obstacle avoidance is required in combination with optimal path planning and control. A wide range of methods including dynamic programming, numerical optimal control, MPC and randomized methods like RRT and A* have been used for motion planning and control of autonomous vehicles in presence of obstacles (e.g. [34, 35, 36, 37]).

The approaches to obstacle avoidance can be broadly separated into two classes based on the obstacle/environment representation used. We will call these: (i) an obstacle centric approach and (ii) an environment centric approach. In an obstacle centric approach each obstacle in the environment is represented using a geometric description of its shape and constraints are imposed to ensure that the vehicle geometry does not collide with any obstacle geometry. In the environment centric approach a geometric description is directly extracted for a feasible region of movement from sensor data without reference to individual obstacle shapes. Constraints are then imposed such that the vehicle geometry remains inside the feasible region to avoid any collisions.

Examples of the obstacle centric approach can be found in works like [34, 38, 39, 40, 41, 42, 43, 44, 45, 46, 35, 36, 47] while the environment centric approach can be found in [48, 49, 37, 50, 51, 52].

Polyhedral obstacle and vehicle geometries are used by [34, 40, 39, 38] with hyperplane separation constraints to impose obstacle avoidance in a nonlinear model predictive control (NMPC) scheme. [35] uses circular obstacles with dynamic programming while [36] uses an RRT* algorithm with polyhedral obstacles. [45, 46] define a potential field for spherical obstacles while [47, 43, 44] use a mixed integer approach for spatially varying constraints for polyhedral obstacles.

The obstacle centric approach imposes an $O(N)$ complexity in the number of constraints if N obstacles are present. In particular for algorithms planning over a horizon length H, $O(NH)$ constraints are needed. This dependence on N creates a problem for real time applications where N can be large and more importantly, can change dynamically, requiring an online update of the optimal control problem structure.

The environment centric approach seeks to circumvent this dependence on N by constructing directly a representation for the feasible region from sensor data. [48, 49] use a Support Vector Machine to learn a non-convex environment representation and perform RRT within the learned region. [37] uses a circular region around the vehicle and LIDAR measurements to partition the circle into sectors not containing any obstacles. A multiphase NMPC scheme is then used to plan trajectories within this circle. [50, 51, 52] use a deep neural network to learn a mapping from features observed in video data of the road to the steering action applied. The features typically correspond to boundaries or markers for the feasible region (e.g. [53]). The approach however does not combine with optimal planning or control and may not always guarantee obstacle avoidance depending on the quality of training and network architecture used.

We present here a manifold learning algorithm to learn feasible environment representations, for an environment centric approach of obstacle avoidance in optimal control methods. The number of constraints introduced is independent of the number of obstacles ($O(1)$ complexity or $O(H)$ for horizon H), while introducing constraints of comparable computational complexity to the linear hyperplane constraints.

The chapter is structured as follows: Section 3.2 presents the manifold learning algorithm and its use for obstacle avoidance in a general optimal control method. Section 3.3 discusses an optimal trajectory planning and path following NMPC scheme for a car parking scenario incorporating the manifold constraints to avoid static and dynamic obstacles in a complex environment. Numerical studies are presented in Section 3.4. Section 3.5 concludes the paper with a few remarks on the method presented and future directions.

Notation

Throughout this chapter, let $\mathcal{H}$ be a Hilbert space of 2π-periodic functions mapping $[0, 2\pi)$ to $\mathbb{R}^2$, with orthonormal basis from a subset of $\cup_{k\geq 1, k\in\mathbb{N}, e_i\in\{(1,0),(0,1)\}}(\{e_i \cos kt\} \cup \{e_i \sin kt\})$. Let $\phi : \mathbb{R}^2 \to [0, 2\pi)$, defined as

$$\phi([x, y]) := \text{arctan2}(y, x)$$

give the angular coordinate of a point in $\mathbb{R}^2$. The angular coordinate for the point $(0, 0)$ is taken to be 0, i.e. $\phi([0, 0]) = 0$.

Let $||\cdot||_{\mathcal{H}}$, $\langle\cdot\rangle_{\mathcal{H}}$ and $||\cdot||_{\mathbb{R}^2}$, $\langle\cdot\rangle_{\mathbb{R}^2}$ be the standard 2-norm and inner product on $\mathcal{H}$ and $\mathbb{R}^2$ respectively. Let $\times$ be the cross product in $\mathbb{R}^2$.

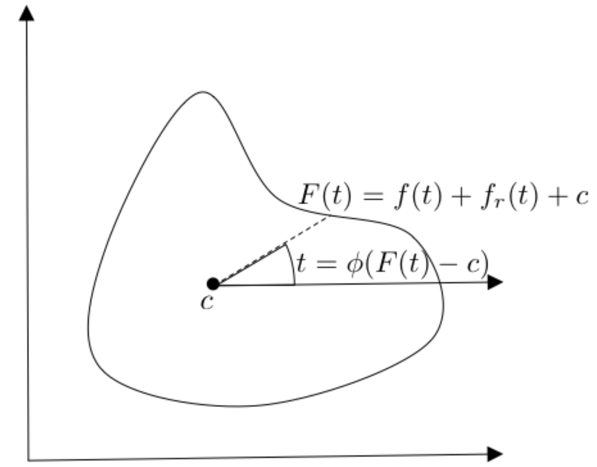


Figure 3.1: A star shaped manifold $\mathcal{M}_c$ centered at $c \in \mathbb{R}^2$. The manifold is parameterized by $t \in [0, 2\pi)$ and satisfies the constraint $t = \phi(F(t) - c)$ for all $t \in [0, 2\pi)$.

3.2 Manifold learning

Manifold learning provides a means to learn a representation for a surface of lower dimensions embedded in a higher dimension space. Several algorithms for manifold learning like local linear embedding, principle component analysis, local tangent space alignment have been presented in machine learning literature. A thorough survey of such methods is available in [54]. The manifold learning methods from [54], aim to learn a surface representation that best fits a given sample data set. We present here a manifold learning algorithm that provides a manifold surface that best fits the given data with two additional constraints. Firstly, the learned manifold is a 2π-periodic surface embedded in $\mathbb{R}^2$ (i.e. a one dimensional manifold) whose interior is a star shaped set. Secondly, all provided samples strictly lie outside or on the boundary of the manifold.

We then present the use of such a learned manifold for an environment centric approach to obstacle avoidance in optimal control. We start by defining the notion of a star shaped manifold.

Definition 3.1. *(Star-shaped Manifold)*
Let c be any point in $\mathbb{R}^2$. Let $f \in \mathcal{H}$ be a 2π periodic function. Let $f_r \in \mathcal{H}$ be defined as $f_r(t) = (r\cos(t), r\sin(t))$ for some fixed $r > 0$ and $\mathcal{M}_c := \{F(t) := (f(t) + f_r(t) + c) : t \in [0, 2\pi)\}$ be a closed curve of points in $\mathbb{R}^2$. Then, for all $t \in [0, 2\pi)$ if $\phi(F(t) - c) = t$ then we say $\mathcal{M}_c$ is a star shaped $\mathcal{S}^1$ isomorphic manifold centered at c.

Note that this definition for star shaped manifold is non-standard and refers to the idea that the interior of such a closed curve will be a star shaped set. For a star shaped manifold $\mathcal{M}_c$, let the interior be defined as $\text{int}(\mathcal{M}_c) := \{p \in \mathbb{R}^2 : \forall m \in [0, 1],\ mp + (1 - m)c \notin \mathcal{M}_c\}$, i.e. the set of points p in $\mathbb{R}^2$ such that a straight line connecting p to c has no intersection with the closed curve of the manifold $\mathcal{M}_c$.

Note that not all 2π-periodic functions $f \in \mathcal{H}$ will represent a star shaped manifold and that the shape of the interior changes as we change the function f. Define

$$L_t f := F(t) = f(t) + f_r(t) + c \tag{3.1}$$

as the affine operator from $\mathcal{H} \to \mathbb{R}^2$ for each fixed t. Similarly define

$$T_t f = \partial_t f(t) + \partial_t f_r(t) \tag{3.2}$$

and

$$N_t f := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} T_t f \tag{3.3}$$

giving a tangent and normal respectively to the manifold at t.

3.2.1 Learning the manifold

Given a point cloud $\mathcal{P}$ of data in $\mathbb{R}^2$, Theorem 3.1 below describes the means to learning a star shaped manifold $\mathcal{M}_c$ such that all obstacle points lie outside the area enclosed by $\mathcal{M}_c$, i.e. $\operatorname{int}(\mathcal{M}_c) \cap \mathcal{P} = \varnothing$.

Theorem 3.1. *Let $\mathcal{P} := \{p_i : i = 1, \dots, M, p_i \in \mathbb{R}^2\}$ be a point cloud of data coordinates and for any $c \in \mathbb{R}^2 \backslash \mathcal{P}$ be some point not included in $\mathcal{P}$. Then a minimizer to the variational problem*

$$\begin{aligned} f_{opt} = \underset{f \in \mathcal{H}}{\arg\min} ||f||^2_{\mathcal{H}} & \qquad (3.4) \\ s.t.\ \forall i \in \{1, \dots, M\},\ t_i = \phi(p_i - c) & \\ (L_{t_i} f - c) \times N_{t_i} f_r = 0 & \qquad (3.4a) \\ \langle p_i - L_{t_i} f, N_{t_i} f_r \rangle_{\mathbb{R}^2} \geq 0 & \qquad (3.4b) \end{aligned}$$

defines a star shaped $\mathcal{S}^1$ isomorphic manifold centered at $c \in \mathbb{R}^2$,

$$\mathcal{M}_c := \{F_{opt}(t) := f_{opt}(t) + f_r(t) + c : t \in [0, 2\pi)\}$$

with $\operatorname{int}(\mathcal{M}_c) \cap \mathcal{P} = \varnothing$.

The proof for the theorem relies on the following two lemmas.

Lemma 3.1. *Let $\mathcal{H}_* := \{f \in \mathcal{H} : \forall t \in [0, 2\pi), \phi(L_t f - c) = t\}$ be the subset of curves in $\mathcal{H}$ that lead to a star shaped manifold. Define*

$$\pi_f = \underset{\hat{f} \in \mathcal{H}_*}{\arg\min}\, ||f - \hat{f}||_{\mathcal{H}}$$

as the projection of f to $\mathcal{H}_*$. Then

(i) $\mathcal{H}_*$ is a closed convex set in $\mathcal{H}$ and

(ii) $\forall f \in \mathcal{H} \backslash \mathcal{H}_*$, $||\pi_f||_{\mathcal{H}} < ||f||_{\mathcal{H}}$

Proof: Firslty, note that any $f \in \mathcal{H}_*$ must be of the form $f(t) = a(t)(\cos(t), \sin(t))$, for some 2π-periodic function $a : [0, 2\pi) \to [0, \infty)$ in $L^2([0, 2\pi))$. $f(t)$ must take this form, because the angle $\phi(L_t f - c) = t$ is given for any $f \in \mathcal{H}_*$. Also the function $a(\cdot)$ must be in $L^2([0, 2\pi))$ to ensure that $||f||_{\mathcal{H}} = ||a(\cdot)(\cos(\cdot), \sin(\cdot))||_{\mathcal{H}} < \infty$. Thus for all $f_1, f_2 \in \mathcal{H}_*$, there exist functions $a_1 : [0, 2\pi) \to [0, \infty)$ and $a_2 : [0, 2\pi) \to [0, \infty)$ such that $f_1(t) = a_1(t)(\cos t, \sin(t))$ and $f_2(t) = a_2(t)(\cos t, \sin(t))$.

Then for all $\alpha, \beta \in [0, \infty)$, $\alpha f_1 + \beta f_2 = (a_1(t) + a_2(t))(\cos(t), \sin(t)) \in \mathcal{H}_*$. As a result for any $\alpha \in [0, 1]$, $\beta = (1 - \alpha)$ and $f_1, f_2 \in \mathcal{H}_*$, we have $\alpha f_1 + (1 - \alpha) f_2 \in \mathcal{H}_*$, implying $\mathcal{H}_*$ is a convex set.

Further for any converging sequence $f_n \in \mathcal{H}_*$, we have a corresponding converging sequence $a_n \in L^2([0, 2\pi))$. Since $L^2([0, 2\pi))$ is a closed Hilbert space (by the Riesz-Fischer theorem), a_n must converge to a point $a \in L^2([0, 2\pi))$ and thus f_n converges to a f in $\mathcal{H}_*$, implying $\mathcal{H}_*$ is closed.

Thus we have shown the first statement ($\mathcal{H}_*$ is a closed convex set).

Then by the Hilbert Projection Theorem [55, Theorem 1.2], $\pi_f \in \mathcal{H}_*$ and $(f - \pi_f) \in \mathcal{H}_*^{\perp}$. Thus $||\pi_f||^2_{\mathcal{H}} + ||f - \pi_f||^2_{\mathcal{H}} = ||f||^2_{\mathcal{H}} \implies$ (ii). □

Lemma 3.2. *Let $f \in \mathcal{H}$ be any feasible solution to (3.4). Then $\pi_f = \arg\min_{\hat{f} \in \mathcal{H}_*} ||f - \hat{f}||$ is also feasible for (3.4).*

Proof: Let $\mathcal{T} = \{t : t = \phi(p_i - c), p_i \in \mathcal{P}\}$ be all the t_is at which constraints in (3.4) are imposed. Note that normal to the circle f_r at any $t \in [0, 2\pi)$, is given by $N_t f_r = (r \cos t, r \sin t)$. Also, as argued in Lemma 3.1, $\pi_f \in \mathcal{H}_*$ must be of the form $\pi_f(t) = a(t)(\cos(t), \sin(t))$ for some 2π-periodic function $a : [0, 2\pi) \to [0, \infty)$ in $L^2([0, 2\pi))$. Thus π_f trivially satisfies (3.4a) for all $t_i \in \mathcal{T}$. Also since both $f(t_i)$ and $\pi_f(t_i)$ satisfy (3.4a), both are in the span of $N_{t_i} f_r$. We can then claim $f(t_i) = \pi_f(t_i)$ for all $t_i \in \mathcal{T}$ as follows.

Suppose $f(t_i) \neq \pi_f(t_i)$ for some $t_i \in \mathcal{T}$, then there must exist a $g \neq 0$ in $\mathcal{H}_*$ and a $g^{\perp} \in \mathcal{H}_*^{\perp}$ such that $f = \pi_f + g + g^{\perp}$. Also since for all $t_i \in \mathcal{T}$, $f(t_i), \pi_f(t_i), g(t_i)$ are all co-linear (satisfying (3.4a)), $g^{\perp}(t_i) = 0$ for all $t_i \in \mathcal{T}$. Then $\pi_f = \arg\min_{\hat{f} \in \mathcal{H}_*} ||f - \hat{f}||_{\mathcal{H}} = \arg\min_{\hat{f} \in \mathcal{H}_*} ||\pi_f + g + g^{\perp} - \hat{f}||_{\mathcal{H}} = \pi_f + g$. But this implies $g = 0$ and hence a contradicts the assumption $f(t_i) \neq \pi_f(t_i)$ for some $t_i \in \mathcal{T}$. Thus for all $t_i \in \mathcal{T}$, we must have $f(t_i) = \pi_f(t_i)$ and thus $L_{t_i} \pi_f = L_{t_i} f$ and π_f satisfies (3.4b) as well. □

The proof for Theorem 1 then follows,

Proof: [Theorem 1] Note that by Lemma 3.2 for any feasible solution $f \in \mathcal{H} \backslash \mathcal{H}_*$, there exists the projection $\pi_f \in \mathcal{H}_*$ as a feasible star shaped solution. Further by Lemma 3.1, $||\pi_f||_{\mathcal{H}} < ||f||_{\mathcal{H}}$. Thus the minimum norm solution $\mathcal{M}_c$ must be star shaped. To show $\text{int}(\mathcal{M}_c) \cap \mathcal{P} = \varnothing$, note that (3.4a) enforces $L_{t_i} f_{opt} - c$ to be in the span of the outward pointing normal $N_{t_i} f_r$, while (3.4b) enforces that the vector pointing from $L_{t_i} f_{opt}$ to p_i is also outward pointing. This ensures that $L_{t_i} f_{opt}$ lies inside the line segment joining c and p_i and since $\mathcal{M}_c$ is star shaped, $\text{int}(\mathcal{M}_c) \cap \mathcal{P} = \varnothing$. Finally, note that for all $c \in \mathbb{R}^2 \backslash \mathcal{P}$ there exists an open neighborhood of c, call it $\text{int}\,(\mathcal{M}_{feas})$, such that $\text{int}\,(\mathcal{M}_{feas}) \cap \mathcal{P} = \varnothing$ (by virtue of Hausdorff separability of $\mathbb{R}^2$). Thus for all $c \notin \mathcal{P}$, there exists a feasible solution to (3.4). □

Note that the minimization problem in (3.4) can be an infinite dimensional one and that the proofs did not rely on any particular definition for the inner product (thus the theorems hold for any inner product definition on $\mathcal{H}$). The infinite dimensional problem is reduced to a finite dimensional one by limiting $\mathcal{H}$ to a space generated by finitely many *sin* and *cosine* basis. Then $f \in \mathcal{H}$ takes the form $f(t) = \sum_{k=0}^{K} a_k \cos kt + b_k \sin kt$ with a finite K and $a_k, b_k \in \mathbb{R}^2$ become the decision variables for (3.4). Another approach to reducing (3.4) to a finite dimensional problem while not reducing $\mathcal{H}$ to a finite basis is to use a reproducing kernel hilbert space $\mathcal{H}$ with a periodic kernel. We avoid this approach in the present work as it would incur a $O(M)$ complexity (M being the size of the point cloud) in evaluating $L_t f_{opt}$ (instead of $O(K)$, K being the size of the basis) making obstacle checking more expensive in Theorem 3.2.

3.2.2 Using the manifold for planning and control

Theorem 3.2. *Let $\mathcal{M}_c$ be the optimal manifold given by Theorem 3.1 with a center $c \in \mathbb{R}^2$. Then a point $p \in \mathbb{R}^2$ is contained in its interior,* $\mathrm{int}(\mathcal{M}_c)$*, if and only if* (3.5) *is satisfied, i.e.,*

$$p \in \mathrm{int}(\mathcal{M}_c) \iff \langle p - L_v f_{opt}, N_v f_r \rangle_{\mathbb{R}^2} \leq 0 \quad \textit{for} \quad v = \phi(p - c) \tag{3.5}$$

Theorem 3.2 provides a simple inequality check for any point being contained in a star shaped manifold. Thus for any point $p \in \mathbb{R}^2$ in order to check for containment in $\mathrm{int}(\mathcal{M}_c)$ a single inequality suffices. To include such a constraint for obstacle avoidance in an optimal control problem simply include (3.5) for each p that needs to be checked. Thus for a horizon H optimal control algorithm where the state for each of the H steps is enforced to be feasible the number of constraints included are of the order $O(H)$. Section 3.3 describes this process in more detail.

3.3 Optimal control and manifold constraints

Three different optimal planning and control algorithms are presented below that are used to accomplish different tasks for an autonomous car parking scenario, using the manifold constraints from Theorem 3.2. Section 3.3.1 describes a corridor planning algorithm over a graph of manifolds using a dynamic programming approach. The corridor plan is then used in Section 3.3.2 to solve a N-phase free end time, numerical optimal control problem for planning a trajectory for the vehicle to move within the free space described by the corridor, while accounting for the vehicle dynamics. The planned trajectory is used as a reference path and a path-following model predictive controller is described in Section 3.3.3 to account for dynamic obstacles and real-time control requirements.

3.3.1 Corridor planning

Given a set of point cloud information pertaining to locations of obstacles in an environment, a single star shaped manifold may not be enough to adequately represent the entire free space configuration in which the vehicle can move. Figure 3.3 shows an example of such an environment (with the point cloud shown in blue). A collection of manifolds (shown in faded black) are then learned with different centers in order to cover the entire free space of interest for the vehicle movement. The collection of manifolds is constructed to ensure that no manifold has a disjoint interior from the rest of the collection and as such there is a path connecting any two manifolds in the collection going through manifolds within the collection.

A undirected graph $\mathcal{G}$ of manifolds is then given by such a collection with each nodes in the graph representing a manifold from the collection and an edge between two nodes indicating that

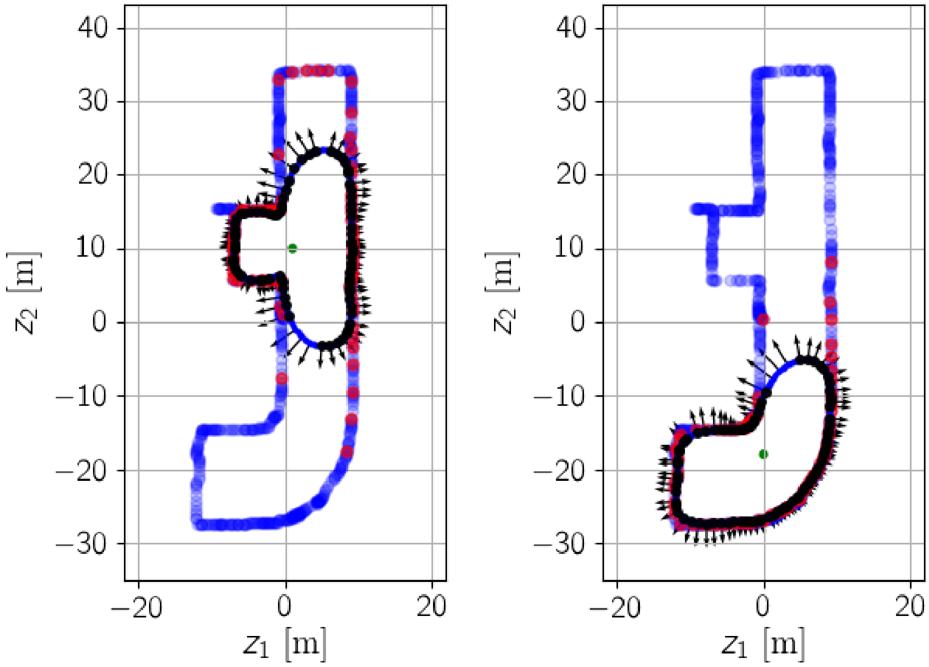


Figure 3.2: Learned Manifolds and Normal Fields. (Green point - center of manifold, red points - point cloud visible from manifold center, blue points - points invisible/out of sensor range from center, black - surface of the manifold, arrows - outward pointing normals on the manifold surface)

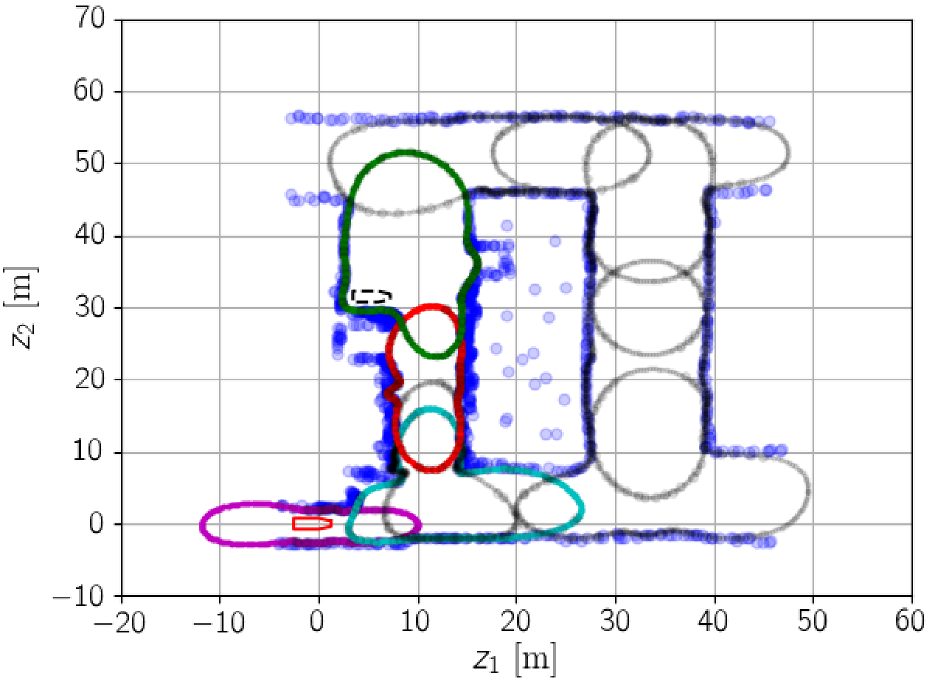


Figure 3.3: Corridor Planning over Manifolds. The shortest sequence of manifolds R_{seq} colored in magenta, cyan, red and green from start to goal. Unused manifolds from $\mathcal{G}$ are in faded black. The car to be controlled is plotted in red and the desired target state is shown with a dashed black profile.

the interiors of the manifolds has non-empty intersection. The complete point cloud of static obstacle points from which the manifolds are learned is denoted as $\mathcal{O}_{static}$. Figure 3.3 shows an example of such an $\mathcal{O}_{static}$ and $\mathcal{G}$. A unit weight is assigned to each edge for simplicity (although other weighting schemes are also possible). Further all manifolds are learned with Theorem 3.1 so that $(\cup_{\mathcal{M}\in\mathcal{G}} \operatorname{int}(\mathcal{M})) \cap \mathcal{O}_{static} = \varnothing$.

Then, given a desired starting and end point, p_{start} and p_{end} respectively, for the vehicle in $\mathbb{R}^2$, we can find the shortest sequence of manifolds in $\mathcal{G}$ connecting p_{start} to p_{end} using a dynamic programming algorithm. The process for constructing the shortest sequence is a standard dynamic programming algorithm is as shown in Algorithm 1. The containment check for any point p in a manifold in Algorithm 1 can be done using Theorem 3.2.

Algorithm 1 Dynamic programming over a graph for corridor planning

Input: graph of manifolds: $\mathcal{G}$, start point: p_{start}, end point: p_{end}

Algorithm:

First setup the costs for traversing the graph from any point to p_{end} as follows:

(i) Let all manifolds in $\mathcal{G}$ be assigned an infinite cost.

(ii) Find all manifolds in $\mathcal{G}$ containing p_{end}, call the set of such manifolds $\mathcal{A}_0$ and set their cost to 0. Set the iteration counter $l = 0$.

(iii) Let, the set of immediate neighbors to $\mathcal{A}_l$ with cost equal to ∞ be called $\mathcal{A}_{l+1}$. If $\mathcal{A}_l$ is an empty set, then terminate. Else, set the cost of all manifolds in $\mathcal{A}_l = l + 1$. Set the iteration counter l to $l + 1$.

(iv) Repeat (iii) till termination. (Note that the steps terminate since we have finitely many nodes in $\mathcal{G}$)

Next find a shortest sequence of manifolds going from p_{start} to p_{end}

(i) Find all nodes in $\mathcal{G}$ containing the point p_{start}, call the set $\mathcal{B}_0$. Set $\mathcal{M}_{c_1}$ as the manifold in $\mathcal{B}_0$ with the lowest assigned cost. Set the iteration counter to $l = 1$.

(ii) Find an immediate neighbor of $\mathcal{M}_{c_l}$ with minimum cost and set $\mathcal{M}_{c_{l+1}}$ to that neighbor. If the cost of $\mathcal{M}_{c_{l+1}}$ is 0, terminate. Else, set the iteration counter l to $l + 1$.

(iii) Repeat (ii) till termination.

Assuming the iteration terminates of the N^{th} step, we have the sequence of manifolds $\{\mathcal{M}_{c_1}, \dots, \mathcal{M}_{c_N}\}$ giving the minimum cost for traversing the graph from p_{start} to p_{end}

Output: $\{\mathcal{M}_{c_1}, \dots, \mathcal{M}_{c_N}\}$

A path connect subset $\mathcal{C} \subseteq \mathbb{R}^2$ such that $\mathcal{C} \cap \mathcal{O}_{static}$ (i.e. not containing any static obstacle points) is called a corridor in the context of autonomous driving. Such a corridor is given by $\mathcal{M}_{c_1} \cup \mathcal{M}_{c_1} \cup \cdots \cup \mathcal{M}_{c_N}$, because each manifold satisfies $\mathcal{M}_{c_l} \cap \mathcal{O}_{static} = \varnothing$, by virtue of Theorem 3.1.

Algorithm 1 thus provides a fast corridor planning algorithm to find the shortest sequence of manifolds, $R_{seq} = \{\mathcal{M}_{c_1}, \dots, \mathcal{M}_{c_N}\}$, that must be traversed to reach the end position. A multiphase optimal trajectory to traverse R_{seq} is then constructed in section 3.3.2, taking the vehicle dynamics into account.

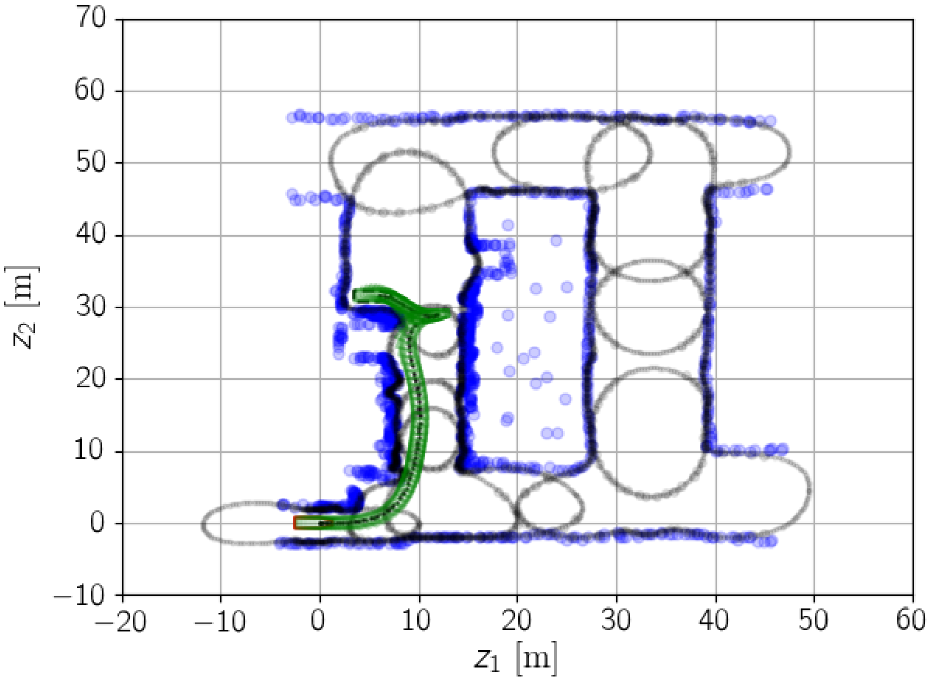


Figure 3.4: Optimal Trajectory Planning over Manifolds. The shortest time trajectory avoiding the corridor plan manifold constraints and adhering to the state dynamics and state-input constraints. The evolution of the car position and orientation is plotted in green over the period of the optimal trajectory.

3.3.2 Optimal trajectory planning

Let $R_{seq} = \{\mathcal{M}_{c_1}, \dots, \mathcal{M}_{c_N}\}$ be the sequence of manifolds given by Algorithm 1 for start and end points, p_{start} and p_{end} respectively. Also let the center of $\mathcal{M}_{c_i}$ be the point $c_i \in \mathbb{R}^2$ for all $i = 1, \dots, N$.

Consider for simplicity, a slip free Dubin's car model (3.6) to describe the non-holonomic vehicle dynamics, with the state $q = (z_1, z_2, \psi, v)$ comprising of (z_1, z_2) giving a coordinate position for the vehicle in $\mathbb{R}^2$, ψ giving a yaw orientation and v giving the car's forward speed. The controls used are a steering input δ and acceleration a. k_δ, k_{acc} are known constants corresponding to the steering and acceleration input gains.

$$\dot{q} = \big(v \cos\psi, v \sin\psi, k_\delta \cdot v \cdot \delta, k_{acc} \cdot a\big)^T \tag{3.6}$$

Assume, now, that the initial state of the vehicle dynamics is given as q_{start}, such that the corresponding position of the vehicle in $\mathbb{R}^2$ is p_{start} and the desired end state q_{end} for the vehicle is such that the corresponding position is p_{end}, as were used to plan the corridor sequence R_{seq}.

The algorithm presented next is agnostic of the exact form and details of the dynamic model used and we will simply denote the state of the vehicle dynamics by q, the inputs to the vehicle as

u and the dynamics to be given by an ordinary differential equation,

$$\dot{q} = Q(q, u) \tag{3.7}$$

Then, a N-phase optimal trajectory satisfying the non-holonomic vehicle dynamics and obstacle avoidance constraints can be generated as follows.

Let $\mathcal{M}_{c_1}$ to $\mathcal{M}_{c_N}$ denote the N manifolds in R_{seq}. Let $q(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m$ be the state and input at time t for the vehicle and $\dot{q}(t) = Q(q(t), u(t))$ be the vehicle dynamics. q_{start} is given as the initial state of the vehicle at time $t = 0$ and q_{end} is the desired end state.

Let $p_j(t) = \Omega_j(q(t))$, $j = 1, \ldots, k$ for some finite k denote a collection points on the vehicle geometry for which to enforce obstacle avoidance, given by selection functions $\Omega_j : \mathbb{R}^n \to \mathbb{R}^2$ (see Definition 3.2 for an example of Ω_j). Also for purposes of brevity we will denote the fact that

$$p_j(t) = \Omega_j(q(t)) \in \text{int}(\mathcal{M}), \quad \forall j = 1, \ldots, k \iff q(t) \in \text{int}(\mathcal{M})$$

by the abuse of notation $q(t) \in \text{int}(\mathcal{M})$.

By construction, R_{seq} is such that $q_{start} \in \text{int}(\mathcal{M}_{c_1})$ and $q_{end} \in \text{int}\left(\mathcal{M}_{c_N}\right)$. Let the state and input be bounded in box constraints $\mathcal{X}_{box}, \mathcal{U}_{box}$ respectively. Let $t_f \in [0, \infty)$ be a free end time for the trajectory and let $t_f^i \in [0, \infty)$ be the time for first exit from $\text{int}(\mathcal{M}_{c_i})$ for $i \in \{1, \ldots, N-1\}$. For notational convenience, let $t_f^0 = 0$ and $t_f^N = t_f$. A general N-phase optimal control problem is described in Algorithm 2 below, variants of which lead to the optimal trajectory generation and MPC path following algorithms to be presented later.

Algorithm 2 N-phase Optimal Control

Input: initial state: q_0, goal state: q_f, manifolds: $\{\mathcal{M}_{c_1}, \ldots, \mathcal{M}_{c_N}\}$ and smooth, strongly convex cost functionals: $\ell : \mathbb{R}^n \times \mathbb{R}^m \to [0, \infty)$, $G : \mathbb{R}^n \to [0, \infty)$

OCP: $q_{opt}, u_{opt}, t^1_{f\,opt}, \ldots, t^N_{f\,opt} :=$

$$\underset{\substack{\hat{q}(\cdot) \in L^2([0,\infty);\mathbb{R}^n) \\ \hat{u}(\cdot) \in L^2([0,\infty);\mathbb{R}^m) \\ t_f^1, \ldots, t_f^N \in [0,\infty)}}{\arg\min} \quad G(\hat{q}(t_f)) + \int_0^{t_f} \ell(\hat{q}(s), \hat{u}(s)) ds + \sum_{i=1}^{N} (t_f^i)^2 \tag{3.8a}$$

$$\text{s.t.} \quad \forall s \in (t_f^{i-1}, t_f^i], i \in \{1, \ldots, N\}, j \in \{1, \ldots, k\} \tag{3.8b}$$

$$\hat{q}(0) = q_0,\ \hat{q}(t_f) = q_f,\ t_f^0 = 0 \tag{3.8c}$$

$$\hat{q}(s) \in \mathcal{X}_{box}, \hat{u}(s) \in \mathcal{U}_{box},\ t_f^i - t_f^{i-1} \geq 0 \tag{3.8d}$$

$$\dot{\hat{q}}(s) = Q(\hat{q}(s), \hat{u}(s)) \tag{3.8e}$$

$$\hat{p}_j(s) := \Omega_j(\hat{q}(s)) \in \text{int}\left(\mathcal{M}_{c_i}\right) \tag{3.8f}$$

Output: $q_{opt},\ u_{opt}, t^1_{f\,opt}, \ldots, t^N_{f\,opt}$

(3.8) describes a general free-end time $N-$phase optimal control problem where the variable t_f^i represents the end time for phase i, $i = 1, \ldots, N$. In each phase of the problem one manifold constraint is active, i.e. $\mathcal{M}_{c_i}$ is active for phase i.

(3.8c) enforces the initial and terminal boundary conditions for the vehicle state and sets the initial $t_f^0 = 0$ for notational convenience of (3.8d). (3.8d) enforces the input and state constraints to be satisfied and states that the switching times should be ordered such that the time at which the manifold is switched from $\mathcal{M}_{c_i}$ to $\mathcal{M}_{c_{i+1}}$ (given by t_f^i) is greater than the switching time t_f^{i-1} when the constraint for $\mathcal{M}_{c_i}$ was first made active. (3.8e) enforces that the solution q_{opt}, u_{opt} satisfy the differential equation for the dynamics considered. (3.8f) enforces that for all times $s \in (t_f^{i-1}, t_f^i]$ the selected points on the vehicle geometry are in the interior of the active manifold $\mathcal{M}_{c_i}$, thus avoiding all obstacles. (3.8f) is equivalent to the inequality constraint given by (3.5).

The optimal reference trajectory q_{ref} and control u_{ref} from q_{start} to q_{end}, given R_{seq} can then be found using Algorithm 3. Note that we are subscripting the optimal solutions as q_{ref} and u_{ref} as these solutions will be used as reference trajectories for a path following model predictive controller in Section 3.3.3.

Algorithm 3 Optimal Trajectory Generation

```
Input: $q_0 = q_{start}$, $q_f = q_{end}$, manifolds: $R_{seq}$, $\ell(\hat{q}(s), \hat{u}(s)) = \gamma||\hat{u}(s)||^2$ for some $\gamma > 0$ and
    $G(\hat{q}(t_f)) = 0$
Solve: OCP (3.8) and get $q_{opt}, u_{opt}$ and $t^i_{f\ opt}$, $i = 1, \ldots, N$
Output: $t_f^{ref} = t^N_{f\ opt}$, $q_{ref} : [0, t_f^{ref}] \to \mathbb{R}^n := q_{opt}$
```

Theorem 3.3. *Let R_{seq}, $\mathcal{X}_{box}$ and $\mathcal{U}_{box}$ be such that the optimization (3.8) is feasible in Algorithm 3. Then the optimal trajectory, q_{ref} is such that for all $t \in [0, t_f^{ref}]$ and all $j \in \{1, \ldots, k\}$, $\Omega_j(q_{ref}(t)) \cap \mathcal{O}_{static} = \varnothing$ and $q_{ref}(t_f^{ref}) = q_{end}$.*

The proof for Theorem 3.3 follows directly from the enforced terminal constraint (3.8c) and manifold constraints (3.8f), which by Theorem 3.1 implies $\Omega_j(q_{ref}(t)) \cap \mathcal{O}_{static} = \varnothing$ in each phase. Thus Theorem 3.3 guarantees that the optimal reference trajectory is such that for all points on the trajectory q_{ref}, selected points of the vehicle geometry are contained in the interior of manifolds in R_{seq}, thus avoiding all known static obstacles in the map.

The next section, describes a real-time path following model predictive controller, using q_{ref} as a reference path in order to address the concerns of fast real time control (as generating an optimal trajectory by (4.6) for ever changing values of N, large values of N can be computationally expensive) and to address the issue that $\mathcal{O}_{static}$ being an offline data set of static obstacle information may not accurately describe the obstacles in the environment. To address the issue of apriori unknown and possibly moving obstacles, we introduce a new dynamic set $\mathcal{O}_{dyn}$ of point cloud data (acquired in real time with a LIDAR like sensor) in the next section and learn a dynamically changing manifold $\mathcal{M}_{c_1(t)}$ centered around a point on the vehicle.

3.3.3 Dynamic obstacles and model predictive control

Let $\mathcal{O}_{dyn}$ be a point cloud of dynamic obstacles not accounted for in $\mathcal{G}$ and let $\mathcal{O} = \mathcal{O}_{static} \cup \mathcal{O}_{dyn}$. While Theorem 3.3 provides an effective method to plan trajectories in presence of static obstacles; for dynamic obstacles we formulate an MPC path following scheme tracking the planned q_{ref} with a reference geometric path to follow. Treating q_{ref} as simply a parametrized geometric path and not a time-bound trajectory allows the controller to move to q_{ref} and follow along q_{ref} at a speed that is feasible for the real vehicle dynamics (as there may be a difference between the real vehicle dynamics and the model used for planning) and for constraints placed by the new dynamic obstacles.

Also since path-following under dynamic obstacles can lead to unforeseeable situations, we address here only the problem in a semi-cooperative setting. Under such a setting, we assume that path q_{ref} is not permanently made infeasible, i.e. we can move to any point to q_{ref} without being hindered by an obstacle permanently (a point may be unreachable for a finite amount of time, but the obstacles will move away in a finite time span, to make the point reachable). In such a setting, we also do not address adversarial obstacles, that are either actively trying to collide with the vehicle or accidentally in a state such that no control action by the MPC controller can avoid collision.

At any time t, let $q(t)$ be the vehicle state. Let $c_1(t) = s_0(q(t))$ be the position of a sensor on the vehicle in $\mathbb{R}^2$, given the state of the car, $q(t)$. Let $\mathcal{M}_{c_1(t)}$ be a dynamic manifold learned using Theorem 3.1 at each time t using new data from the sensor, allowing detection and avoidance of dynamic obstacles in $\mathcal{O}$. Let t_f^{ref} and $q_{ref} : [0, t_f^{ref}] \to \mathbb{R}^n$ be the reference end time and reference geometric curve mapping into the vehicle's state space as given by Algorithm 3.

Let $J : R_{seq} \to \mathbb{R}$ be a map from a manifold in R_{seq} to its cost to go, assigned in the dynamic program from Algorithm 1 to reach the end of the sequence. Then choose a manifold,

$$\mathcal{M}_{c_2(t)} = \underset{\mathcal{M} \in R_{seq}}{\arg\min}\, J(\mathcal{M}) \text{ s.t. } c_1(t) \in \operatorname{int}(\mathcal{M})$$

i.e. $\mathcal{M}_{c_2(t)}$ is the manifold in R_{seq} with the minimum cost to go such that the current sensor position is contained in its interior. Recall that $\mathcal{M}_{c_1(t)}$ is the dynamically learned manifold, centered around the sensor position and thus $\mathcal{M}_{c_2(t)} \cap \mathcal{M}_{c_1(t)} \neq \varnothing$. It is assumed that the initial state of the vehicle in such that $\mathcal{M}_{c_2(t_0)} \cap \mathcal{M}_{c_1(t_0)} \neq \varnothing$ and the recursive feasibility of the MPC shown later will ensure that this remains true for all $t \geq t_0$.

Then for the path following MPC problem considered over a finite horizon T, a heuristic reference terminal state for the horizon $[t, t+T]$ is then chosen as follows.

Let $\mathcal{N}(\mathcal{M}_{c_2(t)})$ be the set of immediate neighbors of $\mathcal{M}_{c_2(t)}$ in R_{seq} including $\mathcal{M}_{c_2(t)}$ itself. Let $t_w(w,t) > 0$ be a positive offset parameter at time and path parameter (t, w), given by

$$t_w(w,t) = \underset{\hat{t}_w \in [0, t_f^{ref}]}{\arg\max}\, \hat{t}_w \text{ s.t. } q_{ref}(w + \hat{t}_w) \in \operatorname{int}(M) \text{ and } M \in \mathcal{N}(\mathcal{M}_{c_2(t)})$$

and let $w_f(w,t) := \min(w + t_w(w,t), t_f^{ref})$ be a forward shift for any $w \in [0, t_f^{ref}]$. Then choose

$$w^*(t) = \underset{w\in[0,t_f^{ref}]}{\arg\min} ||q_{ref}(w) - q(t)|| \text{ s.t. } q_{ref}(w_f(w,t)) \in \cup_{\mathcal{M}\in\mathcal{N}(\mathcal{M}_{c2(t)})} \operatorname{int}(\mathcal{M})$$

This ensures $q_{ref}(w^*(t))$ is the closest point on q_{ref} to the current vehicle state $q(t)$ such that a future point $q_{ref}(w_f(w^*(t),t))$ lies within the neighborhood $\mathcal{N}(\mathcal{M}_{c2(t)})$ and that $q_{ref}(w_f(w^*(t),t))$ is the closest possible point to the end goal that can be selected within the neighborhood $\mathcal{N}(\mathcal{M}_{c2(t)})$. Recall that for any state $q \in \mathbb{R}^n$, we mean by $q \in \operatorname{int}(\mathcal{M})$, that the corresponding selection of vehicle points $p_i \in \mathbb{R}^2$ is in $\operatorname{int}(\mathcal{M})$. Note that this minimization is always feasible, since $\mathcal{M}_{c2(t)}$ belongs to $\mathcal{N}(\mathcal{M}_{c2(t)})$ and it is assumed the current vehicle state is in $\mathcal{M}_{c2(t)}$ and thus one can always trivially choose $q_{ref}(w^*(t))$ and $q_{ref}(w_f(w^*(t),t))$ to be points in $\mathcal{M}_{c2(t)}$ (since $\mathcal{M}_{c2(t)}$ is a manifold chosen at time t from R_{seq} and q_{ref} passes through every manifold in R_{seq} by design).

Now let the heuristic end goal for the MPC over the horizon $[t, t+T]$ be given as

$$q_{end}^{ref}(t) = q_{ref}(w_f(w^*(t),t))$$

and let

$$\mathcal{M}_{c3(t)} = \underset{\mathcal{M}\in\mathcal{N}(\mathcal{M}_{c2(t)})}{\arg\min} J(\mathcal{M}) \text{ s.t. } q_{end}^{ref}(t) \in \operatorname{int}(\mathcal{M})$$

be the manifold in the neighborhood $\mathcal{N}(\mathcal{M}_{c2(t)})$ with minimum cost to go, containing the end goal $q_{end}^{ref}(t)$.

Thus we have three manifolds at each time t; $\mathcal{M}_{c1(t)}$ accounting for new or dynamic obstacles in $\mathcal{O}$, and $\mathcal{M}_{c2(t)}, \mathcal{M}_{c3(t)} \in R_{seq}$, accounting for only static obstacles in $\mathcal{O}_{static}$ for manifolds leading from $q_{ref}(w^*(t))$ to $q_{end}^{ref}(t)$. The following optimal control problem can then be solved at each time t to get a path following NMPC controller.

Algorithm 4 Path Following MPC

Input: $N = 3$, $q_0 = q(t)$, $q_f = q_{end}^{ref}(t)$, manifolds: $\{\mathcal{M}_{c1(t)}, \mathcal{M}_{c2(t)}, \mathcal{M}_{c3(t)}\}$, a safety time margin: $t_{safe} > 0$ and a maximum blocking time: $T_{blocking}$. The cost functionals:

$$\ell(\hat{q}(s), \hat{u}(s)) = \gamma_1 ||\hat{q}(s) - q_{ref}(\min\{w^*(t) + s, w_f(w^*(t),t)\})||^2 + \gamma_2 ||\hat{u}(s)||^2$$

$$G(\hat{q}(t_f)) = \gamma_3 ||\hat{q}(t_f) - q_{end}^{ref}(t)||^2$$

for some constants $\gamma_1, \gamma_2, \gamma_3 > 0$

Solve: OCP (3.8) subject to an additional constraint $T_{blocking} \geq t_f^1 \geq t_{safe}$ and get the optimal solution u_{opt}

Output: Applied control action: $u(t) = u_{opt}(0)$

Algorithm 4 thus proposes a moving horizon version of the free end time, $N-$phase optimal control problem in (4.6) with $N = 3$, tracking a geometric curve q_{ref}, thus giving a path following MPC controller. The MPC controller computes the control signal to follow the geometric reference path as closely as possible starting at $q(t)$ while avoiding the manifold constraints in order to reach a end state $q^{ref}_{end}(t)$ such that $q_{opt}(t^N_{f\ opt}) = q^{ref}_{end}(t)$ (by the imposed terminal constraint in (4.6)). The computed control signal at time t, $u(t)$ is applied to the system, to reach a new state and the optimization problem for Algorithm 4 is resolved again at the new state. Section 3.4 gives more details on the discrete time implementation of such a scheme. With the initial state in (4.6), set as the current state $q(t)$ and the end goal q_f in (4.6) set to the heuristically selected end goal $q^{ref}_{end}(t)$, note that when the true state q_{end} is in $\mathcal{N}(\mathcal{M}_{c_2(t)})$, the heuristic end goal as given above will always be $q^{ref}_{end}(t) = q_{end}$, as it is the closest point in the neighborhood to the end goal. Thus the heuristic end goal over the horizon moves forward towards q_{end} as soon as such a movement is permitted by the obstacle environment, enabling the path following MPC to progress towards the end goal.

The manifolds $\mathcal{M}_{c_1(t)}$ is enforced to ensure that the dynamic obstacles are avoided for all time $[t, t + t^1_f]$. Manifolds $\mathcal{M}_{c_2(t)}$ and $\mathcal{M}_{c_3(t)}$ are used to plan future motion towards the goal once a blocking obstacle has moved away. The switching time $t^1_f > t_{safe}$ ensures that a safety time margin is permitted for the vehicle to come to a halt or reverse its motion to avoid a moving obstacle, during the next iteration of the MPC algorithm. The amount of time the vehicle has to spend within $\mathcal{M}_{c_1(t)}$ before it can proceed with motion through the originally planned manifolds $\mathcal{M}_{c_2(t)}$ and $\mathcal{M}_{c_3(t)}$ is given by t^1_f. If a obstacle is blocking the path of motion for the vehicle then the problem in (4.6) is infeasible as we require for a $t^1_{f\ opt} \le t^N_{f\ opt} < \infty$, $q_{opt}(t^N_{f\ opt}) = q^{ref}_{end}(t)$. Thus, an assumption is made on the obstacles, such that the obstacle will move away in a maximum time $T_{blocking}$ and thus we have $t^f_1 = T_{blocking}$ when the motion is blocked and we plan the motion through $\mathcal{M}_{c_2(t)}$ and $\mathcal{M}_{c_3(t)}$ for the time $[t + T_{blocking}, \infty)$. Note however that if the obstacle does not move away in time $T_{blocking}$, the re-solving of the MPC at the next time iteration again sets $t^f_1 = T_{blocking}$ and thus the MPC can be permanently blocked from making progress by an obstacle and also the controller will not collide with such an obstacle for any time $[0, \infty)$, since there is a safety margin of t_{safe} seconds left from the previous iteration in which the vehicle can be brought to a halt.

The following assumptions on the obstacle environment are made to give formal guarantees on the convergence and recursive feasibility properties for the MPC scheme.

Assumption 3.1.

(i) The dynamic obstacles follow a semi-cooperative policy for their motion such that at any time $t \in [0, \infty]$, the obstacle will remain outside $\mathcal{M}_{c_1(t)}$ for the time interval $[t, t + t_{safe}]$ seconds. Note that this is not exploited by the MPC to behave in a adversarial manner and push obstacles around, since t_{safe} is only a lower bound for t^1_f. This assumption simply means that there is a non-zero safety time margin t_{safe} for the MPC to take a control action such that a collision can be avoided, given the current state of the vehicle.

(ii) Obstacles do not permanently make $q^{ref}_{end}(t)$ unreachable in R_{seq} (i.e. blocking obstacles will eventually move away). (This time may be different from $T_{blocking}$, the assumption is just required to ensure that we do not have infinite iterations of the MPC with $t^1_f = T_{blocking}$ and thus the MPC is prevented from making any progress towards the goal)

(iii) The vehicle dynamics and input constraints are such that the vehicle has a maximum velocity and acceleration/braking such that in t_{safe} seconds it can go from the maximum velocity to zero velocity in $t_{safe}/2$ seconds and the reference trajectory can be tracked with zero position and orientation error for any velocity profile within the limits set by the state and input constraints, given a zero error at the initial state.

Given such assumptions, the following theorem can be established for an MPC scheme for following q_{ref} as a reference path in the presence of dynamic obstacles.

Theorem 3.4. *Given assumption 3.1-(i) and (iii), the closed loop solution $q(t)$ obtained by applying a control signal $u(t) = u_{opt}(0)$ using Algorithm 4 is such that for all $t \in [0, \infty)$ and all $j \in \{1, \ldots, k\}$, $\Omega_j(q(t)) \cap \mathcal{O} = \varnothing$, i.e. selected points on the vehicle geometry avoid all obstacles (dynamic and static) from $\mathcal{O} = \mathcal{O}_{dyn} \cup \mathcal{O}_{static}$.*

Proof: At some time t if (3.8) is feasible then $t^1_{f\,opt}$ is strictly greater than t_{safe} (by the imposed constraint). This implies there is a strictly positive time $t^1_{f\,opt} > t_{safe}$ before any $\Omega_j(q(t))$ exits $\mathrm{int}(\mathcal{M}_{c_1(t)})$. Further by assumption 1-(i), all obstacles are guaranteed to remain outside $\mathrm{int}(\mathcal{M}_{c_1(t)})$ for $[t, t + t_{safe}]$. Thus for all $t' \in [t, t + t_{safe}]$, $\Omega_j(q(t)) \cap \mathcal{O} = \varnothing$. Given the optimal solution at time t, going from $q(t)$ to $q^{end}_{ref}(t)$, for any time $t' \in [t, t + t_{safe}/2]$, Algorithm 4 can be re-solved for which a feasible solution from $q(t')$ to $q^{ref}_{end}(t')$ can be obtained as a motion from $q(t')$ to $q^{end}_{ref}(t)$ (albeit with a different time profile) followed by motion along q_{ref} from $q^{end}_{ref}(t)$ to $q^{end}_{ref}(t')$. (Such a solution exists by virtue of the assumption 3.1-(iii), which state that a given state trajectory is can be tracked with zero error in position and yaw for any velocity profile). Thus (3.8) remains feasible for all $t' \in [t, t + t_{safe}/2]$. By recursively applying this argument for any $t \in [0, \infty)$, (3.8) remains feasible for $[t, t + t_{safe}/2]$ for each $t \in [0, \infty)$. Thus (3.8) remains feasible for all $t \in [0, \infty)$ if (3.8) is feasible at $t = 0$ and $\Omega_j(q(t)) \cap \mathcal{O} = \varnothing$. □

Theorem 3.5. *Given assumption 1-(ii), there exists a finite time t_{end} such that $q(t_{end}) = q_{end}$*

Proof: By assumption 1-(ii), for any time t, $q^{end}_{ref}(t)$ is reachable in some finite time, i.e. there exists a finite time $t' \in [0, \infty)$ such that $q(t') = q^{end}_{ref}(t)$ (note that here by reachable we mean the actual state q reaching the goal $q^{end}_{ref}(t)$ and not just the MPC prediction q_{opt}). Thus there exists a finite time $t' > t$ such that $w^*(t') > w^*(t)$. Thus $q^{ref}_{end}(t')$ is closer to q_{end} along q_{ref} than $q^{ref}_{end}(t)$. Since $w^*(t')$ is upper bound by t^{ref}_f, there must exist a finite t' such that $w^*(t') = t^{ref}_f$, i.e., $q^{ref}_{end}(t') = q_{end}$. Then a finite time $t_{end} > t'$ exists such that $q(t_{end}) = q_{end}$. □

3.4 Numerical results

Figure 3.2 shows the result of learning individual manifolds around different centers using Theorem 3.1. We use $r = 30$ m in f_r and a $\mathcal{H}$ space generated by a finite basis of sine and cosine with $K = 10$. In order to avoid the dependence on a dynamically changing and large M in Theorem 3.1, we preprocess the sensor data to return a single closest point in a sector of resolution θ_{res} by dividing the $[0, 2\pi]$ interval into $N_{part} = 90$ intervals. With N_{part} large enough we are assured that the sensor data accurately enough represents the obstacles. Thus M in Theorem 3.1 is fixed to be N_{part} for fast online optimization. The preprocessed visible points are shown in red in Figure 3.2.

Figure 3.3 shows R_{seq} for a parking scenario with $q_{start} = (0,0,0,0)$ and $q_{end} = (6,31.5,0,0)$. Using this R_{seq} and Theorem 3.3, Figure 3.4 shows the optimal trajectory plan q_{ref} from q_{start} to q_{end}. Figure 3.5 shows the closed loop behavior of the MPC scheme in presence of a dynamic obstacle. A second car (displayed as a green polytope) is added to the environment (not accounted for in $\mathcal{O}_{static}$) and drives out in the opposite direction of the controlled car (displayed as a red polytope). The planned motion at time t ($q_{opt}(\cdot)$) is shown in cyan and the goal state at time t ($q_{end}^{ref}(t)$) is shown as a pink dashed polytope. The final goal state q_{end} is shown as a black dashed polytope. When the new obstacle is encountered, as shown in Figure 3.5b, the path for the vehicle to move forward is blocked. In order to avoid the obstacle (moving in the opposite direction), the car reverses its motion and moves in reverse from Figure 3.5b to 3.5c (the front edge of the car can be seen moving from $z_2 = 20$ m in 3.5b to $z_2 = 17$ m in 3.5c). Eventually space is freed by the moving obstacle (Figure 3.5d) and the car drives forward again to eventually reach the parking state q_{end}.

The slip free Dubin's car model from (3.6) is used for the non-holonomic vehicle with the state $q = (z_1, z_2, \phi, v)$ comprising of the z_1, z_2 coordinate position in the ground plane, yaw orientation ψ and car's forward speed v. The controls used are a steering input δ and acceleration a.

Definition 3.2. *(Vehicle Geometry)*
For describing the vehicle geometry we use an elongated hexagon for the car shape projected on the $z_1 - z_2$ ground plane. Nine vertices are placed on the hexagon corners and side and backward face bisectors. The sensor for the car is placed at its center. The corresponding selection functions Ω_j $(j = 0, \ldots, 9)$ are defined as $\Omega_j(q(t)) = (p_1^j \cos\phi - p_2^j \sin\phi + z_1(t), p_1^j \sin\phi + p_2^j \sin\phi + z_2(t))^T$ if (p_1^j, p_2^j) are coordinates of the point when the car state is $(0,0,0,0)^T$.

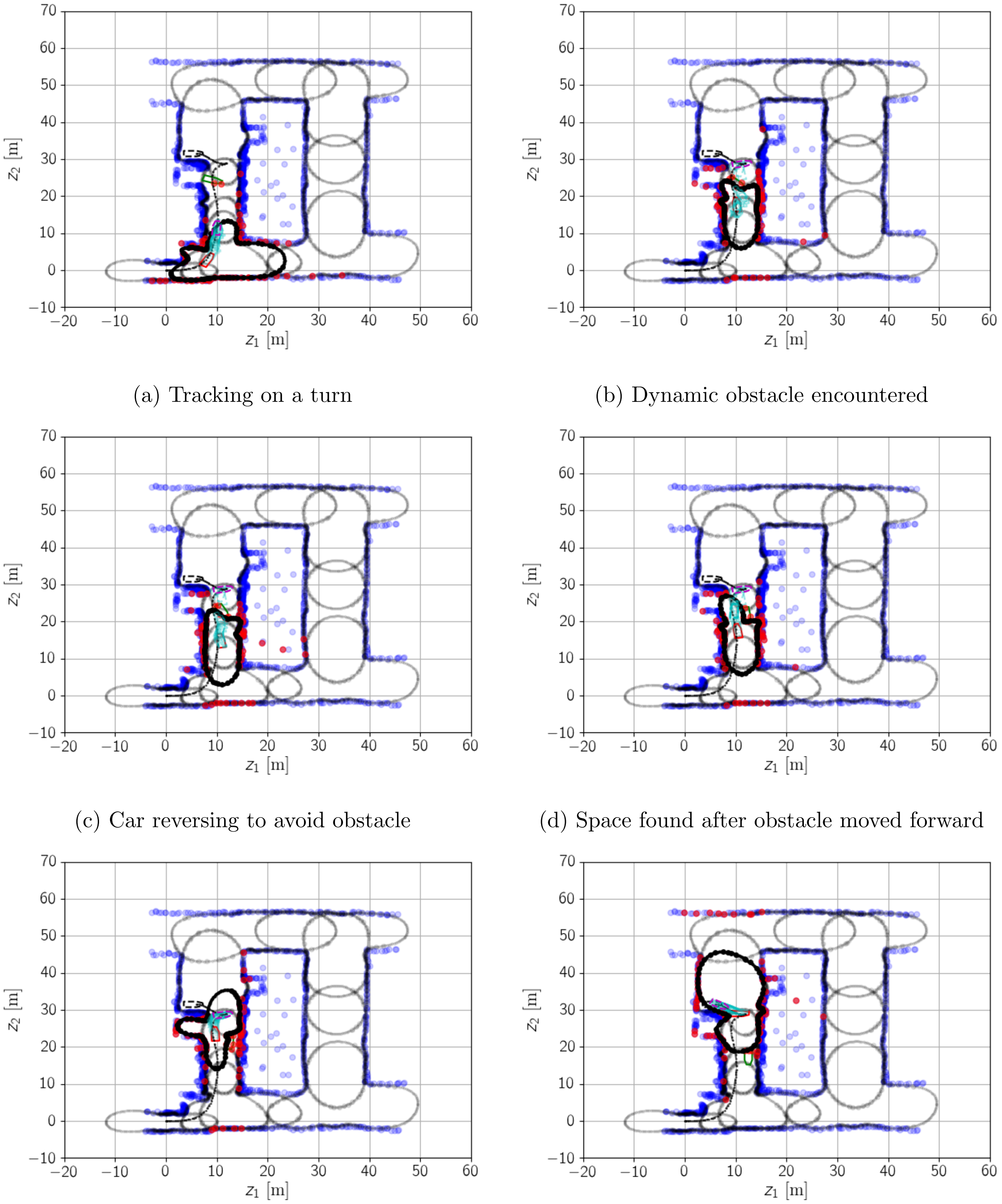
For solving the free end time optimal control problem in (3.8), we use a time scaling input as a decision variable along with time scaled vehicle dynamics. The continuous time problem is converted to discrete time using a multiple shooting approach with RK4 integration of step-size: 0.1.

The control and state bounds imposed were $\mathcal{U}_{box} := \{(-1,-1)^T \leq u \leq (1,1)^T\}$, $\mathcal{X}_{box} := \{(-\infty,-\infty,-\infty,-1) \leq x \leq (+\infty,+\infty,+\infty,4)\}$ and $t_{safe} = 0.05$ seconds with $k_\delta = 0.4$, $k_{acc} = 5$.

On an Intel Core i7, 2.8 GHz processor using an interior point solver (ipopt) the average solve times for the algorithms were as follows: Manifold Learning: 200 ms, free end 4−phase time optimal trajectory generation: 34.9 sec and the free end time 3−phase path following MPC: 754 ms. Note also that the longer solve times for the MPC and optimal trajectory generation are to be expected as we are solving a multiphase, free end time optimal control problem, which is typically computationally expensive compared to a trajectory tracking like approach. Faster implementation schemes thus need to be explored to make the MPC controller compatible for real time implementation.

3.5 Conclusion

A novel manifold learning approach was presented to learn representations of complex and dynamic obstacle environments and to provide computationally tractable constraints for optimal control algorithms. The use of the manifold constraints for obstacle detection and avoidance was demonstrated with three variants of optimal control problems; a dynamic programming approach for corridor planning, an optimal trajectory generation problem and a nonlinear MPC problem for path following. The three variants were deployed to drive a vehicle in a car parking scenario in presence of static and dynamic obstacles. Recursive feasibility of the MPC under semi-cooperative obstacle movements was shown. MPC schemes taking into account obstacle speed and movement plan or adversarial obstacles remains a subject for future investigation and was not covered in this work.



(a) Tracking on a turn

(b) Dynamic obstacle encountered

(c) Car reversing to avoid obstacle

(d) Space found after obstacle moved forward

(e) Transition into the reverse parking position

(f) Executing the reverse parking

Figure 3.5: Path Following MPC with Dynamic Obstacles ($\mathcal{M}_{c_1(t)}$ in bold black, unused manifolds in faded black, sensor data as red point cloud, q_{opt} in cyan, $q_{end}^{ref}(t)$ in dashed pink, q_{end} in dashed black, $q(t)$ in red, q_{ref} as black dotted line)

Chapter 4

Path Following MPC in Airborne Wind Energy Systems

A nonlinear path following model predictive control scheme with application to a kite based airborne wind energy system is presented. A novel terminal convergence field constraint is introduced to guarantee closed-loop stability and convergence of the vehicle to geometric paths of desired shapes. Convergence conditions are investigated and the effectiveness of the approach is demonstrated via numerical simulations for desired path shapes under nominal and perturbed conditions.

4.1 Introduction

A common operational requirement for kite based Airborne Wind Energy (AWE) systems is to track a desired optimal trajectory that maximizes power generation. This requires the kite to reel out at a desired rate as it flies a high energy extraction trajectory, then reel back in with a low energy consumption maneuver to produce a net positive energy generation cycle (see, e.g., [56, 57]).

While the desired trajectory for the vehicle can be pre-computed using numerical optimal control solvers ([58, 59, 60]), the trajectory tracking itself presents numerous challenges due to nonholonomic properties of the system, uncertain wind and system parameters and limited controllability of the vehicle speed. In fact, since the main driving force is provided by the wind, the vehicle can only follow time-profiles along the reference path that are coherent with the wind speed. Previous works, like [61, 62] consider trajectory-tracking Nonlinear Model Predictive Control (NMPC) schemes that can track reference positions on a time parameterized reference trajectory. The effects of unknown wind conditions however limit the applicability of such schemes as the reference trajectory can quickly become incoherent with the wind speed and kite position, leading to non zero tracking errors. [63, 64] overcome this issue of incoherence by changing the reference only on certain position feedback switching events. They do not, however, consider any exact reference trajectory or path shape to be followed. [65] tackles this issue by considering a path following scheme based on feedback linearization of the AWE system. The feedback linearization scheme however provides only a localized region of attraction in the presence of input saturation and leads to suboptimal control demands due to cancellation of all natural dynamics of the vehicle.

Motivated by these observations, a model predictive, path following control (abbreviated

MPFC) scheme is presented to plan for feasible trajectories guaranteeing convergence and tracking of the reference path.

The core idea of a path following MPC scheme is to consider a geometric reference path instead of a time-parametrized reference trajectory [66]. A virtual system is used to control the motion of a reference point along the path. Finally, the input to the virtual system and the real system input are computed by means of receding horizon optimization such that the path is followed as closely as possible.

In order to guarantee path convergence, we consider terminal constraints, which are inspired by vector field control schemes often used in aerial vehicles or mobile robots ([67]). The advantage of incorporating such constraints in the MPC scheme is that we do not need an explicit representation of the vector field.

The organization of the chapter is as follows. In Section 4.2, the problem statement is introduced with the kite and virtual path reference dynamics. Section 4.3 discusses the design of the proposed MPFC scheme. Section 4.4 provides the proof for stability and recursive feasibility of the MPFC scheme subject to a reachability assumption for the convergence field. Section 4.5 presents results for numerical tests of the controller under nominal and perturbed conditions.

Notation

$\|\|v\|\|$	2-norm of a vector ($\|\|v\|\| := \sqrt{v^T v}$)
$\|\|v\|\|_Q$	Q-norm of vector ($\|\|v\|\|_Q := \sqrt{v^T Q v}$)
$\mathrm{atan2}(\cdot,\cdot)$	Four quadrant inverse tangent
$\partial_\tau f$	Partial derivative of a function f w.r.t. τ
$\dot{f}$	Partial derivative of function f with respect to time

4.2 Problem statement

Recall that the core idea of MPFC is to coordinate the control of the real vehicle and the reference speed along a path such that the tracking error is minimized. In order to precisely define this objective for path following in AWE systems, we describe the model of the AWE system and the virtual system used to control the reference motion along the path. Finally, we define the path-following problem.

4.2.1 Kite model

The vehicle for the AWE system moves on a sphere of radius L (tether length). Its position in polar coordinates is given by the elevation and azimuth angles ϑ, φ. We denote by γ the direction of the tangential velocity of the vehicle on the sphere. By simple geometric relations, the angle γ can be written as,

$$\gamma = \mathrm{atan2}(\dot{\varphi}\cos\vartheta, \dot{\vartheta}) \tag{4.1}$$

Let

$$\mathcal{Q} = \{(\vartheta, \varphi, \gamma) : \vartheta \in [0, \pi/2), \varphi \in (-\pi/2, \pi/2), \gamma \in [-\pi, \pi]\}$$

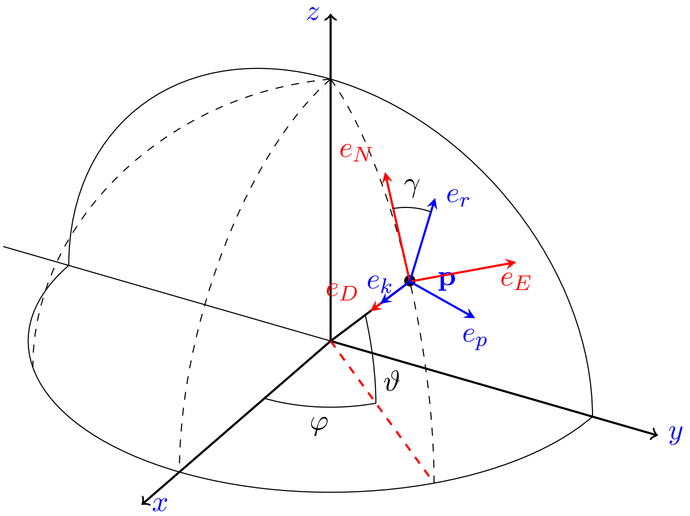


Figure 4.1: Kite coordinate frames representation.

denote the state space for the kite state. Let $\Lambda = (v_w, E) \in \mathbb{R}^2$ be the parameters, wind speed and aerodynamic glide ratio for the kite, respectively. Denoting the state of the vehicle at time t as $q(t) := (\vartheta(t), \varphi(t), \gamma(t)) \in \mathcal{Q}$, we can write the vehicle dynamics as,

$$\dot{q} = s(q, \lambda, L, z) \begin{pmatrix} \cos\gamma \\ \sin\gamma \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ u_\gamma \end{pmatrix} \tag{4.2}$$

where $s(q, \lambda, L, z) : \mathcal{Q} \times \Lambda \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is the physical speed of the kite as a function of the vehicle state $q \in \mathcal{Q}$, physical parameters $p \in \Lambda$ comprising the wind speed and aerodynamic parameters, tether length L and the reel out rate $\dot{L} = z$. u_γ is a steering input to the system that allows us to turn the vehicle ($\dot{\gamma} = u_\gamma$). We refer the reader to Figure 4.1 for a graphical representation of the angles and to [65] for details on this model. The specifics of s, Λ are given in the Appendix.

Remark 4.1. *We assume for the design of the controller, that the tether length L is fixed and the reel out rate $z = 0$. Any deviation from this assumption is treated as a perturbation.*

4.2.2 Reference path

For the reference path we consider any twice continuously differentiable *periodic* mapping $q_{ref}(\tau) : \mathbb{R} \to \mathcal{Q}$ satisfying the assumptions below.

Assumption 4.1 (Nonholonomic constraint). *The reference path is such that* $\gamma_{ref} = \operatorname{atan2}(\partial_\tau \varphi_{ref} \cos\vartheta_{ref}, \partial_\tau \vartheta_{ref})$.

Thus the reference path γ_{ref} satisfies the same geometric relation with $\vartheta_{ref}, \varphi_{ref}$ as (4.1). This assumption is satisfied by any path for which we choose $\vartheta_{ref}(\tau), \varphi_{ref}(\tau)$ and then compute $\gamma_{ref}(\tau)$ as given above.

Assumption 4.2 (Regular curve). *The path is regular in the sense that for all* τ *we have* $||\partial_\tau q_{ref}(\tau)|| \neq 0$.

Thus the reference $q_{ref}(\tau)$ does not remain stationary as τ changes. In other words, locally each point on the path corresponds to a unique τ, cf. [68].

Assumption 4.3 (Compact range). *The path parametrization* $q_{ref} : \mathbb{R} \to \mathcal{Q}$ *has a compact range in* $\mathcal{Q}$.

Thus $q_{ref}(\tau)$ attains a value for each τ within $\mathcal{Q}$, avoiding unbounded reference trajectories and limiting behavior where the limiting value does not lie in $\mathcal{Q}$.

Assumption 4.4 (Input admissibility). *The reference path curvature is limited such that it can be tracked at a given vehicle speed while respecting the input constraints on steering.*

Assumption 4.4 allows only those reference paths which can be tracked by the real vehicle at a given speed s with limited steering input u_γ^{max} when starting with zero tracking error.

We move a virtual point along the path by moving τ with controlled velocity u_τ with the dynamics,

$$\dot{\tau} = u_\tau \tag{4.3}$$

We enforce $u_\tau \geq 0$ to make the virtual vehicle move in a fixed direction along the path.

4.2.3 Kite path-following problem

We consider the augmented system of the virtual point and real vehicle with the state

$$x(t) := (q(t), \tau(t)).$$

The dynamics of $x(t)$ is then given by (4.2), (4.3). We define the path error for q and $q_{ref}(\tau)$ for an arbitrary q, τ as

$$e(q, \tau) = q - q_{ref}(\tau). \tag{4.4}$$

For $q(t), \tau(t)$, at time t, we denote the path-following error as

$$e(t) = q(t) - q_{ref}(\tau(t)). \tag{4.5}$$

For notational convenience, the tangent to the reference path is denoted as $m(\tau) = \partial_\tau q_{ref}(\tau)$. The control objective is then to asymptotically drive $e(t)$ to 0 as $t \to \infty$ subject to the constraints (4.2–4.3) and the actuator constraint set

$$\mathcal{U} = \{(u_\tau, u_\gamma)^T \in \mathbb{R}^2 \,|\, u_\tau \geq 0, \quad |u_\gamma| \leq u_\gamma^{max}\}.$$

4.3 Model predictive path following control

As standard in conventional NMPC, MPFC is based on receding horizon solutions to an Optimal Control Problem (OCP). Here, we consider MPFC based on the following problem:

$$\min_{\substack{x(\cdot)\in L^2([0,T];\mathbb{R}^n)\\ u(\cdot)\in L^2([0,T];\mathbb{R}^m)}} \int_0^T \frac{1}{2}||e(s)||_Q^2 + ||u(s)||_R^2 ds + \frac{1}{2}||e(T)||_{Q_f}^2 \tag{4.6a}$$

$$\text{subject to } (4.2), (4.3) \quad \text{with } x(0) = \hat{x}(t) \tag{4.6b}$$

$$u(s) \in \mathcal{U}, x(s) \in \mathcal{X} \quad \forall s \in [0, T] \tag{4.6c}$$

$$\frac{1}{2}||e(T)||_Q^2 + e(T)^T Q_f \dot{q}(T) \leq 0. \tag{4.6d}$$

$$e(q(T), \tau(T)) Q_f m(\tau(T)) \geq 0 \tag{4.6e}$$

where $\mathcal{X} := \mathcal{Q} \times \mathbb{R}$, and $\hat{x}(t)$ denotes the system's state at time t under the closed-loop control action of the MPFC scheme. The OCP is solved in receding horizon fashion at time t.[1] The actual input applied to the system $\hat{u}(t)$ is given by $\hat{u}(t) = u^\star(0)$ for the optimal solution $u^\star(\cdot)$ of the OCP at time t. As will be shown later, the constraints (4.6e),(4.6d) correspond to the existence of a vector field controller and is used to provide a larger region of attraction to the MPFC scheme. The matrices Q, Q_f are chosen to be symmetric positive definite.

The next result certifies the path convergence properties of the MPFC scheme based on OCP (4.6).

Proposition 4.1 (Path convergence). *Consider the MPFC scheme based on* (4.6). *Let the prediction model* (4.2) *be an exact representation of the kite dynamics, i.e. there is no plant-model mismatch. Suppose that OCP* (4.6) *is feasible for all* $t \geq 0$. *Then, the closed loop satisfies*

$$\lim_{t\to\infty} \|e(t)\| = 0.$$

Proof: Consider the positive semi-definite value function $V : \mathcal{X} \to [0, \infty)$

$$V(\hat{x}(t)) = \int_0^T \frac{1}{2}||e(s)||_Q^2 ds + \frac{1}{2}||e(T)||_{Q_f}^2 \tag{4.7}$$

where the trajectory $e : [0, T] \to \mathbb{R}^3$ is the one predicted by dynamics (4.2),(4.3) under the optimal input trajectory $u^\star(\cdot)$ to (4.6) with initial condition given as $\hat{x}(t)$. Note also that V is positive semidefinite since it only depends on e which lies in a subset of $\mathcal{X}$ and the condition $V = 0$ characterizes the set of points on the reference path. Consider the derivative of V along the closed-loop trajectories of $\hat{x}(t)$,

$$\frac{dV}{dt} = \frac{\partial V}{\partial \hat{x}} \dot{\hat{x}}$$

[1] Note that, for sake of simplified exposition, we consider the nominal case of recomputing the solution to (4.6) in an instantaneous fashion. Furthermore, we assume that, for all $\hat{x}(t)$ and $u(\cdot)$ being piecewise continuous, the OCP admits a locally optimal solution.

We have that

$$\frac{\partial V}{\partial \hat{x}}\dot{\hat{x}} = \int_0^T e^T(s)Q\dot{e}(s)ds + e^T(T)Q_f\dot{e}(T).$$

Integration by parts yields

$$\frac{dV}{dt} = \frac{\partial V}{\partial \hat{x}}\dot{\hat{x}} = \frac{1}{2}e^T(s)Qe(s)\Big|_0^T + e^T(T)Q_f\dot{e}(T). \tag{4.8}$$

The following implication then follows directly

$$\frac{1}{2}||e(T)||_Q^2 + e^T(T)Q_f\dot{e}(T) \leq 0 \tag{4.9}$$

$$\Rightarrow \quad \frac{dV}{dt} \leq -\frac{1}{2}\hat{e}(t)^T Q\hat{e}(t). \tag{4.10}$$

Here $\hat{e}(t)$ $(= e(0)$ in (4.8)) is the closed-loop path-following error corresponding to $\hat{x}(t)$. In Theorem 4.2 we show that if the terminal constraints (4.6d), (4.6e) hold, then (4.9) is satisfied. Then, using LaSalle's invariance principle in conjunction with (4.10) it follows that $\hat{x}(t)$ converges to the largest invariant set such that $\dot{V} = 0 \subset \{x \in \mathcal{X} : e = 0\}$. □

The above proof sketch relies on the quite strong assumption of recursive feasibility of OCP (4.6). In the next section we discuss the existence of a terminal control law enforcing (4.6d),(4.6e) ($\implies$ (4.9)) and recursive feasibility.

4.4 Terminal control and constraints

4.4.1 Global Feasibility of (4.6e):

For any state $q = (\vartheta, \varphi, \gamma)^T$, let

$$\mathcal{H}(q) = \{\tau^\star \,|\, \tau^\star \in \arg\min_{\tau} e(q,\tau)^T Q_f e(q,\tau)\}.$$

Also recall, $m(\tau) := \partial_\tau q_{ref}(\tau)$ and $e(q,\tau) = q - q_{ref}(\tau)$.

Lemma 4.1 (Minimum error points on path).
For all $\tau^\star \in \mathcal{H}(q)$, it holds that $e(q,\tau^\star)^T Q_f m(\tau^\star) = 0$.

Proof: Note that $\tau^\star \in \mathcal{H}(q)$ is a minimizer of $e^T Q_f e$. Hence, symmetry of Q_f and the first-order optimality condition imply $-e(q,\tau^\star)^T Q_f m(\tau^\star) = 0$. □

Lemma 4.2 (Non-emptiness of $\mathcal{H}(q)$).
For all $q \in \mathcal{Q}$, it holds that $\mathcal{H}(q) \neq \varnothing$.

Proof: As $q_{ref}(\tau)$ is twice continuously differentiable map to a compact subset of $\mathcal{Q}$ and is periodic in τ, for any q, the term $e(q,\tau)^T Q_f e(q,\tau)$ has a minimizer. Thus optimizing over $\tau \in \mathbb{R}$ implies $\mathcal{H}(q) \neq \varnothing$. □

Lemma 4.3 (Existence of neighborhoods of $\tau^\star$).
For all $\tau^\star \in \mathcal{H}(q)$, there exists a non-empty and non-singular neighborhood

$$\mathcal{N}(q,\tau^\star) := \left\{\tau \,|\, e(q,\tau)^T Q_f m(\tau) \geq 0\right\} \neq \varnothing$$

and $\mathcal{N}(q,\tau^\star) \setminus \tau^\star \neq \varnothing$.

Proof: By Lemma 4.1, we have $e(q,\tau^\star)^T Q_f m(\tau^\star) = 0$. Furthermore, $e(q,\tau)^T Q_f m(\tau)$ is a continuous function of τ that has a local minimum at $\tau^\star$. Hence, there exists a neighborhood of $\tau^\star$ wherein $e(q,\tau)^T Q_f m(\tau) \geq 0$. □

Let

$$\mathcal{N}(q,\mathcal{H}(q)) := \bigcup_{\tau^\star \in \mathcal{H}(q)} \mathcal{N}(q,\tau^\star),$$

then the following theorem states feasibility of (4.6e).

Theorem 4.1 (Global feasibility of (4.6e)).
For any terminal condition $q(T) \in \mathcal{Q}$, there exists a $\tau(T) \in \mathbb{R}$ such that $(q(T),\tau(T))$ satisfies (4.6e) and is given by $\{q(T),\tau(T) : q(T) \in \mathcal{Q}, \tau(T) \in \mathcal{N}(q(T),\mathcal{H}(q(T)))\}$.

Proof: Observe that, for any given kite state $q \in \mathcal{Q}$,

$$\mathcal{N}(q,\mathcal{H}(q)) = \{\tau | e(q,\tau)^T Q_f m(\tau) \geq 0\}$$

is the set of all τ satisfying the terminal constraint (4.6e). Lemma 4.3 shows that, for all $q \in \mathcal{Q}, \mathcal{N}(q,\mathcal{H}(q)) \neq \varnothing$. Thus, independent of initial condition $q(0)$ and for any terminal state $q(T) \in \mathcal{Q}$, $\tau_T \in \mathcal{N}(q(T),\mathcal{H}(q(T)))$ satisfies (4.6e).

Since the considered path is periodic and we do not impose any input magnitude constraint on u_τ. Hence, for any $\tau(0)$, there exists a positive input such that $\tau(T) = \tau_T$. □

4.4.2 Feasibility of (4.6d):

For sake of readability, we drop the time argument T from vectors like $m(T), e(T), q(T)$. Recall that $\mathcal{X} := \mathcal{Q} \times \mathbb{R}$. We will also use the following short hand notations: $F(q,\tau) := \begin{pmatrix} m & e \end{pmatrix} \in \mathbb{R}^{3\times 2}$. Dropping the arguments (q,τ), we write, $g = F^T Q_f e \in \mathbb{R}^{2\times 1}$, $P = F^T Q_f F$, $S = \mathrm{diag}(1,1,0)$ and $H = F^T S F$. Also note that we can rewrite (4.6d) as

$$e^T Q_f \dot{q} \leq -\frac{1}{2}||e||_Q^2$$

Inspired by the concept of vector field controllers, let us try to find a vector field $\mathbf{v}(q,\tau) : \mathcal{X} \to \mathbb{R}^3$ such that $\dot{q}(T) = \mathbf{v}(q(T),\tau(T))$ satisfies (4.6d). To this end, for $\mathbf{w}(q,\tau) : \mathcal{X} \to \mathbb{R}^2$, we parametrize the desired vector field $\mathbf{v}(q,\tau)$ as

$$\mathbf{v}(q,\tau) = F(q,\tau)\mathbf{w}(q,\tau).$$

and thus (4.6d) can be written as

$$g^T w \leq -\frac{1}{2}||e||_Q^2$$

in the notation defined above.

Furthermore, from (4.2) we have that $||\dot{q}||_S = s(q, \lambda, L, z)$ imposes a magnitude constraint on $S\dot{q}$. Thus, in order to find a $\mathbf{v}(q,\tau)$ that satisfies this magnitude constraint and satisfies (4.6d), we consider a optimization problem to be solved at each (q,τ) to yield a $\mathbf{w}^\star(q,\tau)$

$$\underset{\mathbf{w}\in\mathbb{R}^2}{\text{minimize}} \quad \frac{1}{2}||F\mathbf{w} - m||_{Q_f}^2 + \frac{1}{2}||\mathbf{w}||^2 \tag{4.11a}$$

$$\text{subject to} \quad g^T\mathbf{w} \leq -\frac{1}{2}||e||_Q^2 \tag{4.11b}$$

$$\mathbf{w}^T H\mathbf{w} - s^2 = 0 \tag{4.11c}$$

where $s = s(q, \lambda, L, z)$. Note that, for the sake of readability, we drop the arguments (q,τ) above. The penalization $||F\mathbf{w} - m||_{Q_f}^2$, in the objective function (4.11a), regularizes $\mathbf{w}$ such that $\mathbf{v}$ points along the tangent, m, of the path whenever possible. Equation (4.11b) imposes that $\mathbf{v}$ satisfies (4.6d) and (4.11c) imposes that $\mathbf{v}$ satisfies the magnitude constraint on $S\dot{q}$. Setting $\mathbf{w}(q,\tau) = \mathbf{w}^\star(q,\tau)$ and solving (4.11) yields the desired vector field $\mathbf{v}(q,\tau)$.

Proposition 4.3, presented in Appendix, provides the optimal solution $\mathbf{w}^\star$ to (4.11) and its existence conditions. Specifically, when already on the path ($e = 0$), it turns out that $\mathbf{w}^\star$ is such that $F\mathbf{w}^\star$ points along the tangent direction given by direction of m. Thus, when on the path, the desired vector field pushes the vehicle along the path, rendering the path an invariant set under the vector field. Choosing $\dot{q}(T) = \mathbf{v}(q(T), \tau(T))$, then imposes (4.6d).

Let $\mathcal{S} \subset \mathcal{Q}$ be a compact subset of $\mathcal{Q}$ for which the conditions of Proposition 4.3 are satisfied. Let $\mathcal{V}_\alpha := \{e : \frac{1}{2}||e||^2 \leq \alpha\}$ be the largest level set contained in $\mathcal{S}$ (subject to maximization w.r.t. α). Furthermore, let $\angle\mathbf{v} = \text{atan2}(\mathbf{e}_2\mathbf{v}, \mathbf{e}_1\mathbf{v})$, $\mathbf{e}_1 = \begin{pmatrix}1 & 0 & 0\end{pmatrix}$, $\mathbf{e}_2 = \begin{pmatrix}0 & 1 & 0\end{pmatrix}$, $\mathbf{e}_3 = \begin{pmatrix}0 & 0 & 1\end{pmatrix}$. Then the recursive feasibility for (4.6) can be shown as follows.

Assumption 4.5. *Assume that, starting at the any $q(0) \in \mathcal{Q}$, there exists a reachable point in $\mathcal{Q}$ such that $\mathbf{v}(q(T), \tau(T)) = F(q(T), \tau(T))\mathbf{w}^*(q(T), \tau(T))$.*

Proposition 4.2 (Recursive feasibility of (4.6d)).
Let Assumption 4.5 hold. Then, for any $q(T) \in \mathcal{V}_\alpha$, $\tau(T) \in \mathcal{N}(q(T), \mathcal{H}(q(T)))$ and input $u_\gamma = \mathbf{e}_3\mathbf{v}(q(T), \tau(T))$, $\gamma(T) = \angle\mathbf{v}(q(T), \tau(T))$, the terminal condition (4.6d) holds. Furthermore, the set $\mathcal{V}_\alpha$ is positively invariant and (4.6d) is recursively feasible.

Proof: Observe that $\mathbf{e}_1\dot{q} = s\cos\gamma$, $\mathbf{e}_2\dot{q} = s\sin\gamma$ and $\mathbf{e}_3\dot{q} = u_\gamma$. Thus, for $q(T) \in \mathcal{V}_\alpha$, we have $\gamma(T) = \angle\mathbf{v}(q(T), \tau(T))$, $u_\gamma = \mathbf{e}_3\mathbf{v}(q(T), \tau(T))$, and (4.11c) implies that $\dot{q}(T) = \mathbf{v}$ holds. From (4.11b) we have $\mathbf{v}$ such that $\dot{q}(T) = \mathbf{v}$ satisfies (4.6d). Furthermore, considering the positive semidefinite function $V_T(q,\tau) = \frac{1}{2}||e||_{Q_f}^2$, we see that $\dot{V}_T = e^T Q_f \dot{e} = e^T Q_f \dot{q} - e^T Q_f m(\tau) u_\tau$. Since $\tau(T) \in \mathcal{N}(q(T), \mathcal{H}(q(T)))$ is such that $e^T Q_f m(\tau) \geq 0$ and $u_\tau \geq 0$, we have $e^T Q_f m(\tau) u_\tau \geq 0$ implying $\dot{V}_T \leq e^T Q_f \dot{q}$. Further with $\dot{q} = \mathbf{v}$, from (4.11b), we have $e^T Q_f \dot{q} \leq -\frac{1}{2}||e||_Q^2$ implying $\dot{V}_T \leq -\frac{1}{2}||e||_Q^2$. This implies that $\mathcal{V}_\alpha$ is positively invariant and thus (4.6d) is recursively feasible. □

4.4.3 Convergence Constraints

Theorem 4.2. *If* (4.6d) *and* (4.6e) *hold for a terminal state* $x(T) = (q(T), \tau(T))$*, then* (4.9) *also holds for* $x(T)$*.*

Proof: For sake readability, let us drop the time argument T with the understanding that all quantities in the expressions below are at the terminal time T. Consider

$$\frac{1}{2}||e||_Q^2 + e^T Q_f \dot{e} = \frac{1}{2}||e||_Q^2 + e^T Q_f \dot{q} - e^T Q_f m(\tau) u_\tau.$$

Then from (4.6d), it follows that $e^T Q_f \dot{q} \leq -\frac{1}{2}||e||_Q^2$. Furthermore, from (4.6e) we have that, for all $u_\tau \geq 0$, $-e^T Q_f m(\tau) u_\tau \leq 0$. These statements imply that

$$\frac{1}{2}||e||_Q^2 + e^T Q_f \dot{e} \leq 0.$$

This finishes the proof. □

4.5 Numerical results

For the numerical implementation of our continuous time MPFC scheme we use a sampled data implementation with sampling time δ. The OCP (4.6) in the sampled data setting is solved using a direct multiple shooting approach with N time step horizon ($T = N \cdot \delta$). The nonlinear program (NLP) is setup with automatic differentiation using CasADi ([69]) with a RK4 integrator approximation and solved using an interior point solver (IPOPT, [70]) on a 2.8 GHz Intel Core i7 processor. The values for Q, Q_f, R, N, T are given in (4.12) in the Appendix.

Subsequently, we discuss results for the following scenarios:

1. Nominal simulations: Simulations under zero plant-model mismatch
2. Perturbed simulations:
   (a) Sampled velocity: Speed of the vehicle is sampled at the beginning of the MPC horizon and then assumed constant at that value over the horizon.
   (b) Pumping cycle: The vehicle is reeled in and out with an external controller. The tether length is sampled at the beginning of the horizon and assumed constant over the horizon. The vehicle speed is sampled and assumed constant over the horizon as done in scenario 2a.
3. MPFC without terminal constraints

Note that in the perturbed scenarios 2a,2b, the AWE system is still simulated using the full model in equation (4.2), while the perturbed models as described in 2a,2b are used for predictions in the MPFC controller. Simulations without the terminal constraints is presented to highlight the role of terminal convergence constraints in enforcing faster convergence.

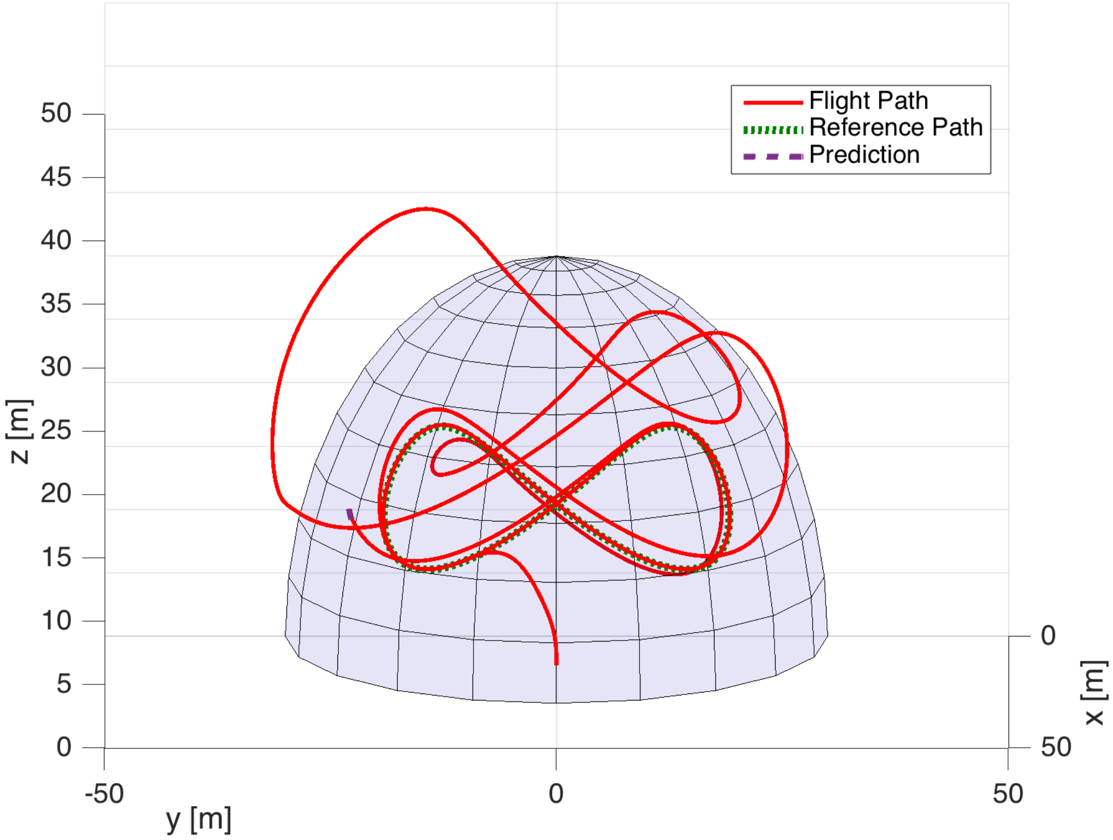


Figure 4.2: Closed loop flight trajectory with convergence constraints for complete pumping cycle.

4.5.1 Nominal simulations

The first 30 seconds of Figure 4.2 and 4.3 provide a typical MPFC closed loop trajectory for the AWE vehicle under nominal simulation when following a lemniscate shaped reference path (see Appendix, (4.14)). Figure 4.3 shows the closed loop evolution of the system state $\hat{x}(t)$ and closed loop inputs $\hat{u}_\gamma(t), \hat{u}_\tau(t)$ when tracking the lemniscate. With $\delta = 0.1s$ and $N = 10$, the average solve time to plan a 1 s long horizon is 0.2 s, suggesting the possibility to apply the nominal MPFC scheme in real time as a higher level planner in a cascaded structure control scheme.

4.5.2 Perturbed simulations

We test our control scheme applying perturbations in the velocity model $s(q, \lambda, L, z)$. Since we do not have an accurate model $s(q, \lambda, L, z)$ due to unknown parametric and structural uncertainties, we choose a simplified model where, $s(q, \lambda, L, z) = s_o$. The constant s_o is updated to the speed estimate of the vehicle at the beginning of the horizon and then held constant for the MPC prediction over the horizon. Figure 4.4 shows the closed loop path following error obtained under the perturbation 2a showing a very close overlap with the nominal case. Thus the MPFC scheme seems to have sufficient inherent robustness with the chosen parameters to plan under imperfect

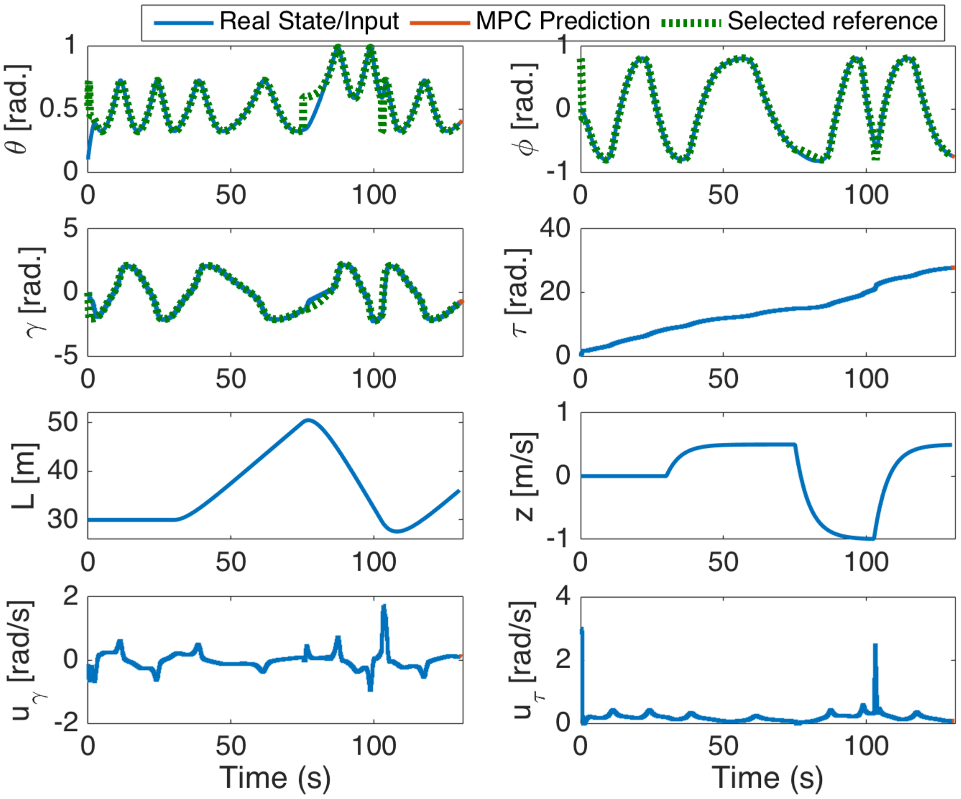


Figure 4.3: Closed loop state and input evolution.

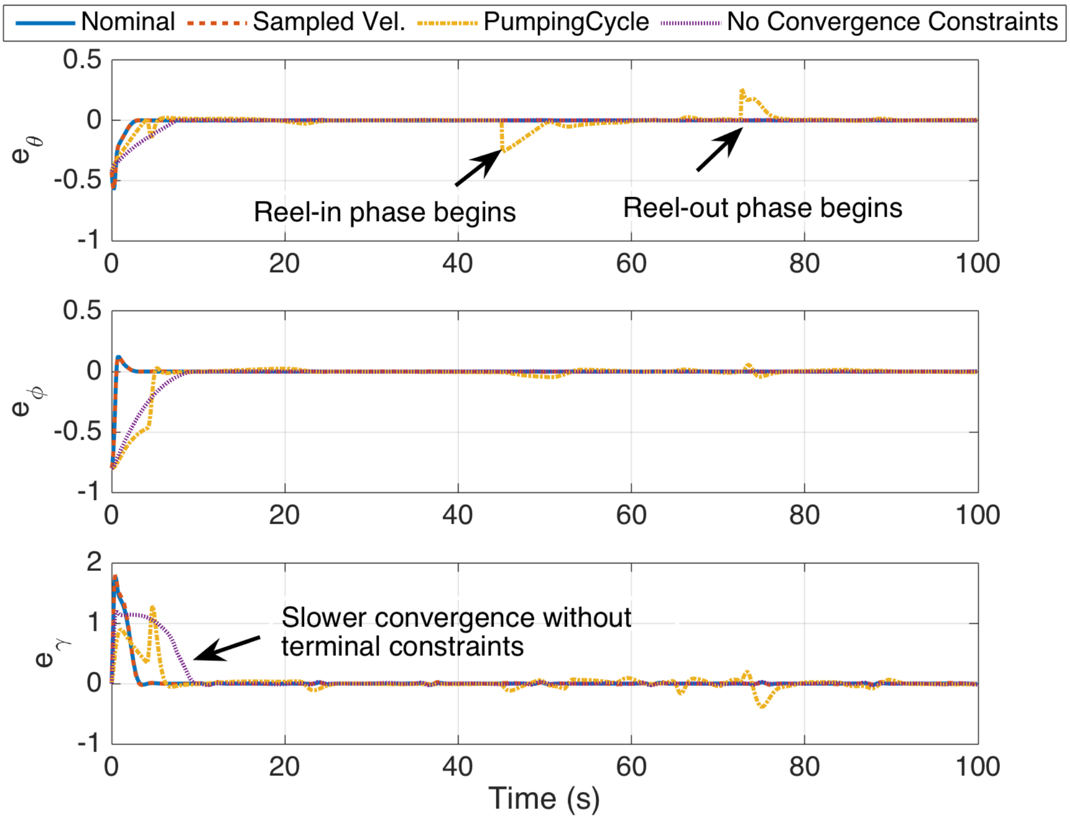


Figure 4.4: Path following errors.

prediction. Further due to the simplification of the dynamics under this sampled speed model, the solve time is significantly reduced with an average solve time of about 0.05 seconds and worst case solve time of about 0.1 second.

The trajectory after 30 seconds in Figure 4.2 and 4.3 shows the closed loop trajectory for the complete pumping cycle operation of an AWE system corresponding to perturbation 2b. Figure 4.4 shows the path following errors for perturbation 2b, where the closed loop trajectory follows a reference for the full pumping cycle which is made up by switching between two different lemniscate paths for the traction and retraction phases. The switching of reference path creates an instantaneous increase in the path following error which is reduced quickly by the MPFC scheme to start tracking the new reference path.

4.5.3 Control without terminal constraints

Figure 4.4 also shows the convergence of path following errors when terminal constraints are not imposed. This shows that the MPFC scheme can also work with only the terminal penalty being imposed. However by comparison, the terminal constraints significantly improve the convergence rates.

4.6 Conclusion

The chapter presented a nonlinear model predictive path following (MPFC) scheme for airborne wind energy systems that may be used as a high-level planner in combination with a low-level steering controller. A novel set of terminal convergence field constraints are introduced to guarantee asymptotic convergence to zero path tracking error. Recursive feasibility and convergence results have been investigated for the proposed scheme under a reachability assumption for the convergence field. Numerical studies under nominal and perturbed conditions indicate good control performance. The proposed MPFC scheme is observed to be computationally viable for real-time application in cascaded kite control schemes.

Appendix

Simulation parameters

MPFC parameters:

$$\begin{aligned} &Q = \mathrm{diag}(1000, 1000, 3), \quad Q_f = \mathrm{diag}(500, 500, 150) \\ &R = \mathrm{diag}(0.1, 0.01), N = 10, \delta = 0.1 sec., u_\gamma^{max} = 20 \end{aligned} \tag{4.12}$$

Model data:

$$\mathcal{Q} = \{(\vartheta, \varphi, \gamma) : \vartheta \in [0, \pi/2), \varphi \in (-\pi/2, \pi/2), \gamma \in [-\pi, \pi]\}$$

$$s(q, \lambda, L, z) = \begin{pmatrix} 1 & 0 \\ 0 & (\cos\vartheta)^{-1} \end{pmatrix} v_w L^{-1} \begin{pmatrix} \cos\gamma \\ \sin\gamma \\ -E \end{pmatrix}^T \begin{pmatrix} -\sin\vartheta\cos\varphi \\ -\sin\varphi \\ -\cos\vartheta\cos\varphi \end{pmatrix} \tag{4.13}$$

$\lambda = (v_w, E)$ are the parameters, wind speed and aerodynamic glide ratio respectively.

Reference path parameterization: For the lemniscate reference path used in numerical simulations, we use,

$$\vartheta_{ref}(\tau) = h + a\sin(2\tau), \quad \varphi_{ref}(\tau) \quad = 4a\cos(\tau) \tag{4.14}$$

In perturbed Sscenario 2b we use $h = \pi/6$ for reel-out and $h = \pi/4$ for reel-in. $a = 0.2$ for all cases. The tether is reeled-out at 0.5 m/s and reeled-in at 1 m/s.

Terminal convergence constraints

Below $adj(\cdot)$ represents the adjoint or adjugate of a matrix and trace$(\cdot)$ gives the trace of a matrix. Let, $F := \begin{pmatrix} m & e \end{pmatrix} \in \mathbb{R}^{3\times 2}$, $g := F^T Q_f e \in \mathbb{R}^{2\times 1}$, $P = F^T Q_f F$, $H = F^T S F$ and

$$\begin{aligned}
\ell &= -2||g||^2_{adj(H)}, \quad c = -||g||^2_{adj(P)} - ||g||^2 \\
b_0 &= -g^T h - g^T adj(P)h - \frac{1}{2}(|P| + 1 + \text{trace}(P))||e||^2_Q \\
b_1 &= -2g^T adj(H)h + (\text{trace}(H) + \text{trace}(H adj(P)))||e||^2_Q \\
b_2 &= 4|H|||e||^2_Q,\ r_0 = b_0 \pm cs, r_1 = b_1 \pm \ell s, \quad r_2 = b_2
\end{aligned}$$

The variable s is the speed as defined in (4.11).

$$\begin{aligned}
\alpha_0 &= -adj(gg^T)h + \frac{1}{2}(I + adj(P))g||e||^2_Q \\
\alpha_1 &= -2adj(H)g||e||^2_Q,\ \beta_0 = h + adj(P)h \\
\beta_1 &= 2adj(H)h,\ h = \begin{pmatrix} ||m||^2_{Q_f} \\ m^T Q_f e \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
k_0 &= |P| + \text{trace}(P) + 1,\ k_2 = 4|H| \\
k_1 &= 2\,\text{trace}(H) + 2\,\text{trace}(H adj(P))
\end{aligned}$$

$$n_0 = k_0 s \pm \beta_0, \quad n_1 = k_1 s \pm \beta_1, \quad n_2 = k_2 s$$

$$\lambda(\nu) \quad = \quad \frac{1}{\ell\nu + c}(b_2\nu^2 + b_1\nu + b_0) \tag{4.15a}$$

$$\text{for } \nu \quad \in \quad \mathcal{P}_1 := \{x \in \mathbb{R} : r_2 x^2 + r_1 x + r_0 = 0\} \tag{4.15b}$$

also let,

$$\mathcal{P}_0 = \{x \in \mathbb{R} : n_2 x^2 + n_1 x + n_0 = 0\} \tag{4.16}$$

Proposition 4.3. *The minimizing solution for* (4.11) *is*

$$\mathbf{w}^{\star} = \frac{1}{\sigma(\mu)}(a_1\mu + a_0) \tag{4.17a}$$

with $\mu \in \mathcal{P}_\lambda$ *such that, for* $e \neq 0, \lambda(\nu) > 0$,

$$\mathcal{P}_\lambda = \mathcal{P}_1, \ \& \ (a_1, a_0) = (\alpha_1, \alpha_0), \quad \sigma(\mu) = \ell\mu + c. \tag{4.17b}$$

Otherwise

$$\mathcal{P}_\lambda = \mathcal{P}_0, \quad (a_1, a_0) = (\beta_1, \beta_0), \quad \sigma(\mu) = k_2\mu^2 + k_1\mu + k_0. \tag{4.17c}$$

The solution exists if the quadratics defining $\mathcal{P}_0$ *and* $\mathcal{P}_1$ *have real roots in a compact set* $\mathcal{S} \subset \mathcal{Q}$*.*

Proof: The proof follows by writing the Lagrangian for (4.11) as

$$\mathcal{L}(w, \lambda, \mu) \quad = \quad \frac{1}{2}||F\mathbf{w} - m||_{Q_f}^2 \ + \ \frac{1}{2}||\mathbf{w}||^2 \ + \ \lambda\left(g^T w + \frac{1}{2}||e||_Q^2\right) \ + \ \mu\left(\mathbf{w}^T H \mathbf{w} - s^2\right). \tag{4.18}$$

Then, writing the KKT conditions and solving for the two cases, $\lambda > 0$ and $\lambda = 0$. For $\lambda > 0$, we can write $\mathbf{w}^{\star}, \lambda$ as a function of μ as given in (4.15a) for $\mu = \nu$. μ itself can be obtained as roots to a quadratic equation (4.15b) with $\mu = \nu$. When $e = 0$, the inequality becomes weakly active and thus the solution can be obtained from the $\lambda = 0$ case. For $\lambda = 0$, μ can be obtained as roots to the quadratic as defined in (4.16). Once μ is taken for the appropriate case, $\mathbf{w}^{\star}$ can be obtained as specified by (4.17). □

Part III

Airborne Wind Energy Systems under Uncertainty

Chapter 5

Optimization of an Airborne Wind Energy System using Constrained Gaussian Processes and Transient Measurements

Airborne wind energy systems are built to exploit the stronger and more consistent wind available at high altitudes that conventional wind turbines cannot reach. This however requires a reliable controller design that can keep the airborne system flying for long durations in varying environmental conditions, while respecting all operational constraints. Such reliability is often delivered by a cascade of low level controllers whose combined behavior is not analytically tractable for performance optimization. An on-line data based method is presented to optimize the towing force of such a system in presence of constraints, varying wind conditions and a constrained low level tracking controller. The approach actively learns Gaussian process models for the objective, constraint and closed loop dynamics of the system and uses transient measurements to optimize over the objective. A chance - constrained optimization problem is posed taking into consideration uncertainty in the learned functions and is used to find potential feasible directions for maximizing of the towing force. Simulation studies are presented showing that we can find optimal set points for the controller without the use of significant assumptions on model dynamics while respecting the unknown constraint function. The results also show an improved performance over a Gaussian process optimization scheme that restricts itself to steady state measurements.

5.1 Introduction

A data based optimization algorithm for an AWE design that is actively used by a commercial company (Skysails) in large marine vessels to increase their fuel savings [71] is presented here. For the system, a tethered flexible airfoil is launched from a mounting station at the front of the ship towards the sky where it performs figure eight loops (lemniscate) using a custom low level controller. In favorable wind conditions the aerodynamic force generated upon the foil is transferred through the tether to pull the ship forward, reducing the load on its engine. However, such a controller can neither automatically guarantee a power optimal steady state trajectory as given by a problem (5.1), nor that the aerofoil will not cross an altitude safety threshold, given by the inequality constraint in (5.1).

$$\begin{aligned} \max_{x_s, u_s} \quad & F(x_s, u_s) \\ \text{s.t.} \quad & x_s = f(x_s, u_s), \quad G(x_s, u_s) \geq 0 \end{aligned} \tag{5.1}$$

[72] addresses (5.1) in the framework of steady state constrained optimization and has demonstrated that a data based approach can be a potential solution. The work utilizes an empirical observation for the controller that shows the closed loop dynamics to be exponentially stable for every controller set point u_s (within an unknown region of attraction), stabilizing the system to a steady state given by $x_s = h(u_s)$ for some unknown function h. Thus under stable wind conditions and using training measurements only after reaching steady state $h(u_s)$, the composed maps $F(h(u_s), u_s)$ and $G(h(x_s), u_s)$ are learned from the steady state measurements as Gaussian process surrogate models. The surrogate models are optimized over with a sampling based Gaussian process optimization technique to trade off between exploring the parameter space to improve information in the model and exploiting the available model for improving the objective value. For the system here, the steady state requires around 10 lemniscate loops with the same inputs applied and fairly constant wind, to be attained and before a measurement can be taken. Such constant wind conditions are difficult to attain in practice, requiring a more frequent wind speed dependent update of the set points u_s and also taking measurements before the system has reached steady state to make predictions about the system at steady state. A more flexible solution where transient measurements after each loop can be used both for learning and optimization is thus presented here. The problem being addressed is of the form,

$$\begin{aligned} \max_{x_s, u_s} \quad & F(x_s, u_s, w_n) \\ \text{s.t.} \quad & x_{n+1} = f(x_n, u_s, w_n), \quad G(x_n, u_s, w_n) \geq 0 \\ & x_s = f(x_s, u_s, w_n), \quad G(x_s, u_s, w_n) \geq 0 \end{aligned} \tag{5.2}$$

where $F : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is an unknown function mapping to the towing force (objective value) of the AWE system. The steady state x_s is attained as a function of the set point u_s and wind w_n, i.e. $x_s = f(x_s, u_s, w_n)$ (for an unknown dynamics function f). $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^n$ describes the discrete time system dynamics and $G : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is an unknown constraint function. The transient state of the system at time n is given by $x_n \in \mathbb{R}^n$ and can be indirectly manipulated through the inputs u_s which are bounded by operation constraints for the controller into a known compact set $\mathcal{X} \subseteq \mathbb{R}^m$. The variables $w \in \mathbb{R}^p$ represents an exogenous signal (here the wind) on which we have no control. The functions considered are assumed to be nonlinear, non-convex and unknown. Gaussian Processes (GP) [73] have been widely used in practice to model and learn such functions and have been extensively used in unconstrained, static optimization [74, 75].

With known constraints, the static problem is addressed in [76]. For unknown constraints the problem is addressed in [77, 78] using joint modeling for the objective and constraint as done here. Under unknown constraints for optimizing processes through physical experiments, however,

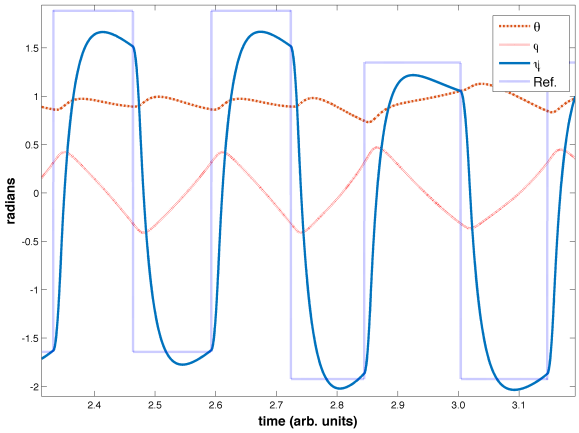


Figure 5.1: Low level tracking control for changing set points. The yaw (ψ - kite orientation) controller tracks the level reference signal which changes value when a crossing occurs. The positional states (θ and ϕ) are affected by ψ.

constraint violations are not as freely permissible as for optimization through simulations, as a result safe learning methods have gained popularity which force the sampling schemes to adhere to a trust region like approach to avoid sampling in highly uncertain regions. Such approaches have been explored in [79]. The present work, utilizes a similar approach while considering additional learned dynamical constraints and transient measurements to make predictions about future steady states and allows more frequent updates of set-points, adapting to changing wind conditions. Simulation results demonstrate that the AWE system can be optimized this way and becomes adaptive to the wind variation while respecting the altitude safety constraint.

The chapter is structured as follows: Section 5.2 briefly describes the AWE system. Section 5.3 provides an overview of the Gaussian process optimization scheme used, and Section 5.4 introduces the algorithm proposed for optimizing in the presence of dynamical constraints and transient measurements, Section 5.5 presents simulation results and Section 5.6 summarizes the findings for the approaches presented.

5.2 System description

The following describes the simulation model and low level controller used to generate closed loop simulations and measurements for the state, tether forces (objective) and constraint (altitude) used in Section 5.5 for the numerical results. Section 5.2.2 presents the low level controller and describes the set point u_s used as a decision variable for the optimization scheme in Section 5.4. Section 5.2.3 describes the discrete time closed loop dynamics used for the transient Gaussian process optimization.

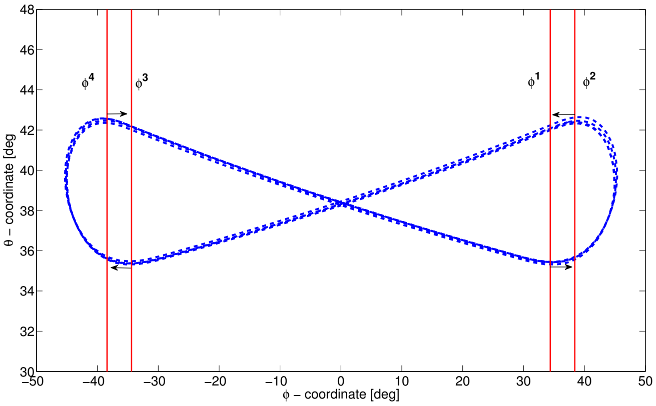


Figure 5.2: Example of kite trajectory (blue) and switching surfaces (red) for the Skysails controller in the spherical coordinate system. Arrows denote the direction of crossing for which a switching surface is active.

5.2.1 Open loop dynamics

The open loop, kinematic equations of motion describing the system [80] can be written as

$$\begin{aligned}\dot{\vartheta} &= \frac{wE}{L}\cos\vartheta\cos\psi - \frac{w}{L}\sin\vartheta\\ \dot{\varphi} &= -\frac{wE}{L}\frac{\cos\vartheta}{\sin\vartheta}\sin\psi \\ \dot{\psi} &= wEg\cos{(\vartheta)}\delta + \dot{\varphi}\cos\vartheta.\end{aligned} \tag{5.3}$$

The spherical coordinates (ϑ, φ) and orientation (ψ) of the airfoil represent the states of the system. The exogenous signal is the wind speed (w). The system coordinates are defined such that one of the horizontal axes is assumed parallel to the wind direction. The uncertain system parameters are the glide ratio (E), the deflection coefficient (g) and the tether length (L). Finally, δ is the deflection applied to the kite affecting its orientation and is used as a control input for a controller, described in Section 5.2.2.

The system exhibits nonlinear behavior even without considering some more complex effects of aerodynamics that have been significantly simplified. It is thus difficult to find closed form expressions for the functions of interest. Numerical optimization results using model based approaches, while useful, are challenged in realistic situations where the model parameters and wind conditions are unknown.

5.2.2 Event triggered control

A simple but robust low level controller developed by [64] is used for flying the “figure eight” loops. The controller uses a model free feedback scheme where a set-point $\psi_s^{(i)}$ for yaw angle ψ is tracked by applying deflections δ to the kite (see Figure 5.1). The set-point $\psi_s^{(i)}$ is taken from a sequence of set-points $\{\psi_s^{(1)}, \dots, \psi_s^{(2r)}\}$ during every loop, advancing through the sequence at predefined switching positions $\{\varphi_s^{(1)}, \ldots, \varphi_s^{(2r)}\}$. This sequence of set points and switching positions is treated as the high level decision variable

$$u_s = \{\psi_s^{(1)}, \ldots, \psi_s^{(2r)}\} \times \{\varphi_s^{(1)}, \ldots, \varphi_s^{(2r)}\}$$

in the Gaussian process optimization scheme defined in Section 5.4.

The controller sets the yaw set point to a value given by $\psi_s^{(i)}$ when the system state φ, crosses $\varphi^{(i)}$, generating an event when,

$$\varphi - \varphi_s^{(i)} = 0 \text{ and } (-1)^i \dot{\varphi} > 0; \tag{5.4}$$

Each event is defined by a constant φ_i giving its position in the azimuth angle (φ) space, considers crossings only in one direction (when $(-1)^i\dot{\varphi} > 0$) and ignores crossings in the other direction. A depiction of the resulting closed loop trajectory can be seen in Figure 5.2. A detailed analysis for the controller can be found in [81, 82].

5.2.3 Closed loop, event based dynamics

Let a fixed controller event point $\varphi_s^{(0)} = 0$ from Section 5.2.2 be given, in addition to the ones defined in u_s. Let the state of the kinematic system $x = (\vartheta, \varphi, \psi)$ from (5.3) at the k^{th} crossing event on $\varphi_s^{(0)}$ be denoted as $x_k = (\vartheta_k, \varphi_k, \psi_k)$ and let the sequence of set points u_s fixed after the k^{th} crossing be denoted u_s^k. Let the time of the k^{th} crossing of $\varphi_s^{(0)}$ be denoted t_k. Then a closed loop, event based dynamical system is written as

$$x_{k+1} = f(x_k, u_s^k, w_k) + \sigma_f^2 \eta \tag{5.5}$$

where f is an unknown function representing the closed loop integration of $\dot{x}$ over the time interval $[t_k, t_{k+1}]$, given an initial condition x_k, sequence of controller set-points u_s^k and wind condition w_k at the t_k (assumed constant over the interval $[t_k, t_{k+1}]$). $\sigma_f^2\eta$ is a zero mean Gaussian vector valued noise with variance σ_f^2 (to model for differences between the simulation model and the real system).

A noisy, minimum altitude measurement over the interval $[t_k, t_{k+1}]$ is given as,

$$z(x_k, u_k, w_k) = \min \left\{ L\cos(\vartheta(t)) : t \in [t_k, t_{k+1}], \vartheta(t) = \int_{t_k}^{t} \dot{\vartheta}(l)dl + \vartheta_k \right\} + \sigma_G^2 \eta_G$$

$\sigma_G^2\eta_G$ being the real valued, zero mean Gaussian random variable with variance σ_G^2.

The unknown constraint function from (5.2) is then taken as

$$G(x_k, u_s^k, w_k) = z(x_k, u_s^k, w_k) - z_{min} \tag{5.6}$$

where z_{min} is a constant for the minimum altitude at which the kite is permitted to fly.

The average towing force measurement can be given by

$$F(x_k, u_s^k, w_k) = (t_{k+1} - t_k)^{-1} \left(\int_{t_k}^{t_{k+1}} \kappa(v(\dot{x}(l)))^2 dl \right) + \sigma_F^2 \eta_F \tag{5.7}$$

where $v(\dot{x}(l))$ gives the kinematic speed of the kite with respect to the wind at some time l and κ is an aerodynamic constant. The average towing force computes the integral of the tether force given by $\kappa(v(\dot{x}(l)))^2$ over the time interval $[t_k, t_{k+1}]$.

This event based dynamics transferring the state from x_k to x_{k+1} according to (5.5) is referred to as the discrete "time" dynamics for the closed loop system and is used in for the GP optimization scheme. Thus we optimize over the system at the end of crossing event of $\psi_s^{(0)}$.

The system is referred as being transient when $x_{k+1} = f(x_k, u_s^k, w_k) \neq x_k$. Likewise, the system is said to be in steady state when the kite returns to the same state at the next crossing event, i.e., the steady state is denoted $x_s = f(x_s, u_s^k, w_k)$.

5.3 Gaussian processes for constrained optimization

A brief description of Bayesian regression for Gaussian processes (GP) and Gaussian process optimization using an expected improvement metric is given below. Further details for GP regression can be found in [83] and expected improvement metric can be found in [84].

5.3.1 Gaussian process regression

For a finite $l \in \mathbb{N}$, let $\mathcal{D}_l = \{x_i, y_i\}_{i=1:l}$ be a set of input-output data for some unknown function $y = h(x)$. Let h_0 be a zero mean Gaussian process with a given covariance function $k(x, x')$ (giving the covariance between the point evaluations $h_0(x)$ and $h_0(x')$). A Bayesian posterior Gaussian process h can be computed given the prior h_0 and data set $\mathcal{D}_l$ with its mean $\mu_h(x)$ and covariance function $\sigma_h(x, x')$ given by

$$\mu_h(x|\mathcal{D}_l) = k(x, x_{1:l})(K_{1:l} + \sigma_n^2 I_l)^{-1}(y_{1:l}) \tag{5.8}$$

and

$$\sigma(x|\mathcal{D}_l) = k(x, x) - k(x, x_{1:l})(K_{1:l} + \sigma_n^2 I_l)^{-1} k(x, x_{1:l})^T \tag{5.9}$$

where $K_{1:l}$ denotes a $l \times l$ matrix with the $(i, j)^{th}$ component given by $k(x_i, x_j)$, I_l denotes a $l \times l$ identity matrix and σ_n^2 is the covariance of a noise process, modeling the input-output relation $y = h(x) + \sigma_n^2 \eta$ (for a standard normal random variable η). $k(x, x_{1:l})$ denotes a $1 \times l$ matrix for which the $(1, j)^{th}$ component is $k(x, x_j)$. A similar extension to $\mathbb{R}^n-$valued Gaussian processes is given by considering the kernel $k(x, x')$ to be a matrix valued kernel [85].

A squared exponential $\mathbb{R}-$valued kernel of the following form is used below

$$k_{SE}(x, x') = \sigma_y^2 \exp(-\frac{(x - x')^T \Lambda^{-1} (x - x')}{2})$$

where Λ is a symmetric positive definite matrix determined by known constants, called hyperparameters for the Gaussian process. The parameters $\theta = \{\sigma_y, \Lambda, \sigma_n\}$ can be optimized over, using the measurements $\mathcal{D}_l$ through a likelihood maximization scheme (although not a necessary step for simple Bayesian inference and the GP optimization scheme defined next, as long as reasonable estimate for the hyper-parameters is fixed).

5.3.2 Gaussian process optimization

The Gaussian process optimization schemes [84, 77, 78, 79, 72] consider a maximization problem on a given domain set $\mathcal{X}$ and an unknown objective function $F : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$. Including the $\{-\infty\}$ in the range of F allows the consideration of constrained problems. While objective function is unknown, given a finite sample set of observations $\mathcal{D}_l = \{(x_i, F(x_i)) : i = 1, \ldots, l\}$ and a Gaussian process prior F' as considered in Section 5.3.1, a Bayesian posterior, Gaussian process $F'|\mathcal{D}_l$ can be considered as a surrogate model representing F. Since the mean of the Bayesian posterior acts like a least squares regressor fitting to the observation in $\mathcal{D}_l$, the surrogate model represents F with higher accuracy at the sample points x_i than in other places. The posterior also has lower variance at $\{x_i\}_{i=1:l}$ than at points away from the sampled data.

A Gaussian process optimization scheme then constructs a metric to select a candidate sampling point x^* to meet a dual objective, firstly to improve the objective value at the new sample point and thus optimizing the function F, secondly to explore new regions for improving the model in order to not miss any maximizers due to inaccuracies in $F'|\mathcal{D}_l$. Thus a metric J is designed to balance between information gained by sampling at x^* and the improvement in objective value achieved by sampling at x^*.

$$x^* = \underset{x \in \mathcal{X}}{\arg\max}\, J((F'|\mathcal{D}_l)(x)) \tag{5.10}$$

Since the goal is to maximize F with the fewest such samples, the bias in the designed metric is not on exploration but on maximization and exploration is done only to degree that is necessary to avoid missing maximizer due to poorly sampled information. Several such metric have been considered in the works mentioned above. Two common variants of such metrics are given by the GP-upper confidence method (GP-UCB) [86] and the Expected Improvement metric [84] (which is used in Section 5.4).

In applications requiring experimental sampling, the candidate point x^* has to ensure that it does not violate the constraint upto a certain level of confidence. This is often called a safe GP optimization algorithm and relies on choosing points within a certain trusted region of the model $F'|\mathcal{D}_l$. This is the approach followed in [79, 72] and in Section 5.4.

Without considering trust region constraints required for safe learning, convergence of GP optimization schemes to the global optimum was shown in [87, Theorem 2] for the EI metric as being $O(n^{-\max\{\nu,1\}/d})$ where n is the number of sample points, ν is a smoothness parameter for the RKHS space of the kernel (for smooth, i.e. infinitely differentiable kernels ν tends to ∞ and the convergence rate is $O(n^{-1/d})$), and d is the dimension of the space $\mathcal{X}$. Similarly for the GP-UCB approach regret bounds of the form $O(\sqrt{n\gamma(n,d)})$ have been shown [88] under different assumptions on the regularity of the RKHS space, where $\gamma(n,d)$ is a factor that gets larger as the d (dimension of $\mathcal{X}$) grows.

In Section 5.4 we utilize the Expected improvement metric (denoted J_{EI}) is computed as follows for a Gaussian process $F'|\mathcal{D}_l$

$$J_{EI}(F'|\mathcal{D}_l, x) = \mathbb{E}[\max\{(F'|\mathcal{D}_l)(x) - F_{max}, 0\}] = \sigma_F(x)[v(x)\Phi(v(x)) + \phi(v(x))], \tag{5.11}$$

where $v(x) = (\mu_F(x) - y_{max})/\sigma_F(x)$ and $F_{max} = \max\{F(x_i) : x_i \in \mathcal{D}_l\}$. The GP optimization scheme then relies on finding a candidate point $x^* = \arg\max_{x\in\mathcal{X}} J_{EI}(x)$, performing an experiment to sample $F(x^*)$ and constructing $\mathcal{D}_{l+1} = \mathcal{D}_l \cup \{(x^*, F(x^*))\}$. Finally update the Bayesian posterior model to $F'|\mathcal{D}_{l+1}$. The process is then repeated till $\max_{x\in\mathcal{X}} J_{EI}(x)$ converges to 0, implying the F_{max} has converged to maximizer.

5.4 Learning and optimization with transient measurements

The approach with steady state measurements applied to the AWE system is first presented in Section 5.4.1, followed by the algorithm for incorporating transient measurements and dynamic constraints into a GP optimization scheme, in Section 5.4.2.

5.4.1 Steady state optimization

For an AWE system, a maximization of the unknown function F given by (5.7) is considered, where F represents the average tether force over a lemniscate loop for the closed loop system. A unknown constraint function G, (5.6), representing the minimum altitude attained during a loop in considered. Assuming that the wind is constant through out the experiment, a measurement for is taken by applying a set point u_s and waiting for the system to reach its steady state x_s. The objective and constraints functions are then all implicitly dependent on a single variable u_s and thus taking the steady state measurements for F and G, a initial data set $\mathcal{D}_l = \{(u_s)_i, F((u_s)_i), G((u_s)_i) : i = 1, \dots, l\}$ is considered for some finite l. Separate Gaussian process surrogate models are then constructed for the objective and constraint functions, denoted as $F'|\mathcal{D}_l$ and $G'|\mathcal{D}_l$ respectively. The constraint function is included into the GP optimization scheme by formulating it as a chance constraint over the Gaussian process thus giving a candidate selection problem,

$$\begin{aligned} u_s^* &= \arg\max_{u\in\mathcal{X}} && J_{EI}(F'|\mathcal{D}_l, u) && (5.12a)\\ & s.t. && \mathbb{P}[(G'|\mathcal{D}_l)(u) \geq 0] \geq 1 - \beta && (5.12b)\\ & && \sigma_{G'}(u) \leq \alpha\sigma_y^2 && (5.12c) \end{aligned}$$

for some constant tuning constants $\beta \in (0, 1)$ and $\alpha \in (0, 1)$. (5.12a) gives the candidate for maximizing the expected improvement according to the surrogate $F'|\mathcal{D}_l$, (5.12b) restricts the points admissible as candidates to a subset of $\mathcal{X}$ where the surrogate $G'|\mathcal{D}_l$ predicts with high confidence ($1 - \beta$ for $\beta \to 0$) that the constraint $G \geq 0$ is satisfied. The σ_y^2 is the variance for the prior Gaussian process and for $\alpha < 1$, (5.12c) restricts the candidate points to a subset of $\mathcal{X}$ such that the variance is at-least reduced by a factor α for the posterior by the data present in $\mathcal{D}_l$. Thus (5.12c) provides a trust region in which the surrogate model for G can be trusted. As α tends to 0, the trust region shrinks to only points where the function ahs already been sampled, and as α tends to 1, the trust region expands to the entire space $\mathcal{X}$. By tuning α and β we can thus tune the aggressiveness of the sampler in order to control the amount and number of violations that will be tolerated in the experiments.

The expected improvement metric J_{EI} is a non smooth and discontinuous function in u and thus optimization over J_{EI} is done by using a Monte-Carlo approach of sampling values of J_{EI} for a number of candidate u points and choosing a candidate that maximizes J_{EI} while satisfying (5.12b) and (5.12c).

Algorithm 5 Steady State Optimization

```
1: **Initialization.** Start with $l$ samples of feasible solutions ($\mathcal{D}_l$)
2: **Training.** Update the posteriors for the Gaussian process models $G'|\mathcal{D}_l$ and $F'|\mathcal{D}_l$
3: **EI maximization** Find $u_s^*$ as a maximizer for (5.12)
4: **Update** Set $\mathcal{D}_{l+1} = \mathcal{D}_l \cup F(u_s^*)$, $F_{max} = \max\{F_{max}, F(u_s^*)\}$, $l = l + 1$, $u_{best} = u_s^*$ if $F(u_s^*) > F_{max}$ and $G(u_s^*) > 0$
5: **if** $J_{EI}(u_s^*) < \epsilon$ **then**
6:     Use $\tilde{u} = u_{best}$
7: **else**
8:     Use $\tilde{u} = u_s^*$
9: Go to 2, and repeat
```

Algorithm 5 gives a non terminating version of the GP optimization such that it maximizes the objective function under the unknown constraints and then continues to use the best solution u_{best} as an output $\tilde{u}$, once the expected improvement has fallen below a small threshold ϵ. The output of the algorithm, $\tilde{u}$, is used to update the set point, $u_s = \tilde{u}$, after each optimization iteration. The algorithm thus also acts as a higher level optimizing controller that continuously monitors the performance and updates the set points to the lower level controller for the AWE system.

Section 5.5 provides simulation results for this steady state scheme under constant wind and changing wind conditions. The approach shows fast convergence to the optimum, under constant wind conditions, as the assumptions for this approach are met. Under changing wind conditions, the assumptions are violated, and performance degradation and constraint violation are observed.

The transient measurement based algorithm presented next increases the frequency at which the optimization provides a feedback $\tilde{u}$ to the low level controller and thus mitigates the issue observed here under changing wind conditions.

5.4.2 Optimization with transient measurements

A transient version of the GP optimization algorithm is constructed by incorporating the unknown dynamics $x_{k+1} = f(x_k, u_s^k, w_k)$ from (5.5) as part of the optimization problem.

Recalling the notation from Section 5.2.3, the system is said to be in steady state x_s when $x_s = f(x_s, u_s, w_k)$ for a wind condition w_k assumed constant for the period of one lemniscate loop (this is a much more relaxed assumption than the steady state, constant wind counterpart as a loop lasts over a time scale of about 10 seconds in practice). The transient dynamics for the k^{th} loop are given as $x_{k+1} = f(x_k, u_s^k, w_k)$.

The surrogate models for the dynamics, tether force and altitude constraint, f, F and G are learned as functions of x_k, u_s^k and w_k, thus being able to predicting the transient behavior for the system at any state, input and wind combination.

The expected improvement maximization problem is then posed as,

$$
\begin{aligned}
x_s^*, u_s^* \quad &= \arg\max_{(x,u)\in\mathcal{X}} \quad J_{EI}(F'|\mathcal{D}_l, (x,u,w_k)) && \text{(5.13a)}\\
&\quad s.t. \quad \mathbb{P}[(G'|\mathcal{D}_l)((x,u,w_k)) \geq 0] \geq 1-\beta && \text{(5.13b)}\\
&\quad \mathbb{P}[(G'|\mathcal{D}_l)((x_k,u,w_k)) \geq 0] \geq 1-\beta && \text{(5.13c)}\\
&\quad \mathbb{P}[||x-(f'|\mathcal{D}_l)((x,u,w_k))|| \leq \epsilon_s] \geq 1-\beta && \text{(5.13d)}\\
&\quad \sigma_{G'}((x,u,w_k)) \leq \alpha\sigma_y^2 && \text{(5.13e)}\\
&\quad \sigma_{G'}((x_k,u,w_k)) \leq \alpha\sigma_y^2 && \text{(5.13f)}\\
&\quad \mathbb{P}[||x-(f'|\mathcal{D}_l)((x_k,u,w_k))|| \leq \epsilon_r] \geq 1-\beta && \text{(5.13g)}
\end{aligned}
$$

The discussion for the constraint in (5.13) is presented as follows:

(i) The $\alpha\sigma_y^2$ in (5.13b) and (5.13c) play the same role as in the steady state case of reducing the search space for $(x,u)\in\mathcal{X}$ to points the surrogate model has been learned with a high confidence and thus can be relied upon. Thus (5.13b) and (5.13c) play the role of providing a trust region for the optimization.

(ii) (5.13d) imposes as a high confidence chance constraint, that the pair (x,u) is chosen such that it satisfies the steady state equation $x-f(x,u,w_k)$ to a high accuracy ϵ_s.

(iii) (5.13b) imposes that the steady state pair (x,u) chosen is such that it satisfies the altitude constraint G with high confidence.

(iv) (5.13c) similarly imposes that the u chosen is such that it does not lead to a constraint violation in the transient, given the current state of the system x_k.

(v) Finally, (5.13g), imposes the chosen steady state to be close to the current state x_k. This enforces a continuity in search from one iteration to the next. Without such a constraint the optimizer would pick target steady states jumping to far away points on each iteration and none of them would ever be reached. Thus the optimization will not make any progress as there is no continuity in what the optimizer tries to do from one iteration to the next. This constraint heuristically imposes a continuity plan for the optimizer between successive iterations.

(vi) Extra trust region constraints are not imposed on $f'|\mathcal{D}_l$ as the process shares the same data set with $G'|\mathcal{D}_l$, for which the trust region constraints are included.

Algorithm 6 then presents the GP optimization utilizing (5.13) as its candidate selection scheme. The output $\tilde{u}$ is used as before to set the controlled set point $u_s = \tilde{u}$, however this update is made at the end of every lemniscate loop, taking into account a new measurement of the state, objective, constraint and wind. Thus the model is updated much more frequently compared to the steady state optimization scheme, acquiring more data regarding the system quickly and adapting in its decisions to the changing wind conditions (thus providing a faster rate of feedback to the low level controller). Simulation results in Section 5.5 show the transient optimization algorithm to be significantly more robust to the changing wind conditions while also optimizing the performance to its maximum.

The F_{max} used for the expected improvement computation is a predicted value from the surrogate model $F'|\mathcal{D}_l$ and is computed as,

$$F_{max} = \max_{(x,s)\in\mathcal{X}} \mu_F(x,u,w_k), \quad \text{s.t.} \qquad (5.13b), (5.13d), (5.13e) \tag{5.14}$$

and

$$x_{best}, u_{best} = \arg\max_{(x,s)\in\mathcal{X}} \mu_F(x,u,w_k), \quad \text{s.t.} \qquad (5.13b), (5.13d), (5.13e) \tag{5.15}$$

Thus (5.14) predicts a best value for F at a steady state pairing (x,u), feasible for the constraint G with high confidence, at wind condition w_k and within the trust region of the altitude constraint.

Algorithm 6 Transient Optimization

```
1: **Initialization.** Start with $l$ samples of feasible solutions ($\mathcal{D}_l$)
2: **Training.** Update the posteriors for the Gaussian process models $f'|\mathcal{D}_l$, $G'|\mathcal{D}_l$ and $F'|\mathcal{D}_l$.
   Find $F_{max}$ using (5.14) and $u_{best}$ from (5.15)
3: **EI maximization** Find $x_s^*, u_s^*$ as a maximizer for (5.13)
4: **Update** $\mathcal{D}_{l+1} = \mathcal{D}_l \cup F(x_k, u_s^*, w_k)$, $l = l+1$
5: **if** $J_{EI}(u_s^*) < \epsilon$ and $\mathbb{P}[G'(x_k, u_{best}, w_k) > 0] \geq 1-\beta$ **then**
6:     Use $\tilde{u} = u_{best}$
7: **else**
8:     Use $\tilde{u} = u_s^*$
9: Go to 2, and repeat
```

5.5 Results

The dynamics model and controller described in Sections 5.2.1 and 5.2.2 are implemented as continuous time simulations to generate data corrupted with noise for testing the algorithms. The set points updated by the optimization schemes are applied to the low level controller in this continuous time simulation with a real time implementation, i.e., the simulation does not stop for the optimizer to finish computation. The two run in parallel processes on a computer and when the result for the optimizer is ready, it is passed to the controller immediately which updates its set points on the next crossing event for $\varphi_s^{(0)}$.

For constant wind conditions, we start both Algorithms 5 and 6 with $l = 15$ initial training points in the feasible set. This is possible from prior experience of the operator. For Algorithm 5, set points are repeated for 10 loops, until the kite reaches a steady trajectory. On the other hand, for Algorithm 6 we take measurements after every loop. Both algorithms show convergence (within 20-30 samples) to within 5% of the optimum calculated using a numerical optimal control solver GPOPS-II, see [89], which makes use of the model and has full control throughout the trajectory and not just at the switching surfaces. The wind dependence of the objective value has been normalised so that the results are comparable across different wind conditions. The normalization is done by dividing the tether force by w_k^2. This approximately gets rid of the wind

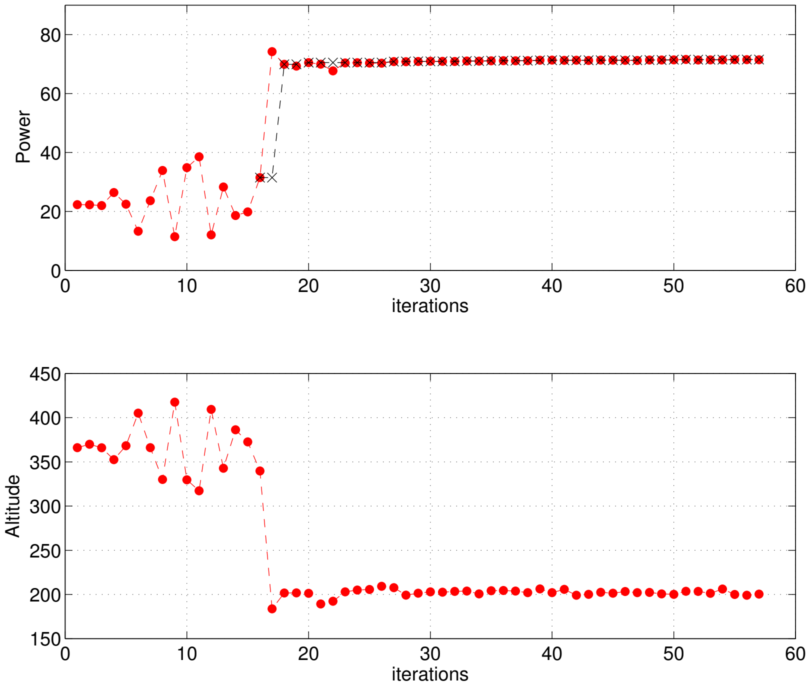


Figure 5.3: Convergence of steady state optimization (Algorithm 5) in constant wind condition and 200m altitude constraint. Red circles represent measured values and black ones represent F_{max} (predicted)

speed squared dependence in the force term, as the speed of the kite relative to the wind is known to be proportional to the wind speed.

Figure 5.3 shows the performance of the steady state optimization (Algorithm 5) in constant wind. Table 5.1 compares with Algorithm 6 for the same conditions. Both algorithms converge close to the optimum, however Algorithm 6 using transient predictions and frequent updates can avoid large constraint violations and converges much faster to the optimum. Also Algorithm 6 guarantees constraint satisfaction both in predicted stationary orbit and during transients. The zero constraint violation in Algorithm 6 is due to the fact that the algorithm remains conservative and the minimum altitude achieved over all the iterations in maintained strictly above the $z_{min} = 200$ threshold.

Method	Final value	Max Violation (m)	Loops
Algorithm 1 (2 surfaces)	71.17	29.	450
Algorithm 1 (4 surfaces)	72.60	4.77	300
Algorithm 2 (2 surfaces)	71.05	0.00	65
GPOPS-II	74.61	0.00	-

Table 5.1: Performance comparisons in constant wind

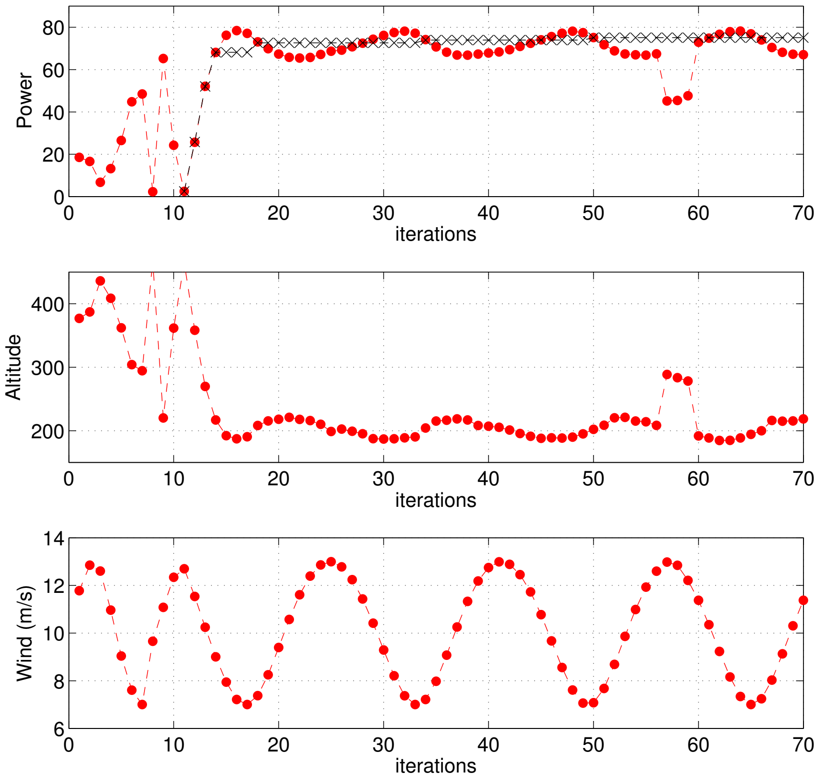


Figure 5.4: Maximum force tracking with steady state optimization (Algorithm 5) under varying wind conditions.

Algorithm 5 was mainly developed under the assumption of constant wind and is not well suited for handling varying wind conditions. This is due to the large difference between the time at which the decision is made and the time at which the system reaches the corresponding steady state. Algorithm 6 overcomes this limitation by sampling in transients and thus increasing the frequency at which decisions are revised.

Table 5.2 summarizes performance results in varying wind conditions. Algorithm 6 can track the optimal power without significant violations and is able to converge to different optimal set-points as the exogenous conditions vary. Figure 5.5 shows the progress of the transient algorithm under varying wind conditions (and Figure 5.4 for the steady state one), while Figure 5.7 shows the tracked trajectory under these conditions. Figure 5.6 shows the evolution of the system states and set points with time and wind. Algorithm 5 performs slightly better because it incorporates a large constraint violation (15.35m) for which the power production (or towing force) is much favorable.

Decision space	Final value	Max Violation (m)	Loops
Algorithm 1 (2 surfaces)	75.09	15.35	320
Algorithm 2 (2 surfaces)	73.84	2.90	47
GPOPS-II	74.61	0.00	-

Table 5.2: Performance comparisons in varying wind

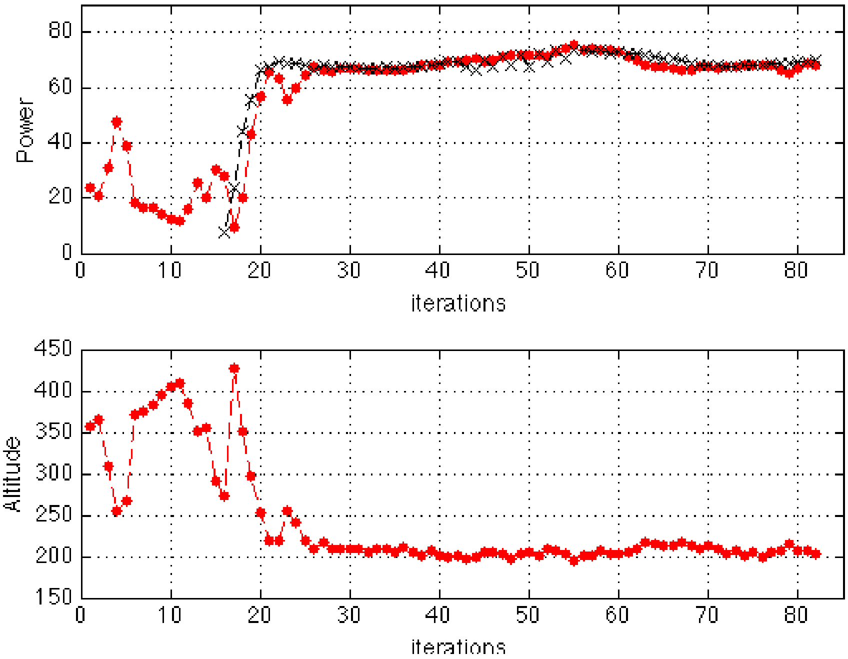


Figure 5.5: Power and Altitude tracking for 200m altitude constraint and varying wind condition using transient optimization (Algorithm 6)

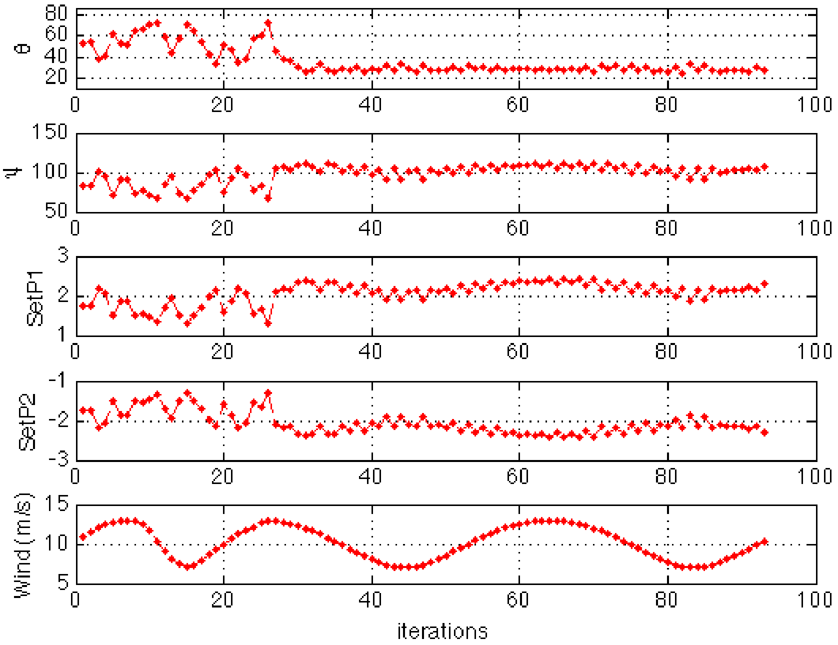


Figure 5.6: System states, set-points and wind measurements through the iterations of Algorithm 6

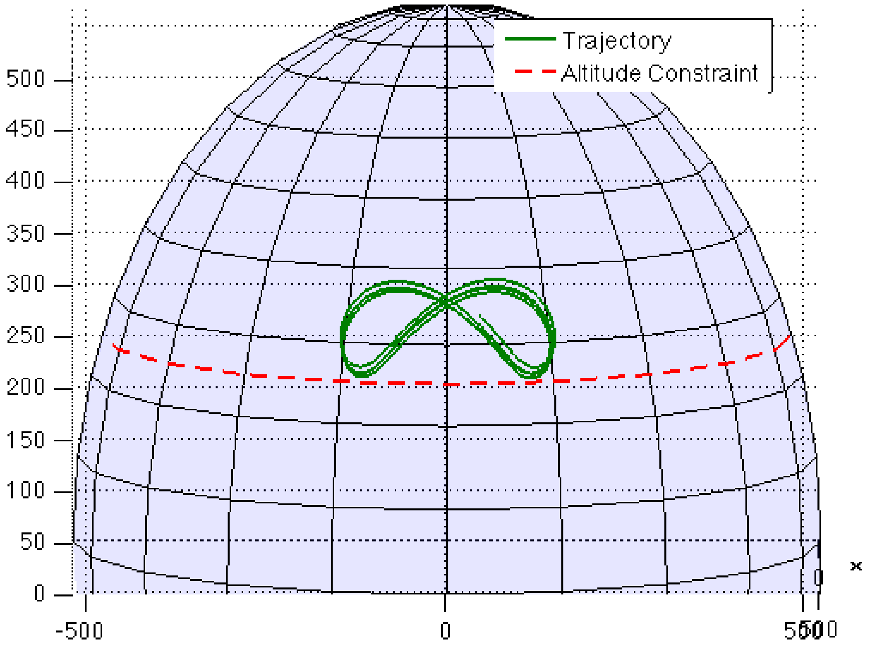


Figure 5.7: Trajectory of kite with Algorithm 2 in varying wind conditions

5.6 Conclusions

Two algorithms for optimizing the towing force produced by an AWE system under minimum altitude constraints were presented. Both methods use a model free Gaussian process optimization based approach to progressively learn the system dynamics, objective and constraint functions. The second method can utilize transient measurements to converge faster to the optimum and better adapt to changing wind conditions. The results are compared to an off-line optimal control numerical solver (with knowledge of the model and full input control) and the optimal values found by the GP optimization methods were found to be with 5% of the predicted values by the optimal control solver.

Chapter 6

A Nonlinear Adaptive Controller for Airborne Wind Energy Systems

A direct-adaptive, nonlinear path following controller for a kite based airborne wind energy system is presented in presence of system and environmental parametric uncertainties. For a given reference geometric path, necessary conditions for closed-loop convergence of the kite to a tube centered around the reference path are provided. An adaptive control law for the case of unknown wind vector and kite parameters is presented. The effectiveness of the approach is demonstrated via numerical simulations for multiple shapes of geometric paths and for varying tether length references.

6.1 Introduction

A common operating mode for kite based Airborne Wind Energy (AWE) systems is the "pumping cycle". This operation requires the kite to reel out at a desired rate as it flies a high energy extraction manoeuvre, then reel back in with a low energy consumption manoeuvre to produce a net positive energy generation cycle (see, e.g., [56, 57]).

While the desired trajectory for the vehicle can be pre-computed using numerical optimal control solvers ([58, 59, 60]), the motion control of the system presents numerous challenges due to the nonholonomic properties of the system and the limited control inputs available. In fact, since the main driving force is provided by the wind, the vehicle can only follow time-profiles along the reference trajectory that are coherent with the wind. Due to this reason, most of the control schemes as explored in [90],[81],[63],[64], focus on tracking motion of the vehicle in a plane perpendicular to the tether and control the tether length in a decoupled fashion. This alone, cannot guarantee closed loop bounded tracking of an optimal/desired trajectory. Nonlinear MPC schemes for trajectory tracking have also been explored in literature (e.g., [61, 62]) which require real-time estimation of wind speed and vehicle parameters.

Motivated by these observations, a path-following controller, where the reference is not a time-parametrised trajectory but rather a parameterized geometric trajectory is proposed. The path parameter is then driven by the controller resulting in feasible trajectories coherent with the wind field. The controller is also extended to the case of unknown wind velocity vector and vehicle

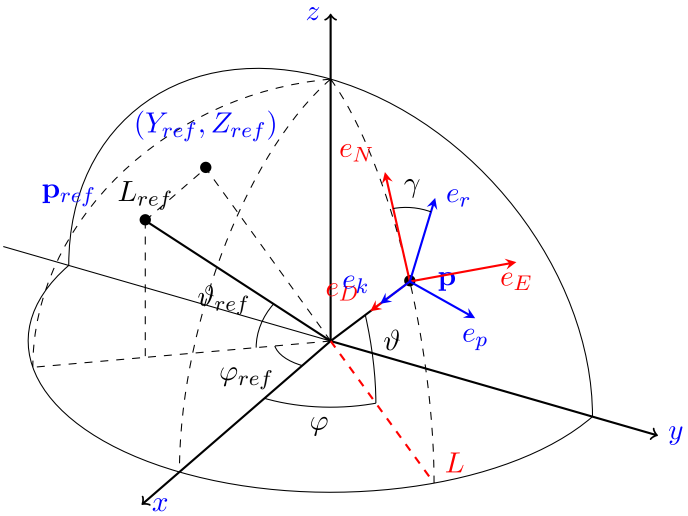


Figure 6.1: Coordinate frames, Kite and Reference Path States

parameters a direct-adaptive control scheme that guarantees bounded errors for path tracking and bounded parameter estimation via dynamic parameter update law. In both cases, necessary conditions for convergence of the vehicle position to an arbitrary small neighbourhood of the desired trajectory are provided.

The organization of the chapter is as follows. In Section 6.2 a kinematic model for the kite is introduced. Section 6.3 discusses the main result, controller design, for the nominal and adaptive cases. Section 6.4 gives the results from the numerical tests of the controller. Section 6.5 summarizes the results presented. A collection of the exact expressions and background computation required for the controller are presented in the Appendix.

6.2 Kite model

6.2.1 Coordinate frames

Coordinate frames consistent with those used in [64] are presented below. Fig. 6.1 shows a graphical illustration for the same.

An inertial frame $\{G\}$ is attached to the ground, with basis vectors (x, y, z), and a moving frame $\{K\}$ is attached to the body of the kite, with basis vectors (e_r, e_p, e_k). Let $\mathbf{p}$ denote the position of the origin of the kite frame written in the ground frame, and let (L, ϑ, φ) denote its polar coordinate representation. Here, L represents the tether length and ϑ and φ denote the elevation and azimuth angle, respectively.

For any fixed tether length the kite moves on a sphere. An intermediate right handed coordinate frame $\{N\}$ centered in $\mathbf{p}$ with basis vectors (e_N, e_E, e_D) is considered, with e_N pointing in the direction of the sphere's apex and e_D pointing towards the sphere's center. Using this intermediate frame, and assuming always non-zero velocity, we denote by γ the angle that the kite's velocity vector projected on the $e_N - e_E$ plane, tangent to the sphere, forms with e_N. Then, the kite frame

$\{K\}$ is obtained by rotating the frame $\{N\}$ about e_D by the angle γ.

Let R_{GN}, R_{NK} denote the rotation transformation matrices from the ground frame $\{G\}$ to local north frame $\{N\}$, and from the local north $\{N\}$to the kite's body fixed frame $\{K\}$, respectively, i.e.,

$$\begin{aligned} R_{GN} &= \begin{pmatrix} -\sin\theta\cos\varphi & -\sin\varphi & -\cos\vartheta\cos\varphi \\ -\sin\vartheta\sin\varphi & \cos\varphi & -\cos\vartheta\sin\varphi \\ \cos\vartheta & 0 & -\sin\vartheta \end{pmatrix} \\ R_{NK} &= \begin{pmatrix} \bar{R}_{NK} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}, \; \bar{R}_{NK} = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix}. \end{aligned}$$

6.2.2 Kinematic model

The model presented below is similar to the model used in [81], but extended to a variable tether length setting and expressed in different coordinate frames. The kinematic model of the kite can be written as

$$\begin{aligned} \begin{pmatrix} L & 0 \\ 0 & L\cos\vartheta \end{pmatrix} \begin{pmatrix} \dot{\vartheta} \\ \dot{\varphi} \end{pmatrix} = \; & \bar{R}_{NK} \begin{pmatrix} 1 & 0 & -\mathbf{E} \\ 0 & 0 & 0 \end{pmatrix} R_{NK}^T R_{GN}^T v_w - \bar{R}_{NK} \begin{pmatrix} \mathbf{E}z \\ 0 \end{pmatrix} \\ & \dot{L} = z \qquad \dot{z} = u_z \\ & \dot{\gamma} = k v_s(\vartheta, \varphi, \mathbf{E}, v_w)\delta \end{aligned} \tag{6.1}$$

where $\mathbf{E}$ is an aerodynamic parameter of the kite, called glide ratio, and v_w is the wind velocity vector in the ground frame and v_s is some nonlinear function involving the state, parameter and wind velocity, that affects the steering gain for the vehicle. The control inputs of the physical system are the reel-out rate z and kite deflection δ for turning. The controller is designed taking $(u_z, \dot{\gamma})$ as virtual inputs to the kite. A proportional controller is used to get the input δ that approximately tracks the resulting γ demanded by our controller and thus the effect to the complicated steering gain is addressed using this cascading of controllers for the steering.

6.2.3 Reference path

The reference path is described by specifying its projection on the $y-z$ plane of the ground frame $\{G\}$ and denoted as the smooth parameterized curve $(Y_{ref}(\tau), Z_{ref}(\tau))$ for a scalar parameter τ. Note that for any tether length there exists a trajectory that, projected in the $y-z$ plane of $\{G\}$, satisfies the desired assignment. Also note that τ is a parameter controlled by the controller and as a result the speed with which the reference moves along the path is controlled through τ.

The desired path is assumed to have no stationary points, i.e., $||\frac{\partial(Y_{ref}(\tau), Z_{ref}(\tau))}{\partial\tau}|| \neq 0$ for any $\tau \in \mathbb{R}$, or in other words, as τ changes the point should move along the curve with a non-zero speed. Two such path projections, used in our numerical studies, for a "figure of eight" lemniscate trajectory and an ellipsoidal trajectory are shown in (6.38),(6.39).

The length of the tether is controlled specifying the desired velocity of the tether $z_{ref}(t)$, which is time parameterized unlike the Y_{ref}, Z_{ref}.

In view of these observations, the output of the system is defined as

$$y = \begin{pmatrix} L & 0 & 0 \\ 0 & L\cos\vartheta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \vartheta \\ \varphi \\ z \end{pmatrix}, \quad y_{ref} = \begin{pmatrix} L & 0 & 0 \\ 0 & L\cos\vartheta_{ref} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \vartheta_{ref} \\ \varphi_{ref} \\ z_{ref} \end{pmatrix} \tag{6.2}$$

6.3 Controller design

6.3.1 Error definition

This section introduces the error space utilized for the design of the path following controller. Similar to [91, 92], we consider the tracking error vector

$$\mathbf{e} = R_{NK}^T(y - y_{ref}) - \epsilon \tag{6.3}$$

for a given vector $\epsilon \in \mathbb{R}^3$ with non-zero norm. Note that as the norm of the error vector goes to zero, the distance $y - y_{ref}$ converges to $\|\epsilon\|$, which can be made arbitrarily small.

6.3.2 Error dynamics

Taking the derivative for the error as defined in (6.3), we get error dynamics in the form,

$$\dot{\mathbf{e}} = \dot{\gamma}\tilde{S}\mathbf{e} + f(\mathbf{x}) + f_\lambda(\mathbf{x})\lambda + g(\mathbf{x})u + R(\beta)\tilde{S}\epsilon_o\dot{\beta} \tag{6.4}$$

where $\mathbf{x} = (\vartheta, \varphi, \gamma, L, z, \dot{z}_{ref}, \tau)^T$ is the system state, $u = (\dot{z}, \dot{\gamma}, \dot{\tau})^T$ is the control input and $\lambda = (\mathbf{E}, v_w, \mathbf{d})^T$ is a vector of unknown system parameters. The bilinear dependence on $\mathbf{E} \cdot v_w$ in (6.1) is denoted and estimated as an independent parameter $\mathbf{d}$. The exact expressions for $\tilde{S}, f, f_\lambda, g$ are presented in the appendix in (6.43). We use

$$\epsilon = R(\beta)\epsilon_o \tag{6.5}$$

where ϵ_o is a constant vector in R^3 with non-zero norm and $R(\beta)$ shown in (6.40) is a rotation matrix with state dependent β. From the expression for the determinant of g presented in (6.44) it can be seen that using $\beta = \zeta - \gamma + \pi/2$, $g(\mathbf{x})$ is guaranteed to be invertible at all times. ζ is again a state dependent term defined in (6.42). Without the state varying β, $g(\mathbf{x})$ will lose rank for certain states and the system will lose feedback linearizability at those states. For the state varying β chosen above, we can show, $\dot{\beta}$ to be of the form,

$$\dot{\beta} = W(\mathbf{x})^T u \tag{6.6}$$

where $W(\mathbf{x})$ is a vector in $\mathbb{R}^3$. The error dynamics thus take the form,

$$\dot{\mathbf{e}} = f(\mathbf{x}) + f_\lambda(\mathbf{x})\lambda + (g(\mathbf{x}) + h(\mathbf{x}))u \tag{6.7}$$

where $h(\mathbf{x}) = R(\beta)\tilde{S}\epsilon_o W(\mathbf{x})^T + \begin{pmatrix} \mathbf{0} & \tilde{S}\mathbf{e} & \mathbf{0} \end{pmatrix}$.

This brings the error dynamics in the class of systems for which we will prove local convergence of the closed loop system to an ultimate bound in the following theorem, for cases of known and unknown system parameter vector λ.

6.3.3 Main result

We consider systems of the form,

$$\dot{\mathbf{e}} = f(\mathbf{x}) + f_\lambda(\mathbf{x})\lambda + G(\mathbf{x})u \tag{6.8}$$

where $\mathbf{x} \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ are the system states and inputs respectively. $\mathbf{e} = l(\mathbf{x}) \in \mathbb{R}^m$ is some nonlinear output of the system state. $G(\mathbf{x})$ is possibly non-invertible at certain states, but can be written in the form,

$$G(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}) \tag{6.9}$$

with $g(\mathbf{x})$ always guaranteed to be invertible and $||g|| \in [a, b]$ for some finite positive constants a, b and $||h|| \leq \Delta$ for a finite positive constant Δ. f, f_λ, g are known functions satisfying $||\frac{df(\mathbf{x})}{dt}|| \leq \Delta_1 ||\mathbf{e}||\,||u||$, $||\frac{df_\lambda(\mathbf{x})}{dt}|| \leq \Delta_2 ||\mathbf{e}||\,||u||$, $||\frac{dg}{dt}|| \leq \Delta_3 ||\mathbf{e}||\,||u||$ and $||f_\lambda|| < \Delta_4$ where $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ are scalar positive and finite constants.

λ is a vector of system parameters in $\mathbb{R}^p$. When the parameters λ are unknown we use an online estimate of the parameter denoted by $\bar{\lambda}$ and design for an update rule $\dot{\bar{\lambda}}$ assuming λ to be unknown constants. We also denote any offline a priori estimates of the parameters by $\hat{\lambda}$. In the nominal case when λ is known we simply set $\bar{\lambda} = \hat{\lambda} = \lambda$ and $\dot{\bar{\lambda}} = 0$ in our adaptive control law.

Note, since $f(\mathbf{x}), f_\lambda(\mathbf{x}), g(\mathbf{x}), h(\mathbf{x})$ are functions of only $\mathbf{x}$ we will drop the explicit notation and denote the function evaluated at $\mathbf{x}$ as f, f_λ, g, h. Also we use the following vector operations: $\mathrm{Tanh}(\Theta)$ acting on a vector Θ represents an element-wise $\tanh(\cdot)$ operating on the elements of Θ. $\mathrm{Cosh}(\Theta), \mathrm{Sech}(\Theta)$ are diagonal matrices with diagonal entries being the corresponding element-wise operations on elements of vector Θ.

Theorem 6.1. *For a system of the form* (6.8)*, a control law,*

$$u = q + N \cdot \mathrm{Tanh}(\Theta) \tag{6.10}$$

with

$$\begin{aligned}
\dot{\Theta} &= \mathrm{Cosh}^2(\Theta) N^{-1} g^{-1} \nu && (6.11)\\
\nu &= -K_r \mathbf{r} - m + \chi && (6.12)\\
\mathbf{r} &= f + f_\lambda \bar{\lambda} + gu + K\mathbf{e} && (6.13)\\
m &= \dot{f} + \dot{f}_\lambda \bar{\lambda} + \dot{g}u + Khu + f_\lambda \dot{\bar{\lambda}} && (6.14)\\
\chi &= -k_\Theta g M \,\mathrm{Tanh}(\Theta) && (6.15)
\end{aligned}$$

and parameter update law

$$\dot{\bar{\lambda}} = K_\lambda f_\lambda^T (\mathbf{e} + K^{-1}\mathbf{r}) - \sigma(\bar{\lambda} - \hat{\lambda}), \tag{6.16}$$

enforces $\mathbf{e}, \mathbf{r}, \Theta, \bar{\lambda}$ to converge to a bounded set tuned using the tuning variables in the control scheme, v.i.z., $K, K_r, k_\Theta, N, K_\lambda, \sigma$.

Proof: Consider the error vector defined in (6.3) with

$$\dot{\mathbf{e}} = f + f_\lambda \lambda + gu + hu. \tag{6.17}$$

We would like $f + f_\lambda \lambda + gu$ to be close to $-K\mathbf{e}$, for a diagonal positive definite constant matrix $K \succ 0$, with bounded inputs u. We thus proceed in a backstepping fashion by defining the backstepping variable

$$\mathbf{r} = f + f_\lambda \bar{\lambda} + gu + K\mathbf{e} \tag{6.18}$$

and driving it to zero. Note that, since we only have the estimate $\bar{\lambda}$ of the parameter vector λ, such variable is defined using the estimate, and later we will design a suitable estimator to compensate for the effect of such discrepancy. Combining (6.17) with (6.18) in the nominal case of known parameters, i.e., $\bar{\lambda} = \lambda$, results in

$$\dot{\mathbf{e}} = -K\mathbf{e} + \mathbf{r} + hu \tag{6.19}$$

whereas in the case of unknown parameters we have

$$\dot{\mathbf{e}} = -K\mathbf{e} + \mathbf{r} + f_\lambda(\lambda - \bar{\lambda}) + hu. \tag{6.20}$$

Differentiating the backstepping variable $\mathbf{r}$ results in

$$\dot{\mathbf{r}} = \dot{f} + \dot{f}_\lambda \bar{\lambda} + \dot{g}u + K\dot{\mathbf{e}} + f_\lambda \dot{\bar{\lambda}} + g\dot{u} \tag{6.21}$$

where the terms $\dot{f}$, $\dot{f}_\lambda$, and $\dot{g}$ denote the time derivatives of functions f, f_λ, and g, respectively. Plugging (6.20) into (6.21),

$$\dot{\mathbf{r}} = m - K^2\mathbf{e} + K\mathbf{r} + g\dot{u} + Kf_\lambda(\lambda - \bar{\lambda}) \tag{6.22}$$

where the term m is defined as

$$m = \dot{f} + \dot{f}_\lambda \bar{\lambda} + \dot{g}u + Khu + f_\lambda \dot{\bar{\lambda}}.$$

Now considering the Lyapunov function,

$$\begin{aligned} V_1 &= \frac{1}{2}(\mathbf{e}^T\mathbf{e} + \mathbf{r}^T K^{-2}\mathbf{r}) && (6.23)\\ \dot{V}_1 &= \mathbf{e}^T\dot{\mathbf{e}} + \mathbf{r}^T K^{-2}\dot{\mathbf{r}} \\ &= -\mathbf{e}^T K\mathbf{e} + \mathbf{e}^T hu + (\mathbf{e}^T + K^{-1}\mathbf{r}^T) f_\lambda(\lambda - \bar{\lambda}) \\ &\quad + \mathbf{r}^T K^{-2}(m + K\mathbf{r} + g\dot{u}) && (6.24) \end{aligned}$$

In what follows, we proceed by defining a suitable input $\dot{u}$ to enforce the desired decrease of the lyapunov function. Although, the term $e^T hu$ in the inequality (6.24) cannot be cancelled, we design $\dot{u}$ explicitly enforcing a bounded u. This can be achieved by defining

$$u = q + N \cdot \mathrm{Tanh}(\Theta)$$

with first time derivative

$$\dot{u} = N\,\mathrm{Sech}^2(\Theta)\dot{\Theta} \tag{6.25}$$

where the constants q and N are design parameters and Θ is an internal state of the controller.

Therefore, designing $\dot{u}$ such that

$$g\dot{u} = -K_r\mathbf{r} - m + \chi,$$

for some term χ, is equivalent to choosing

$$\dot{\Theta} = \mathrm{Cosh}^2(\Theta)N^{-1}g^{-1}(-K_r\mathbf{r} - m + \chi), \tag{6.26}$$

The term χ, in the following, is used to maintain the internal state of the controller Θ bounded and avoid numerical integration problems.In fact, an unstable internal state Θ will eventually drive $\mathrm{Sech}(\Theta)$ to $\mathbf{0}$, and the $\dot{\Theta}$ resulting from

$$gN\,\mathrm{Sech}^2(\Theta)\dot{\Theta} = -K_r\mathbf{r} - m + \chi$$

will be numerically infeasible to integrate. Toward this goal, we update the Lyapunov function introducing an extra term,

$$V_2 \;=\; V_1 + \frac{1}{2}k_{\Theta}^{-1}\,\mathrm{Tanh}^T(\Theta)\,\mathrm{Tanh}(\Theta) \tag{6.27}$$

where $k_\Theta > 0$ is a positive scalar constant. Computing the first time derivative combining with (6.24) and (6.26) results in

$$\begin{aligned}
\dot{V}_2 &= \dot{V}_1 + k_{\Theta}^{-1}\,\mathrm{Tanh}^T(\Theta)N^{-1}g^{-1}(-K_r\mathbf{r} - m + \chi) \\
&= -\mathbf{e}^T K\mathbf{e} - \mathbf{r}^T K^{-2}(K_r - K)\mathbf{r} + \mathbf{e}^T hq \\
&\quad +\mathbf{e}^T hN\,\mathrm{Tanh}(\Theta) + \mathbf{r}^T K^{-2}\chi \\
&\quad +k_{\Theta}^{-1}\,\mathrm{Tanh}^T(\Theta)N^{-1}g^{-1}(-K_r\mathbf{r} - m + \chi) \\
&\quad +(\mathbf{e}^T + K^{-1}\mathbf{r}^T)f_\lambda(\lambda - \bar{\lambda}).
\end{aligned}$$

Choosing

$$\chi = -k_\Theta gM\,\mathrm{Tanh}(\Theta) \qquad M = \min(N, I)$$

where M is the element wise minimum of the matrices N and I (the identity matrix), such that,

$$0 < N^{-1}M \leq I \qquad 0 < M \leq I$$

results in

$$\begin{aligned}\dot{V}_2 &= -\mathbf{e}^T K\mathbf{e} - \mathbf{r}^T K^{-2}(K_r - K)\mathbf{r} + \mathbf{e}^T hq \\ &\quad +\mathbf{e}^T hN \operatorname{Tanh}(\Theta) - k_\Theta \mathbf{r}^T K^{-2} gM \operatorname{Tanh}(\Theta) \\ &\quad -k_\Theta^{-1} \operatorname{Tanh}^T(\Theta) N^{-1} g^{-1} K_r \mathbf{r} \\ &\quad -k_\Theta^{-1} \operatorname{Tanh}^T(\Theta) N^{-1} g^{-1} m \\ &\quad - \operatorname{Tanh}^T(\Theta) N^{-1} M \operatorname{Tanh}(\Theta) \\ &\quad +(\mathbf{e}^T + K^{-1}\mathbf{r}^T) f_\lambda(\lambda - \bar{\lambda}). \end{aligned} \tag{6.28}$$

Known parameter case. For the case of known parameter λ we have $\dot{\bar{\lambda}} = 0, \bar{\lambda} = \lambda$ Using the norm inequalities,

$$\begin{aligned}\dot{V}_2 &\leq -\mathbf{e}^T K\mathbf{e} - \mathbf{r}^T K^{-2}(K_r - K)\mathbf{r} \\ &\quad - \operatorname{Tanh}^T(\Theta) N^{-1} M \operatorname{Tanh}(\Theta) \\ &\quad +||\mathbf{e}||\,||h||\,||q|| + ||\mathbf{e}||\,||h||\,||N|| \\ &\quad +||\mathbf{r}||\,||K^{-2}||\,||g|| \\ &\quad +k_\Theta^{-1}||N^{-1}||\,||g^{-1}||\,||K_r||\,||\mathbf{r}|| \\ &\quad +k_\Theta^{-1}||N^{-1}||\,||g^{-1}||\,||m|| \end{aligned} \tag{6.29}$$

By earlier assumptions on bounds for $||\dot{f}||, ||\dot{f}_\lambda||, ||\dot{g}||, ||h||$

$$\begin{aligned}||m|| &\leq ||\dot{f}|| + ||\dot{f}_\lambda||\,||\bar{\lambda}|| + ||\dot{g}||\,||u|| + ||K||||h||||u|| \\ &\leq \Delta_1||\mathbf{e}|| + \Delta_2||\mathbf{e}||\,||\bar{\lambda}|| + \Delta_3||\mathbf{e}|| + ||K||\,\Delta \end{aligned} \tag{6.30}$$

Implying,

$$\begin{aligned}\dot{V}_2 &\leq -k_1||\mathbf{e}||^2 - k_2||\mathbf{r}||^2 - k_3||\operatorname{Tanh}(\Theta)||^2 \\ &\quad +||\mathbf{e}||(\Delta\,(||q|| + ||N||) + k_4(\Delta_1 + \Delta_2||\bar{\lambda}|| + \Delta_3)) \\ &\quad +||\mathbf{r}||(k_1^{-2}\,||g|| + k_4\,k_6\,) + k_4||K||\Delta \end{aligned} \tag{6.31}$$

where

$$\begin{aligned} k_1 &:= \lambda_{min}(K) \\ k_2 &:= \lambda_{min}(K^{-2}(K_r - K)) \\ k_3 &:= \lambda_{min}(N^{-1}M) \\ k_4 &:= k_\Theta^{-1}||N^{-1}||\,||g^{-1}|| \\ k_5 &:= ||K_r|| \end{aligned}$$

where for a generic matrix A, the term $\lambda_{min}(A)$ denotes the minimum singular value of A and where the constant terms K, K_r, N, k_Θ, with the restriction $K_r > K$, are design parameters introduced earlier. Note that, in the Lyapunov inequality (6.31), as the terms e and r grow, the quadratic negative terms will eventually dominate the positive linear and bounded terms, resulting in the standard ultimately bounded behaviour of e and r. Similar applies to the term Θ, although here, since the $\mathrm{Tanh}(\cdot)$ is not a radially unbounded function, an excessive magnitude of the positive terms might cause Θ to be unbounded and therefore care should be taken in selection of the design parameters.

Further increasing k_1, k_Θ and k_2 allows us to reduce the ultimate bound on $\mathbf{e}$ and $\mathbf{r}$ axes.

Unknown parameter case. For the case of unknown parameter λ, we consider the Lyapunov function,

$$\begin{aligned} V_3 &= V_2 + \frac{1}{2}(\lambda - \bar{\lambda})^T K_\lambda^{-1}(\lambda - \bar{\lambda}) && (6.32)\\ \dot{V}_3 &= \dot{V}_2 - (\lambda - \bar{\lambda})^T K_\lambda^{-1}\dot{\bar{\lambda}} && (6.33)\end{aligned}$$

Using (6.28),

$$\begin{aligned} \dot{V}_3 = & -\mathbf{e}^T K\mathbf{e} - \mathbf{r}^T K^{-2}(K_r - K)\mathbf{r} + \mathbf{e}^T hq \\ & + \mathbf{e}^T hN\,\mathrm{Tanh}(\Theta) - k_\Theta \mathbf{r}^T K^{-2} gM\,\mathrm{Tanh}(\Theta) \\ & - k_\Theta^{-1}\,\mathrm{Tanh}^T(\Theta)N^{-1}g^{-1}K_r\mathbf{r} \\ & - k_\Theta^{-1}\,\mathrm{Tanh}^T(\Theta)N^{-1}g^{-1}m \\ & - \mathrm{Tanh}^T(\Theta)N^{-1}M\,\mathrm{Tanh}(\Theta) \\ & + (\lambda - \bar{\lambda})^T(f_\lambda^T(\mathbf{e} + K^{-1}\mathbf{r}) - K_\lambda^{-1}\dot{\bar{\lambda}}). \end{aligned} \qquad (6.34)$$

Choosing the parameter update law,

$$\dot{\bar{\lambda}} = K_\lambda f_\lambda^T(\mathbf{e} + K^{-1}\mathbf{r}) - \sigma(\bar{\lambda} - \hat{\lambda}) \qquad (6.35)$$

where K_λ and σ are positive definite, diagonal matrices, we obtain

$$\begin{aligned} \dot{V}_3 = & -\mathbf{e}^T K\mathbf{e} - \mathbf{r}^T K^{-2}(K_r - K)\mathbf{r} + \mathbf{e}^T hq \\ & + \mathbf{e}^T hN\,\mathrm{Tanh}(\Theta) - k_\Theta \mathbf{r}^T K^{-2} gM\,\mathrm{Tanh}(\Theta) \\ & - k_\Theta^{-1}\,\mathrm{Tanh}^T(\Theta)N^{-1}g^{-1}K_r\mathbf{r} \\ & - k_\Theta^{-1}\,\mathrm{Tanh}^T(\Theta)N^{-1}g^{-1}m \\ & - \mathrm{Tanh}^T(\Theta)N^{-1}M\,\mathrm{Tanh}(\Theta) \\ & - \sigma(\lambda - \bar{\lambda})^T(\bar{\lambda} - \hat{\lambda}). \end{aligned} \qquad (6.36)$$

The last term in $\dot{V}_3 = -\sigma(\lambda - \bar{\lambda})^T(\bar{\lambda} - \hat{\lambda})$ is always negative definite outside a box in $\mathbb{R}^p$ defined by the values of λ and $\hat{\lambda}$ and this keeps the estimates $\bar{\lambda}$, bounded. The norm of m will now have the bound

$$\begin{aligned} ||m|| &\leq ||\dot{f}|| + ||\dot{f}_\lambda||\,||\bar{\lambda}|| + ||\dot{g}||\,||u|| + ||K||||h||||u|| \\ &\quad +||f_\lambda||^2\,||K_\lambda||\,(||\mathbf{e}|| + ||K^{-1}||\,||\mathbf{r}||) + ||\sigma||\,||\lambda - \hat{\lambda}|| \\ &\leq \Delta_1||\mathbf{e}|| + \Delta_2||\mathbf{e}||\,||\bar{\lambda}|| + \Delta_3||\mathbf{e}|| + ||K||\,\Delta \\ &\quad +\Delta_4^2||K_\lambda||\,||\mathbf{e}|| + \Delta_4^2||K_\lambda||\,||K^{-1}||\,||\mathbf{r}|| \\ &\quad +||\sigma||\,||\lambda - \hat{\lambda}|| \end{aligned} \tag{6.37}$$

Since the $||m||$ is still bounded linearly in terms of $||\mathbf{e}||, ||\mathbf{r}||$, $\dot{V}_3$ also takes the same form as (6.31) with a added constant $||\sigma||\,||\lambda - \hat{\lambda}||$. Thus with some reasonable a priori estimate of parameters $\hat{\lambda}$ such that $||\lambda - \hat{\lambda}||$ is bounded, the controller will converge to a bounded ellipsoid in the $(\mathbf{e}, \mathbf{r}, \Theta, \bar{\lambda})$ space. The boundary of the ellipsoid satisfies the equation $\dot{V}_3 = 0$ ($\dot{V}_2 = 0$, when λ is known). As long as for the chosen tuning variables and apriori estimate $\hat{\lambda}$, the ellipsoid boundary satisfies the strict inequality $||\,\mathrm{Tanh}(\Theta)||_\infty < 1$, the convergence of states $(\mathbf{e}, \mathbf{r}, \Theta, \bar{\lambda})$ to the ellipsoid is guaranteed. □

6.4 Numerical results

We test the control scheme described in Theorem 6.1 for our kite system, with the path following error system dynamics as described in (6.4) under constant but unknown wind vector $v_w = (10, 0, 0)$, constant unknown glide ratio $\mathbf{E} = 5$ with apriori guesses $\hat{v}_w = (9, 0, 0), \hat{\mathbf{E}} = 6, \hat{\mathbf{d}} = (45, 0, 0)$.

The controller tuning parameters were chosen as, $\epsilon_o = (-0.1, -0.1, 0)^T$, $K = \mathrm{diag}(4, 4, 10)$, $K_r = \mathrm{diag}(20, 20, 20)$, $k_\Theta = 10$, $\sigma = 0.1$, $K_\lambda = \mathrm{diag}(0.5, 0.5, 0.5, 0.5, 0.5, 0.01, 0.01)$, $q = (0, 0, 0)$, $N = \mathrm{diag}(2, 50, 20)$, where $\mathrm{diag}(x)$ represents a diagonal matrix with diagonal entries given by x.

The kite is initialized at an initial condition close to the ground, to show the behaviour of the controller for a large starting error and a long transient phase. The reel-out reference rate is set to be 0.5 m/s during traction and at -1 m/s during the reel-in phase. The reference path sizes are different during the two phases and we switch the reference paths when the phases are switched. The kite is set to be in the reel out phase initially. When the tether length exceeds 50 meters, we switch to retraction mode and reel-in till the tether length becomes less than 35 meters, at which point we switch back to the traction phase, completing a full pumping cycle.

Figure 6.5 shows the evolution of the states for the kite as it flies figures of eight during several pumping cycles. The tether length tracks the different reference slopes during the cycle and maintains tracking of the kite reference position. The minimum elevation angle ϑ decreases as the tether length increase and vice versa. This occurs because we have demanded a reference path with a constant minimum height characteristic which is desirable to higher power generation.

Figures 6.2,6.3,6.4 show path tracking for different reference paths in the traction and retraction phase. In the paths tested the controller shows fast convergence of the errors to a small bound with good tracking performance.

The virtual control inputs given by the controller $u = (\dot{z}, \dot{\gamma}, \dot{\tau})$ are shown in Fig.6.6. None of the virtual controllers become saturated at any time as we had allowed for a maximum of amplitude of 2,50,20 for $\dot{z}, \dot{\gamma}, \dot{\tau}$, respectively. Thus the internal states of the controller Θ also remain bounded (Fig.6.7).

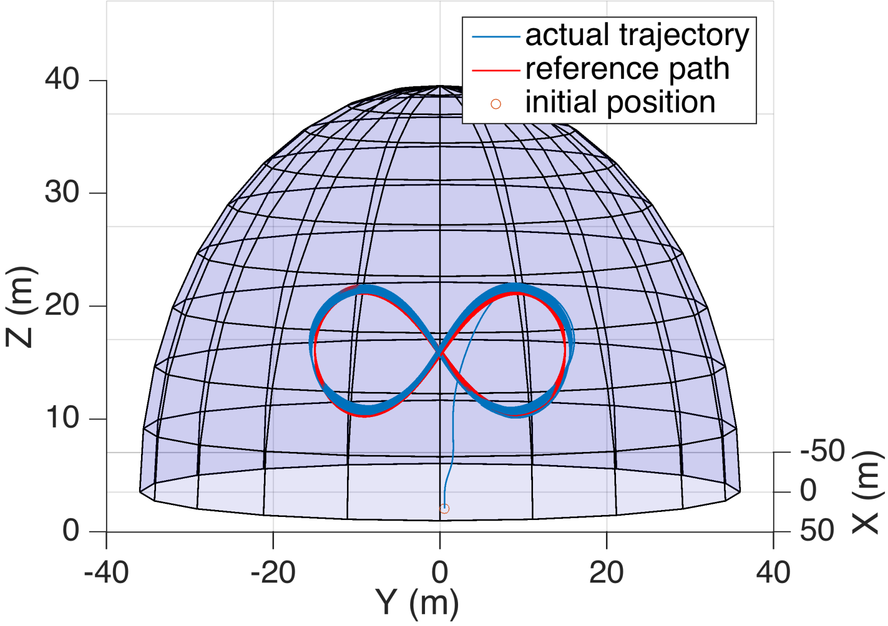


Figure 6.2: Path following of lemniscate figures in traction phase

The estimates for the parameters show bounded values as well. Note that the controller only guaranteed closed loop stability of the system and does not require or guarantee the convergence of the estimates to their true values. This is seen in Fig. 6.8

6.5 Conclusion

An adaptive path following controller for kite systems with parameter mismatch and with unknown wind velocity vector was presented. Under mild assumptions, the controller steers kite to a tube centered around a predefined geometric path. The tube diameter is determined by the choice of the design parameters of the controller. The effectiveness of the proposed strategy is demonstrated via numerical results on multiple geometric desired paths and pumping cycle flights.

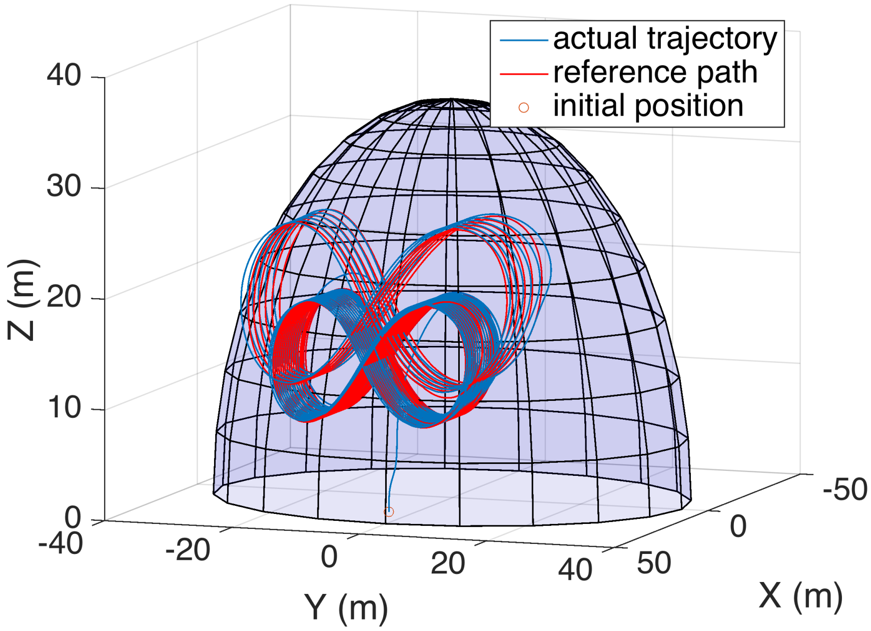


Figure 6.3: Path following of lemniscate in full pumping cycle

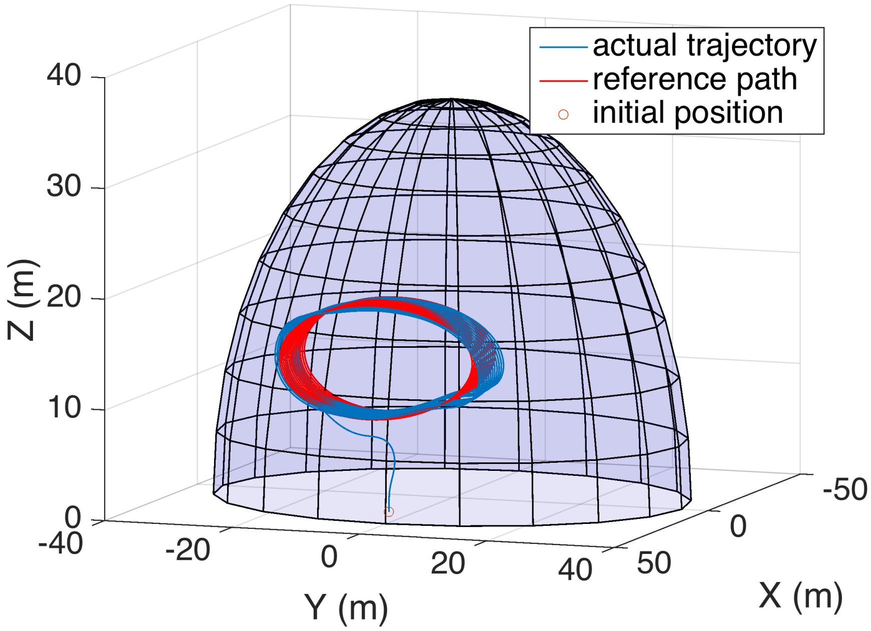


Figure 6.4: Path following of ellipsoidal orbits in traction phase

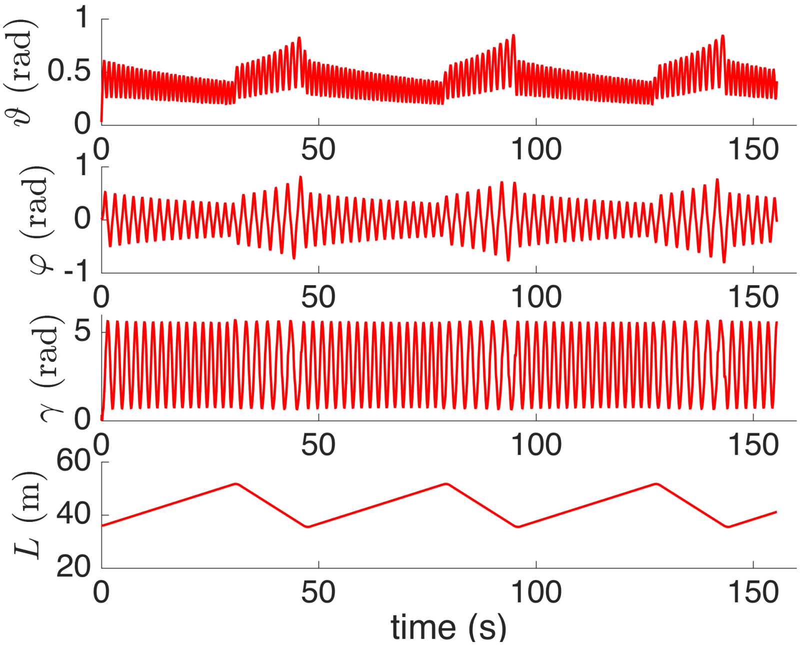


Figure 6.5: Kite state evolution through the pumping cycle

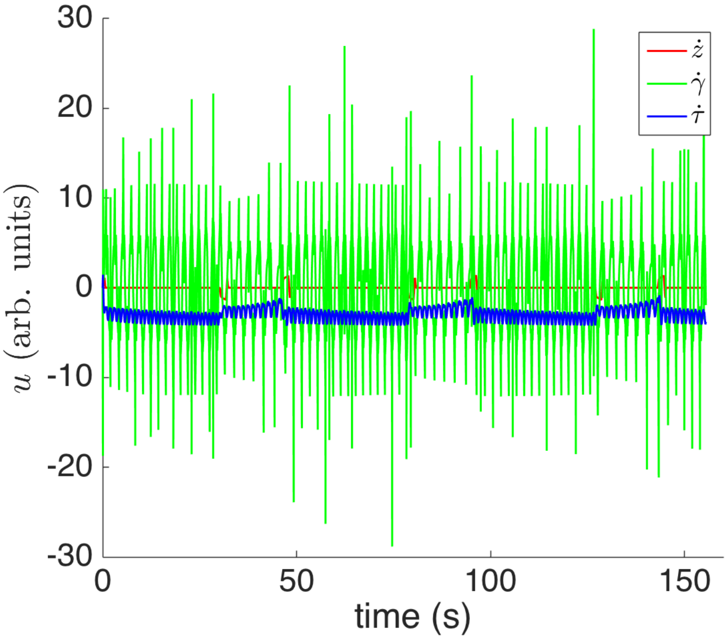


Figure 6.6: Control inputs through the pumping cycle

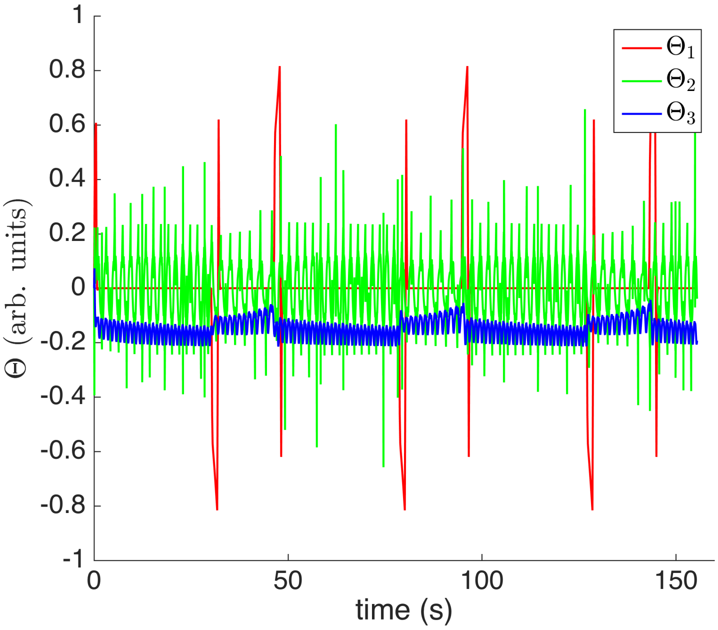


Figure 6.7: Internal states of the controller Θ

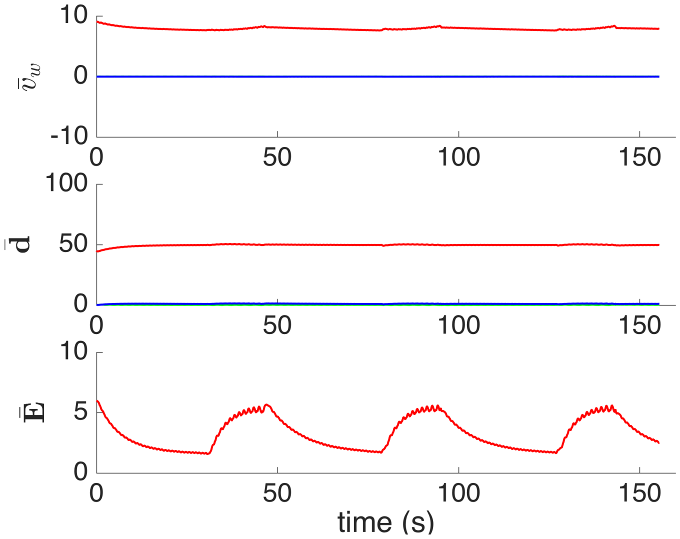


Figure 6.8: Parameter estimates $\bar{\lambda} = (\bar{v}_w, \bar{\mathbf{d}}, \bar{\mathbf{E}})$. Vector components of $\bar{v}_w, \bar{\mathbf{d}}$ in red, blue and green colors.

6.6 Appendix

The figure of eight trajectory as tracked in figures 6.2,6.3 has the following projection on the $y - z$ plane,

$$Y_{ref} = \frac{a \cos \tau}{1 + \sin^2 \tau} \qquad Z_{ref} = h + \frac{a \sin \tau \cos \tau}{1 + \sin^2 \tau} \tag{6.38}$$

with a being the width of the lemniscate and h the height of the center of the lemniscate. We use the values for $(a, h) = (15, 15)$ during the reel out phase and use a larger figure with $(a, h) = (20, 20)$ during the reel in phase to have a different retraction path.

The ellipsoidal trajectory as tracked in 6.4 has the following projection on the $y - z$ plane,

$$Y_{ref} = a \cos \tau \qquad Z_{ref} = h + \frac{a}{e} \sin \tau \tag{6.39}$$

with a being the width of the major axis for the ellipse, e being its eccentricity and h the height of its center.

For the reference reel-out rate $z_{ref}(t)$ we use a constant positive reel out rate $c_o = 0.5$ and a constant reel-in rate $c_o = -1$. The reference length $L_{ref}(t)$ can then be written as, $L_{ref}(t) = L_{ref}(0) + c_o \cdot t$ where $L_{ref}(0)$ is the initial tether length of the kite.

$$R(\beta) = \begin{pmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \tilde{S} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{6.40}$$

$$T = || \begin{pmatrix} \sin\vartheta_{ref}\varphi_{ref}\partial_\tau\vartheta_{ref} - \cos\vartheta_{ref}\partial_\tau\varphi_{ref} \\ \partial_\tau\vartheta_{ref} \end{pmatrix} || \tag{6.41}$$

$$\zeta = \angle \begin{pmatrix} \sin\vartheta_{ref}\varphi_{ref}\partial_\tau\vartheta_{ref} - \cos\vartheta_{ref}\partial_\tau\varphi_{ref} \\ \partial_\tau\vartheta_{ref} \end{pmatrix} \tag{6.42}$$

$$
\begin{aligned}
\dot{\mathbf{e}} &= R_{NK}^T \begin{pmatrix} 0 & & -LTsin(\zeta) \\ 0 & R_{NK}\tilde{S}\epsilon & LT\cos(\zeta) \\ 1 & & 0 \end{pmatrix} \begin{pmatrix} \dot{z} \\ \dot{\gamma} \\ \dot{\tau} \end{pmatrix} \\
& \quad + \dot{\gamma}\tilde{S}\mathbf{e} + R(\beta)\tilde{S}\dot{\beta}\epsilon_o \\
& \quad - R_{NK}^T \begin{pmatrix} L & 0 & 0 \\ -L\sin\vartheta_{ref}\varphi_{ref} & L\cos\vartheta_{ref} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \partial_t\vartheta_{ref} \\ \partial_t\varphi_{ref} \\ \ddot{L}_{ref} \end{pmatrix} \\
& \quad + R_{NK}^T \begin{pmatrix} \vartheta - \vartheta_{ref} - \mathbf{E}cos\gamma \\ \cos\vartheta\varphi - \cos\vartheta_{ref}\varphi_{ref} + \mathbf{E}\sin\vartheta\varphi\cos\gamma - \mathbf{E}\sin\gamma \\ 0 \end{pmatrix} z \\
& \quad + R_{NK}^T \begin{pmatrix} 1 & 0 & 0 \\ -sin\vartheta\varphi & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} R_{NK} \begin{pmatrix} 1 & 0 & -\mathbf{E} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} R_{NK}^T R_{GN}^T v_w \\
&= \dot{\gamma}\tilde{S}\mathbf{e} + R(\beta)\tilde{S}\dot{\beta}\epsilon_o + f + f_\lambda\lambda + gu
\end{aligned} \tag{6.43}
$$

$$
det(g) = LT\epsilon_{o_1} \tag{6.44}
$$

Part IV

Conclusion

Chapter 7

Contributions of the thesis

Chapter 2 presents a generalized representer theorem for variational problems with loss functionals and regularizers defined over arbitrary separable Hilbert spaces. The relation between adjoints for closed, densely defined operators and kernel methods in RKHS is highlighted, and simplified assumptions on subspace valued maps are presented to allow for non-r-regular maps. This extends previous work on the topic from [6, 8, 9] to allow for unbounded linear operators, like the differential operator to be considered in the loss functionals of the variational problem. Furthermore, considering loss functionals over arbitrary Hilbert spaces allows the consideration of loss functionals over the space of square integrable random vectors and stochastic processes as done in Section 2.4.2, allowing for representer theorems for such problems. The non-r-regular subspace valued maps allow for examples of ℓ_1-regularization to be considered over infinite dimensional spaces, and an example for this is presented in Section 2.4.3. Section 2.4.1 shows the application of the generalized representer theorem to deep neural networks for end-to-end learning to provide an explicit infinite dimensional linear representer for the neural network solution. The linear representer for the neural network is shown to be determined by finitely many signed vector measures and thus transforms the problem from optimization over a space of functions to optimization over a space of signed measures. For the case of optimization in the space of RKHS functions given by smooth kernels, a gradient-based method is used to solve for a locally optimal solution.

Chapter 3 presents a manifold learning approach to obstacle avoidance in autonomous driving. A variational problem for learning manifolds with star-shaped interiors is presented over a Hilbert space of periodic functions. The variational problem can be solved using both, kernel methods and with finite basis approximations with the techniques presented in Chapter 2. A collection of learned star-shaped sets are used to represent the free space in which a vehicle can move, and a corresponding inequality constraint is included in optimal control problems for obstacle avoidance in autonomous driving. The formulation with star-shaped sets allows for a reduction in the number of constraints required from $O(N)$ constraints [34, 38, 39, 40, 41, 42, 43, 44, 45, 46, 35, 36, 47] to $O(1)$ number of constraints required in the presence of N obstacles. This constant complexity helps to deal effectively with a large number and dynamically changing number of obstacles, which otherwise would have required an online reformulation of the optimization problem. Having a fixed structure for the optimization problem irrespective of the number of obstacles in the environment allows for a simpler implementation that is suitable for running on embedded systems. The use of the manifold constraints for obstacle avoidance is presented at different levels of planning and control by including it in a dynamic programming approach for corridor planning (Section 3.3.1), an N-phase free-end-time optimal control problem for optimal trajectory planning (Section 3.3.2) and a real-time obstacle avoidance and path following model predictive control scheme (MPC) in Section 3.3.3. The convergence and recursive feasibility of the MPC scheme is shown under mild assumptions on obstacle behavior and vehicle dynamics.

Chapter 4 presents a variational formulation for a continuous time, finite horizon model predictive controller for path following in an Airborne Wind Energy system. A convergent vector field control inspired terminal constraint is introduced to guarantee recursive feasibility and convergence of the path following error to zero. As the exogenous wind condition determines the flight speed for the vehicle, the vehicle speed is uncontrollable. As a result, a virtual vehicle is introduced on the desired geometric path with controlled speed. The virtual and real vehicles are then driven into consensus by the model predictive controller to achieve path following of the given geometric path. By using a vector field based terminal constraint, the region of attraction for the MPC scheme is expanded to allow for control with short prediction horizons, which makes the numerical implementation for the MPC scheme amenable for real-time application. A multiple shooting based implementation is provided for a real-time path following MPC controller.

Chapters 5 and 6 address problems pertaining to uncertainty in the wind and aerodynamic characteristics for optimization and control in airborne wind energy systems. Chapter 5 presents a method for data-driven performance optimization in the airborne wind energy system using Gaussian process surrogate models for the system objective, constraints and dynamics. The closed-loop performance is optimized by experimental sampling of candidate controller set-points, selected using the surrogate models. An information gain and bias trade-off metric called the expected improvement (Section 5.3.2) is used to search the candidate optimization points in a data efficient manner. Additional sampling constraints are introduced in the optimization problem to allow transient measurements to be used to update the model and candidate set points at a faster rate (Section 5.4). The performance under varying wind condition is shown to improve under the faster feedback provided by the Gaussian process optimization scheme with transient measurements, under constant and varying wind conditions (Section 5.5).

Chapter 6 presents a nonlinear direct adaptive control scheme for path following in AWE systems given parametric uncertainties of wind velocity and aerodynamic coefficients in the system dynamics. The path following error system for the AWE is shown to be affine in the control inputs and uncertain parameters. The system is not feedback linearizable, but under bounded control inputs, is shown to be approximately feedback linearizable. The online parameter update laws and an approximate feedback linearizing controller is shown to drive the path following error to a small neighborhood of zero. The inputs are bounded using a sigmoidal bounding transform for the inputs. The parameter estimates are shown to remain in a bounded neighborhood around some prior estimates provided to the controller.

The thesis thus uses Chapter 2, 3 and 4 to demonstrate the use of variational problems in learning and control and presents two different techniques for adaptive optimization and control of dynamical systems under uncertainty in Chapters 5 and 6 respectively. Chapter 8, points to some open questions and future directions for the work presented.

Chapter 8

Open research questions and future directions

It was shown in Chapter 2 that the differential operator is a closable, densely defined operator on a Sobolev space with a closed, densely defined adjoint operator (Example 2.3). Using the generalized representer theorem for the solving learning and control problems involving constraints given as ordinary and partial differential equations is thus made possible over an RKHS space embedded in such Sobolev spaces. The properties of kernels for such RKHS embedding in Sobolev spaces need further exploration and along with its relation to non-positive kernels as studied in [93, 94]. Further, a study of properties of the differential and integral operators over the Hilbert space of square integrable stochastic processes can help develop numerical collocation methods based on the representer theorem for numerical stochastic optimal control and model predictive control approaches. Working with the definition for adjoints over Banach spaces it may further be possible to extend the generalized representer theorem to Banach spaces and reproducing kernel Banach spaces [95].

Chapter 3 presented the use of manifold learning for obstacle avoidance in autonomous driving. The algorithm presented generalizes to $\mathbb{R}^n$ in a straightforward manner, and its application to obstacle avoidance and robust optimal control can be further considered. The multiphase nature of the optimal control formulation considered for trajectory generation and path following in Sections 3.3.2 and 3.3.3 leads to increased numerical complexity for the problem, and alternative formulations for the optimal control problem should be considered for faster numerical implementation.

The convergent vector field-terminal constraints in Chapter 4 are explicitly presented for the AWE dynamics. The idea can be further generalized to general dynamical systems for which a sliding mode or vector field controller can be designed. The issue of characterizing some systems and a general design method for the terminal constraints can be further looked into.

Finally, the Gaussian process optimization scheme presented in Chapter 5, with transient measurements being used can be combined with recent safe Gaussian process optimization techniques from [79, 96, 97].

Part V

Appendix

Appendix A

Some notes on functional analysis

A.1 Topology

The notion of topology is fundamental to defining the notion of continuity and convergence over spaces. As such continuity makes sense only with respect to a given topology and has no meaning if a topology is not specified (topologies are often left unspecified when they are understood to be a standard underlying topology for the given space). Convergence on the other hand can have two variants: (i) convergence in terms of a topology and (ii) convergence without reference to any topology. In fact a topology independent notion of convergence can then be used to establish a topology (see for example [98]).

Definition A.1. *(Topology)*
For an arbitrary set $\mathcal{Z}$, let Θ be a collection of subsets of $\mathcal{Z}$ such that

1. *Both $\varnothing$ and $\mathcal{Z}$ belong to Θ*

2. *Any union of elements in Θ belongs to Θ*

3. *Any intersection of finitely many elements in Θ belongs to Θ*

*Then the tuple $(\mathcal{Z}, \Theta)$ is called a **topological space** with **topology** Θ. The elements of Θ are called **open sets**.*

A comprehensive reference for topological spaces can be found in [99].

A.2 A note on dual norm

Let $\mathcal{Z}$ be a Banach space and $\mathcal{Z}^\star$ be the topological dual, i.e., the space of continuous linear functionals on $\mathcal{Z}$.

Definition A.2. *(Dual norm)*
The norm in the dual space is defined as $||h||_{\mathcal{Z}^\star} := \sup\{|h(x)| : ||x||_{\mathcal{Z}} \leq 1\}$.

The above definition for the dual norm is shown to provide the continuity bound for any linear functional $h \in \mathcal{Z}^\star$ in Lemma A.1 below.

Lemma A.1. $\forall x \in \mathcal{Z}$, $h \in \mathcal{Z}^\star$, $|h(x)| \leq ||h||_{\mathcal{Z}^\star}||x||_{\mathcal{Z}}$

Proof: *Note that for $||x||_{\mathcal{Z}} = 0$, the inequality is trivially satisfied since $||x||_{\mathcal{Z}} = 0$ implies $x = 0$ in a Banach space which in turn implies $h(x) = 0$ (by linearity of h).*

For $||x||_{\mathcal{Z}} \neq 0$, note that $||h||_{\mathcal{Z}^\star} \geq |h(x/||x||_{\mathcal{Z}})|$ (since $x/||x||_{\mathcal{Z}}$ is a norm 1 vector and $||h||_{\mathcal{Z}^\star}$ by definition is the supremum of a set containing $|h(x/||x||_{\mathcal{Z}})|$). By linearity of h, $|h(x/||x||_{\mathcal{Z}})| = |h(x)|/||x||_{\mathcal{Z}}$ and thus $||h||_{\mathcal{Z}^\star} \geq |h(x)|/||x||_{\mathcal{Z}} \implies |h(x)| \leq ||h||_{\mathcal{Z}^\star}||x||_{\mathcal{Z}}$. □

A.3 A note on reflexive Banach spaces

A more comprehensive note on reflexive Banach spaces can be found in the corresponding section of [17] or [22, Chapter 3, Section 11].

Let $\mathcal{Z}$ be a Banach space and $\mathcal{Z}^\star$ denote its dual Banach space. Since $\mathcal{Z}^\star$ is also a Banach space, one can find its dual Banach space $\mathcal{Z}^{\star\star}$ and so on. This naturally leads to the question of what such a sequence of duals looks like. Does the dualization terminate at some point by returning to $\mathcal{Z}$. In particular, is $\mathcal{Z}^{\star\star}$ isomorphic to $\mathcal{Z}$? Or does this lead to an infinite sequence of dual Banach spaces? Reflexivity ($\mathcal{Z}^{\star\star} \overset{iso}{=} \mathcal{Z}$) addresses this question.

Note at the outset that not all Banach spaces are reflexive, i.e., $\mathcal{Z}^{\star\star} \overset{iso}{\neq} \mathcal{Z}$ and thus this question doesn't resolve to a trivial yes or no answer. The answer is a trivial yes (i.e. $\mathcal{Z}$ is reflexive) when $\mathcal{Z}$ is a finite dimensional Banach space or when $\mathcal{Z}$ is a Hilbert space. In general, however, $\mathcal{Z} \overset{iso}{\subseteq} \mathcal{Z}^{\star\star}$; i.e. every $x \in \mathcal{Z}$ corresponds to a $x^{\star\star} \in \mathcal{Z}^{\star\star}$ with a natural inclusion map $i : \mathcal{Z} \to \mathcal{Z}^{\star\star}$ such that $i(x) = (x^{\star\star} : \mathcal{Z}^\star \to \mathbb{R})$ given by $x^{\star\star}(h) = h(x)$ for all $h \in \mathcal{Z}^\star$.

Using [22, Corollary 6.7] however one can show that not only is $i : \mathcal{Z} \to \mathcal{Z}^{\star\star}$ a natural isomorphic inclusion but is actually an isometric inclusion, i.e., for all $x \in \mathcal{Z}$ and $x^{\star\star} = i(x)$, $||x^{\star\star}||_{\mathcal{Z}^{\star\star}} = ||x||_{\mathcal{Z}}$ (irrespective of whether $\mathcal{Z}$ is reflexive or not). The proof for this isometry is rewritten here for reference. Note that the proof in the case of a general Banach space is fairly deep with arguments leading back to the Hahn-Banach theorem, we thus break the proof down into smaller lemmas (which actually are corollaries of the Hahn-Banach theorem) and present the proof top-down, i.e. present arguments for the theorem first invoking the lemma and then proving the lemma required.

Theorem A.1. $||x^{\star\star}||_{\mathcal{Z}^{\star\star}} = ||x||_{\mathcal{Z}}$

Proof: *Recall that $||h||_{\mathcal{Z}^\star} := \sup\{|h(x)| : ||x||_{\mathcal{Z}} \leq 1\}$ and $||x^{\star\star}||_{\mathcal{Z}^{\star\star}} := \sup\{|x^{\star\star}(h)| : ||h||_{\mathcal{Z}^\star} \leq 1\} = \sup\{|h(x)| : ||h||_{\mathcal{Z}^\star} \leq 1\}$. By Lemma A.1, $|h(x)| \leq ||h||_{\mathcal{Z}^\star}||x||_{\mathcal{Z}}$ and thus $||x^{\star\star}||_{\mathcal{Z}^{\star\star}} = \sup\{|h(x)| : ||h||_{\mathcal{Z}^\star} \leq 1\} \leq \sup\{||h||_{\mathcal{Z}^\star}||x||_{\mathcal{Z}} : ||h||_{\mathcal{Z}^\star} \leq 1\} = ||x||_{\mathcal{Z}}$, i.e. $||x^{\star\star}||_{\mathcal{Z}^{\star\star}} \leq ||x||_{\mathcal{Z}}$. By Lemma A.2 below, for every $x \in \mathcal{Z}$ there exists a $h_x \in \mathcal{Z}^\star$ such that $h_x(x) = ||x||_{\mathcal{Z}}$ and $||h_x||_{\mathcal{Z}^\star} = 1$. Thus $||x^{\star\star}||_{\mathcal{Z}^{\star\star}} = \sup\{|h(x)| : ||h||_{\mathcal{Z}^\star} \leq 1\} = |h_x(x)| = ||x||_{\mathcal{Z}}$ (the supremum being attained for $h_x \in \mathcal{Z}^\star$).* □

Note that existence of such an $h_x \in \mathcal{Z}^\star$ is somewhat simpler to see when $\mathcal{Z}$ is an Hilbert space since we can define $h_x(y) = \langle x/||x||_{\mathcal{Z}}, y\rangle$ as the functional (when $||x||_{\mathcal{Z}} \neq 0$) and note that $h_x(x) = ||x||_{\mathcal{Z}}$ and also $|h_x(y)|$ is maximum with $y = x/||x||$ (for $||y||_{\mathcal{Z}} \leq 1$) implying $||h_x||_{\mathcal{Z}^\star} = 1$. For $||x||_{\mathcal{Z}} = 0$ note that any linear functional with norm 1 ($||h||_{\mathcal{Z}^\star} = 1$) trivially satisfies this requirement $h_x(x) =$

$||x||_{\mathcal{Z}}$ (since $h_x(x) = ||x||_{\mathcal{Z}} = 0$). This result is formalized by Lemma A.2 for a general Banach space where the result may not be as simple to see.

Lemma A.2. *($\exists\, h_x \in \mathcal{Z}^\star$: $h_x(x) = ||x||_{\mathcal{Z}}$, $||h_x||_{\mathcal{Z}^\star} = 1$)*
For every $x \in \mathcal{Z}$, there exists a $h_x \in \mathcal{Z}^\star$ such that $h_x(x) = ||x||_{\mathcal{Z}}$ and $||h_x||_{\mathcal{Z}^\star} = 1$.

Proof: *Let $\mathcal{M} = \{\lambda x : \lambda \in \mathbb{R}\}$ be the subspace of $\mathcal{Z}$ spanned by x. Let $g : \mathcal{M} \to \mathbb{R}$ denote the linear functional in $\mathcal{M}^\star$ given by $g(\lambda x) = \lambda ||x||_{\mathcal{Z}}$. Note that $||g||_{\mathcal{M}^\star} = \sup\{|g(\lambda x)| : |\lambda|||x||_{\mathcal{Z}} \leq 1\} = \sup\{|\lambda|||x||_{\mathcal{Z}} : |\lambda|||x||_{\mathcal{Z}} \leq 1\} = 1$. By Lemma A.3 it is known that there will exist a $h_x \in \mathcal{Z}^\star$ such that h_x restricted to $\mathcal{M}$, denoted $h_x|\mathcal{M}$, is such that $h_x|\mathcal{M} = g$ and $||h_x||_{\mathcal{Z}^\star} = ||g||_{\mathcal{M}^\star}$. Then for all $x \in \mathcal{M}$, we have $h_x|\mathcal{M}(x) = g(x) = ||x||_{\mathcal{Z}}$ and $||h_x||_{\mathcal{Z}^\star} = ||g||_{\mathcal{M}^\star} = 1$. Thus there exists a $h_x \in \mathcal{Z}^\star$ such that $h_x(x) = ||x||_{\mathcal{Z}}$ and $||h||^\star_{\mathcal{Z}} = 1$.* □

Lemma A.3. *Let $p : \mathcal{Z} \to [0, \infty)$ be a semi-norm on $\mathcal{Z}$. Let $\mathcal{M} \subseteq \mathcal{Z}$ be a linear subspace of $\mathcal{Z}$ and $g \in \mathcal{M}^\star$ be a bounded linear functional on $\mathcal{M}$, then*

1. *$\forall x \in \mathcal{M}^\star$, $|g(x)| \leq p(x) \implies$ there exists a $h \in \mathcal{Z}^\star$ such that $h|\mathcal{M} = g$ and $\forall x \in \mathcal{Z}$, $|h(x)| \leq p(x)$.*

2. *there exists a $h \in \mathcal{Z}^\star$ such that $h|\mathcal{M} = g$ and $||h||_{\mathcal{Z}^\star} = ||g||_{\mathcal{M}^\star}$.*

Proof: *For the first statement, note that $|g(x)| \leq p(x)$ implies $g(x) \leq p(x)$ for all $x \in \mathcal{M}$. Then by the Hahn Banach theorem there exists an extension $h \in \mathcal{Z}^\star$ such that $h(x) \leq p(x)$ for all $x \in \mathcal{Z}$ and $h|\mathcal{M} = g$. Since p is a semi-norm we have $h(-x) \leq p(-x) = p(x)$ and $h(-x) = -h(x)$ (by linearity of h). Thus both $h(x) \leq p(x)$ and $-h(x) \leq p(x)$ for all $x \in \mathcal{Z}$ and thus $|h(x)| \leq p(x)$.*

For the second statement, note that $p : \mathcal{Z} \to [0, \infty)$ defined as $p(x) = ||g||_{\mathcal{M}^\star}||x||_{\mathcal{Z}}$ is a semi-norm (rather a norm) on $\mathcal{Z}$. Then from the first statement there exists a $h \in \mathcal{Z}^\star$ such that $h|\mathcal{M} = g$ and for all $x \in \mathcal{Z}$, $|h(x)| \leq ||g||_{\mathcal{M}^\star}||x||_{\mathcal{Z}}$. Then $||h||_{\mathcal{Z}^\star} = \sup\{|h(x)| : ||x||_{\mathcal{Z}} \leq 1\} = ||g||_{\mathcal{M}^\star}$. □

Theorem A.1 essentially establishes the isometry of $\mathcal{Z}^{\star\star}$ to $\mathcal{Z}$ even when $\mathcal{Z}$ may not be reflexive. It does not however address the question of what the necessary and sufficient conditions are for $\mathcal{Z}$ to be reflexive. Theorem A.2 (restatement of [22, Theorem 4.2]) states this necessary and sufficient condition.

Theorem A.2. *(Conditions for reflexivity)*
Let $\sigma(\cdot,\cdot)$ denote the weak topology. A Banach space $\mathcal{Z}$ is reflexive if and only if*

- *$\mathcal{Z}^\star$ is reflexive, or equivalently*

- *$\sigma(\mathcal{Z}^\star, \mathcal{Z}) = \sigma(\mathcal{Z}^\star, \mathcal{Z}^{\star\star})$*

A.4 Adjoints for bounded operators

The following discussion follows along the corresponding sections in [17].

Definition A.3. *(Adjoint on Banach spaces)*
Let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ be two Banach spaces. Let $\mathcal{Z}_1^\star$ and $\mathcal{Z}_2^\star$ be the corresponding dual spaces. Then for a bounded (continuous) linear operator $L : \mathcal{Z}_1 \to \mathcal{Z}_2$ the adjoint is defined as the bounded (continuous) linear operator, $L^\star : \mathcal{Z}_2^\star \to \mathcal{Z}_1^\star$ that satisfies

$$\forall x \in \mathcal{Z}_1, h \in \mathcal{Z}_2^\star, \quad h(Lx) = (L^\star h)(x) \tag{A.1}$$

On Hilbert spaces the simplified definition for the adjoint can be written as follows

Definition A.4. *(Adjoint on Hilbert space)*
Let $\mathcal{Z}_1$, $\mathcal{Z}_2$ be Hilbert spaces. Let $L : \mathcal{Z}_1 \to \mathcal{Z}_2$ be a bounded linear operator then the adjoint is the bounded linear operator $L^\star : \mathcal{Z}_2 \to \mathcal{Z}_1$ satisfying,

$$\forall x \in \mathcal{Z}_1, y \in \mathcal{Z}_2, \quad \langle Lx, y\rangle_{\mathcal{Z}_2} = \langle x, L^\star y\rangle_{\mathcal{Z}_1} \tag{A.2}$$

The existence and uniqueness of the adjoint operator for bounded linear operators in Hilbert spaces and Banach spaces is shown in [17, Theorem 5.4.2] and [17, Proposition 9.1.3] respectively. It is also shown that $L^{\star\star}|\mathcal{Z}_1 = L$.

A.5 Adjoints for unbounded, densely defined operators

Definition A.5. *(Dense subset)*
In a topological space, $(\mathcal{Z}, \Theta_{\mathcal{Z}})$ a subset $\mathcal{D}$ is said to be dense in $\mathcal{Z}$ if for all $z \in \mathcal{Z}$ and all open neighborhoods of z, $\mathcal{N}(z) \in \Theta_{\mathcal{Z}}$, there exists a $d \in \mathcal{D}$ such that $d \in \mathcal{N}(z)$. In a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ this is equivalent to saying that each $z \in \mathcal{Z}$, there exists a sequence $d_n \in \mathcal{D}$ such that $\lim_{n\to\infty} d_{\mathcal{Z}}(d_n, z) = 0$ (i.e. d_n approaches z arbitrarily close in distance).

Remark A.1. *In a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, if $\mathcal{D}$ is a dense subset of $\mathcal{Z}$ then the closure of $\mathcal{D}$ with respect to $d_{\mathcal{Z}}$, denoted $\overline{\mathcal{D}}$, is equal to $\mathcal{Z}$, i.e., $\overline{\mathcal{D}} = \mathcal{Z}$.*

Definition A.6. *(Densely defined operator)*
Let $\mathcal{Z}_1$, $\mathcal{Z}_2$ be topological spaces and let $\mathcal{D}$ be a dense subset of $\mathcal{Z}_1$. Then a linear operator $L : \mathcal{D} \to \mathcal{Z}_2$ is said to be densely defined on $\mathcal{Z}_1$ and denoted $L : \mathcal{Z}_1 \to \mathcal{Z}_2$.

Remark A.2. *We denote the domain of an operator $L : \mathcal{Z}_1 \to \mathcal{Z}_2$ as dom(L). In the definition above, dom$(L) = \mathcal{D}$.*

Definition A.7. *(Adjoint of densely defined operators in Hilbert spaces)*
Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a densely defined operator. Let dom$(A^\star) := \{k \in \mathcal{H}_2 : f : \text{dom}(A) \to \mathbb{R} : f(h) = \langle Ah, k\rangle_{\mathcal{H}_2}$is bounded linear functional on dom$(A)\}$. Then for all $h \in$ dom(A) and $k \in$ dom$(A^\star)$ there exists a unique $f \in \mathcal{H}_1$ such that $\langle Ah, k\rangle_{\mathcal{H}_2} = \langle f, h\rangle_{\mathcal{H}_1}$ (by Riesz representer theorem). The adjoint is defined as the operator $A^\star : \text{dom}(A^\star) \subseteq \mathcal{H}_2 \to \mathcal{H}_1$.

By [22, Chapter 10, Proposition 1.6], if the operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ is closable and densely defined then the adjoint $L^\star$ is also densely defined, i.e., dom$(A^\star)$ is a dense subset of $\mathcal{H}_2$.

Definition A.8. *(Extension of an operator)*
Let A, B be operators from sets $\mathcal{Z}_1$ to $\mathcal{Z}_2$. Then A is said to be an extension of B and denoted $B \subseteq A$, if dom$(B) \subseteq$ dom(A) and for all $h \in$ dom(B), $Ah = Bh$.

Definition A.9. *(Closed and closable operator)*
An operator $A : \mathcal{Z}_1 \rightarrow \mathcal{Z}_2$ is called closed if its graph $\tau(A) := \{(h, Ah) \in \mathcal{Z}_1 \times \mathcal{Z}_2 : h \in \text{dom}(A)\}$ is a closed set in the topology of $\mathcal{Z}_1 \times \mathcal{Z}_2$. It is called closable if there exists an extension with domain being $\mathcal{Z}_1$ that is closed.

Definition A.10. *(Continuous extension of closable operator)*
Let $\mathcal{Z}^\star$ be a dense subset of $\overline{\mathcal{Z}^\star}$. For a closable operator $J : \mathcal{Z}^\star \rightarrow \mathcal{Z}$ there exists a closed extension to $\overline{\mathcal{Z}^\star}$ such that for any $h \in \overline{\mathcal{Z}^\star}$ and a Cauchy sequence $h_n \in \mathcal{Z}^\star$ such that $\lim_{n\rightarrow\infty} ||h_n - h||_{\overline{\mathcal{Z}^\star}} = 0$, $Jh = \lim_{n\rightarrow\infty} Jh_n$.

A.6 Self Adjoint operators

Using the general definition of the adjoint in Banach spaces, a self adjoint operator in $\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}$ can be defined as follows,

Definition A.11. *(Self adjoint bounded operator on Banach spaces)*
Let $J : \mathcal{Z}^\star \rightarrow \mathcal{Z}$ be a bounded linear operator and $J^\star : \mathcal{Z}^\star \rightarrow \mathcal{Z}^{\star\star}$ denote the corresponding adjoint. J is said to be self adjoint if $J^\star(\mathcal{Z}^\star) \overset{iso}{\subseteq} \mathcal{Z}$ and $\forall y \in \mathcal{Z}^\star$, $J^\star(y) \overset{iso}{=} J(y)$, i.e., $J^\star = J$.

On Hilbert spaces the above definition simplifies to the following,

Definition A.12. *(Self adjoint bounded operator on Hilbert spaces)*
Let $J : \mathcal{H} \rightarrow \mathcal{H}$ be a bounded linear operator and $J^\star : \mathcal{H} \rightarrow \mathcal{H}$ denote the corresponding adjoint. J is said to be self adjoint if $\forall y \in \mathcal{Z}^\star$, $J^\star(y)$=$J(y)$, i.e., $J^\star = J$.

Lemma A.4. *(Self adjoint densely defined operator on Banach space)*
Let $J : \mathcal{Z}^\star \rightarrow \mathcal{Z}$ be a bounded operator with dom$(J) = \mathcal{Z}^\star$. Let $\mathcal{Z}^\star$ be a dense subset in $\overline{\mathcal{Z}^\star}$, i.e., J is densely defined on $\overline{\mathcal{Z}^\star}$, then the adjoint $J^\star : \mathcal{Z}^\star \rightarrow \mathcal{Z}^{\star\star}$ is densely defined on $\overline{\mathcal{Z}^\star}$.

A.6.1 Hilbert space induced by self-adjoint, positive semi-definite operators

Definition A.13. *(Self adjoint operator)*
Let $J : \mathcal{Z}^\star \rightarrow \mathcal{Z}$ be a bounded linear operator and $J^\star : \mathcal{Z}^\star \rightarrow \mathcal{Z}^{\star\star}$ denote the corresponding adjoint. J is said to be self adjoint if $J^\star(\mathcal{Z}^\star) \overset{iso}{\subseteq} \mathcal{Z}$ and $\forall y \in \mathcal{Z}^\star$, $J^\star(y) \overset{iso}{=} J(y)$, i.e., $J^\star = J$.

Definition A.14. *(Positive (semi)definite operators)*
An operator $J : \mathcal{Z}^\star \rightarrow \mathcal{Z}$ is said to be positive semidefinite if $\forall h \in \mathcal{Z}^\star$, $h(Jh) \geq 0$. Further, the operator is called positive definite if $\forall h \in \mathcal{Z}^\star$, $h \neq 0$, $h(Jh) > 0$. We denote a positive semidefinite operator as $J \geq 0$ and a positive definite operator as $J > 0$.

Definition A.15. *(Partially ordered set of self adjoint, positive semidefinite operators)* *Let $\mathcal{S}^+(\mathcal{Z}) = \{J \in \mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}} : J = J^\star, J \geq 0\}$ be a collection of all positive semidefinite, self adjoint continuous linear operators from $\mathcal{Z}^\star$ to $\mathcal{Z}$. We say $J_1 \geq J_2$ when $J_1 - J_2 \geq 0$. For positive definite operators we use the notation $\mathcal{S}^{++}(\mathcal{Z}) = \{J \in \mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}} : J = J^\star, J > 0\}$.*

Theorem A.3. *(Hilbert space induced by self adjoint, positive semidefinite operators)* *Let $J \in \mathcal{S}^+(\mathcal{Z})$ be a self adjoint, positive semidefinite operator and let $\mathcal{H}_0 = J(\mathcal{Z}^\star)$. Let $J^{-1}\zeta$ denote the pre-image set $J^{-1}\zeta = \{h \in \mathcal{Z}^\star : Jh = \zeta\}$. Let $(J^{-1}\zeta)(\eta)$ denote the set $(J^{-1}\zeta)(\eta) = \{h(\eta) : h \in J^{-1}\zeta\}$ for any $\zeta, \eta \in \mathcal{Z}$. Then,*

1. *for all $\zeta, \eta \in \mathcal{H}_0$, $(J^{-1}\zeta)(\eta)$ is a non empty singleton set*
2. *$\langle\cdot,\cdot\rangle_{\mathcal{H}_0} : \mathcal{H}_0 \times \mathcal{H}_0 \to [0,\infty)$ defined as $\langle\zeta,\eta\rangle_{\mathcal{H}_0} = (J^{-1}\zeta)(\eta)$ is an inner product on $\mathcal{H}_0$*
3. *The completion of $\mathcal{H}_0$ under the inner product induced norm $||\cdot||_{\mathcal{H}_0}$ is a Hilbert space, denoted $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$.*

Proof: *Let $\mathcal{H}_0 = J(\mathcal{Z}^\star)$ be the range of J. Then for any $\zeta \in \mathcal{H}_0$ there exists a $h \in \mathcal{Z}^\star$ such that $\zeta = Jh$ and thus the pre-image set $J^{-1}\zeta$ is non-empty for any $\zeta \in \mathcal{H}_0$. Then for any $\zeta, \eta \in \mathcal{H}_0$ note that for all $h \in J^{-1}\zeta$ and $g \in J^{-1}\eta$, $h(\eta) = h(Jg) \overset{(1)}{=} g(Jh) = g(\zeta)$, the equality in $\overset{(1)}{=}$ following from the self adjoint property of J. Thus for any $h_1, h_2 \in J^{-1}\zeta$ and any $g \in J^{-1}\eta$, $h_1(\eta) = h_1(Jg) = g(Jh_1) = g(\zeta) = g(Jh_2) = h_2(Jg) = h_2(\eta)$ and $(J^{-1}\zeta)(\eta) = \{h(Jg) : h \in J^{-1}\zeta, g \in J^{-1}\eta\}$ is a singleton set.*

We can thus define a bilinear symmetric operation on $\mathcal{H}_0$, $\langle\cdot,\cdot\rangle_{\mathcal{H}_0} : \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{R}$ given by $\langle\zeta,\eta\rangle_{\mathcal{H}_0} = (J^{-1}\zeta)(\eta) = (J^{-1}\eta)(\zeta) = \langle\eta,\zeta\rangle_{\mathcal{H}_0}$. Further since J is positive semidefinite, then the symmetric bilinear operation is positive semidefinite, i.e., $\forall\zeta \in \mathcal{H}_0$, $\langle\zeta,\zeta\rangle_{\mathcal{H}_0} \geq 0$. To see this, note that $\langle\zeta,\zeta\rangle_{\mathcal{H}_0} = (J^{-1}\zeta)(\zeta)$ and for any $h \in J^{-1}\zeta$, $(J^{-1}\zeta)(\zeta) = h(Jh) \geq 0$ (by positive semidefiniteness of J). Also by linearity of h, $h(Jh) = 0$ if and only if $Jh = 0$, i.e., $\zeta = 0$. Thus $\langle\zeta,\zeta\rangle_{\mathcal{H}_0} = 0 \iff \zeta = 0$. Thus $\langle\cdot,\cdot\rangle_{\mathcal{H}_0} : \mathcal{H}_0 \times \mathcal{H}_0 \to [0,\infty)$ defines a valid inner product on $\mathcal{H}_0 \subseteq \mathcal{Z}$.

The completion under the inner product norm of $\mathcal{H}_0$ is the space $\mathcal{H}$ such that all Cauchy sequences in $\mathcal{H}_0$ converge in $\mathcal{H}$. For any $\eta, \zeta \in \mathcal{H}$, there thus exist Cauchy sequences η_n, ζ_n such that $\lim_{n\to\infty} ||\eta_n - \eta||_{\mathcal{H}_0} = 0$ and $\lim_{n\to\infty} ||\zeta_n - \zeta||_{\mathcal{H}_0} = 0$, i.e. $\mathcal{H}_0$ is dense in $\mathcal{H}$. The inner product $\langle\eta,\zeta\rangle_{\mathcal{H}} = \lim_{n\to\infty}\langle\eta_n,\zeta_n\rangle_{\mathcal{H}_0}$ is well defined since $\langle\eta_n,\zeta_n\rangle_{\mathcal{H}_0}$ forms a Cauchy sequence on $\mathbb{R}$. Further for any Cauchy sequence ζ_n in $\mathcal{H}$, there exists a Cauchy sequence ζ_n' in $\mathcal{H}_0$ such that $\lim_{n\to\infty} ||\zeta_n' - \zeta_n||_{\mathcal{H}} = \lim_{n\to\infty} ||\zeta_n' - \zeta_n||_{\mathcal{H}_0} = 0$. Since ζ_n' must converge in $\mathcal{H}$, so must ζ_n and thus $\mathcal{H}$ is complete under the norm $||\cdot||_{\mathcal{H}}$, thus forming a Hilbert space. □

Using the fact that $\mathcal{H}_0$ is dense in $\mathcal{H}$, we can treat the inner product and norm on $\mathcal{H}$ as being identical to the inner product and norm on $\mathcal{H}_0$. Thus henceforth we will not distinguish between the two spaces for the inner product and norm computations.

Definition A.16. *(Continuous embedding of a Hilbert space)* *A Hilbert space $\mathcal{H}$ is said to be continuously embedded in a Banach space $\mathcal{Z}$ if $\mathcal{H} \subseteq \mathcal{Z}$ and there exists a natural inclusion $i : \mathcal{H} \to \mathcal{Z}$ given by $i(x) = x$ for all $x \in \mathcal{H}$ and a constant $\kappa \in [0,\infty)$ such that for all $x \in \mathcal{H}$, $||x||_{\mathcal{Z}} \leq \kappa||x||_{\mathcal{H}}$.*

Theorem A.4. *(Continuous embedding of Hilbert space induced by $J \in \mathcal{S}^+(\mathcal{Z})$)*
The Hilbert space $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$ induced by a self adjoint, positive semidefinite operator $J \in \mathcal{S}^+(\mathcal{Z})$ is continuously embedded in $\mathcal{Z}$.

Proof: *By the application of Cauchy Schwartz inequality, $\forall \zeta, \eta \in \mathcal{H}$ and $\forall h \in J^{-1}\zeta$, $g \in J^{-1}\eta$, $|\langle \zeta, \eta \rangle_{\mathcal{H}}| \leq ||\zeta||_{\mathcal{H}} ||\eta||_{\mathcal{H}} = ||\zeta||_{\mathcal{H}} |g(Jg)|^{1/2} \leq ||\zeta||_{\mathcal{H}} ||g||^{1/2}_{\mathcal{Z}^\star} ||Jg||^{1/2}_{\mathcal{Z}} \leq ||\zeta||_{\mathcal{H}} ||J||^{1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||g||_{\mathcal{Z}^\star}$. Thus for all $\zeta \in \mathcal{H}$ and $g \in \mathcal{Z}^\star$, $|\langle \zeta, Jg \rangle_{\mathcal{H}}| \leq ||\zeta||_{\mathcal{H}} ||J||^{1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||g||_{\mathcal{Z}^\star}$*

Further (by Theorem A.1) $||\zeta||_{\mathcal{Z}} = ||\zeta||_{\mathcal{Z}^{\star\star}} = \sup_{\{g \in \mathcal{Z}^\star : ||g||_{\mathcal{Z}^\star}=1\}} |g(\zeta)| = \sup_{\{g \in \mathcal{Z}^\star : ||g||_{\mathcal{Z}^\star}=1\}} |g(Jh)| = \sup_{\{g \in \mathcal{Z}^\star : ||g||_{\mathcal{Z}^\star}=1\}} |h(Jg)| = \sup_{\{g \in \mathcal{Z}^\star : ||g||_{\mathcal{Z}^\star}=1\}} |\langle \zeta, Jg \rangle_{\mathcal{H}}| \leq ||\zeta||_{\mathcal{H}} ||J||^{1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}}$. Thus $||\zeta||_{\mathcal{Z}} \leq ||J||^{1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||\zeta||_{\mathcal{H}}$, implying $(\mathcal{H}, ||\cdot||_{\mathcal{H}})$ is continuously embedded in $(\mathcal{Z}, ||\cdot||_{\mathcal{Z}})$. □

Theorem A.5. *(Properties of J induced inner product)*
For $J \in \mathcal{S}^+(\mathcal{Z})$, the induced inner product $\langle \zeta, \eta \rangle_{\mathcal{H}} = J^{-1}\zeta(\eta)$ is such that,

1. *$\forall h, g \in \mathcal{Z}^\star$, $\langle Jh, Jg \rangle_{\mathcal{H}} = h(Jg) = g(Jh)$*
2. *$\forall h \in \mathcal{Z}^\star$, $\zeta \in \mathcal{H}$, $\langle Jh, \zeta \rangle_{\mathcal{H}} = h(\zeta)$*

Proof: *For the first statement, note that $\langle Jh, Jg \rangle_{\mathcal{H}} = (J^{-1}Jh)(Jg) = h(Jg)$ and by symmetry of inner product $\langle Jh, Jg \rangle_{\mathcal{H}} = \langle Jg, Jh \rangle_{\mathcal{H}} = g(Jh)$. For the second statement, note that $\langle Jh, \zeta \rangle_{\mathcal{H}} = (J^{-1}Jh)(\zeta) = h(\zeta)$. Recall that the well defined (uniqueness) nature of $J^{-1}Jh(\zeta)$ follows from Theorem A.3-1.* □

Theorem A.6. *Let $\mathcal{H}$ be a dense subspace in $\mathcal{Z}$, induced by a $J \in \mathcal{S}^+(\mathcal{Z})$, then $\forall h \in \mathcal{H}^\star$, there exist a unique bounded linear extension to $h' \in \mathcal{Z}^\star$ such that $||h||_{\mathcal{H}^\star} = ||h'||_{\mathcal{Z}^\star}$ and for all $\zeta \in \mathcal{H}$, $\langle Jh', \zeta \rangle_{\mathcal{H}} = h(\zeta)$*

Proof: *The existence of the unique, bounded linear extension $h' \in \mathcal{Z}^\star$ follows from a specialization of the Hahn-Banach theorem for dense linear subspaces as given by [**?**, Theorem 3]. Then for any $h \in \mathcal{H}^\star$ the unique bounded linear extension $h' \in \mathcal{Z}^\star$ is such that $h(f) = h'(f)$ for all $f \in \mathcal{H}$ and for all $\zeta \in \mathcal{Z}$ $|h'(\zeta)| \leq ||h||_{\mathcal{H}^\star} ||\zeta||_{\mathcal{Z}}$. From statement two of Theorem A.5, we know that for all $h' \in \mathcal{Z}^\star$ and $f \in \mathcal{H}$, we have $\langle Jh', f \rangle_{\mathcal{H}} = h'(f) = h(f)$.* □

Appendix B

Some notes on probability theory

Section B.1 presents the preliminary notions of sigma algebra, probability measure and a probability measure space with their conventional definitions. Section B.2 defines the notion of a $(\mathcal{F}|\mathcal{B})$-measurable function and Section B.3 reviews the notion of Lebesgue integration of measurable functions. Section B.4 defines the moments and expectation for measurable functions. Section B.5 and B.6 review the notions of a density function and the Radon-Nikodym theorem respectively. Finally, Section B.7 defines the notion of a Gaussian measure on separable Banach spaces.

B.1 Probability Measure Space

Definition B.1. *(σ-algebra)*
Let Ω be a set and let $\varnothing$ denote an empty set. A σ-algebra on Ω is a collection $\mathcal{F}$ of subsets of Ω satisfying

1. $\varnothing \in \mathcal{F}$
2. *If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$*
3. *If $\{A_n \in \mathcal{F} : \forall n = 1 \dots \infty\}$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$*

The following additional properties can then be derived from the above three axioms

1. Both $\varnothing$ and Ω belong to $\mathcal{F}$
2. For a countable sequence $\{A_n \in \mathcal{F} : n = 1 \dots \infty\}$, both $\cup_{n=1}^{\infty} A_n$ and $\cap_{n=1}^{\infty} A_n$ belong to $\mathcal{F}$
3. For any $A, B \in \mathcal{F}$ and $A \subset B$, $B \backslash A \in \mathcal{F}$

A tuple $(\Omega, \mathcal{F})$ of the set Ω and a σ-algebra $\mathcal{F}$ on it is called a *measurable space*. A σ-additive m

Definition B.2. *(Probability measure)*
Given a measurable space $(\Omega, \mathcal{F})$, Probability measure is a map $\mathbb{P} : \mathcal{F} \to [0, 1]$ that satisfies:

1. $\forall \vartheta \in \mathcal{F}$, $\mathbb{P}(\vartheta) \in [0, 1]$

2. $\mathbb{P}(\varnothing) = 0$ *and* $\mathbb{P}(\Omega) = 1$

3. For any countable collection $\{A_n \in \mathcal{F}\}_{n=1}^{\infty}$ *of disjoint measurable sets such that* $\forall n \neq m$, $A_n \cap A_m = \varnothing$, $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$

A tuple $(\Omega, \mathcal{F}, \mathbb{P})$ of the event set Ω, σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$ on $\mathcal{F}$ is called a *Probability Measure space.*

B.2 Measurable Functions

Definition B.3. *($(\mathcal{F} \mid \mathcal{B})$-measurable function)*
Let $(\Omega, \mathcal{F})$ *and* $(\mathcal{Z}, \mathcal{B})$ *be two measurable spaces and let* $X : \Omega \to \mathcal{Z}$ *be a function map between them.* X *is said to be a* $(\mathcal{F} \mid \mathcal{B})$*-measurable function if for all* $\beta \in \mathcal{B}$*, the preimage set* $X^{-1}(\beta) := \{\omega \in \Omega : X(\omega) \in \beta\} \in \mathcal{F}$.

Often $\mathcal{Z}$ is taken to be a topological space with topology $\Theta_{\mathcal{Z}}$ and $\mathcal{B}$ is taken to be the Borel sigma algebra generated by the open sets in $\Theta_{\mathcal{Z}}$, denoted as $\mathcal{B}(\mathcal{Z})$.

Definition B.4. *(Law of a measurable function)*
Let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability measure space,* $(\mathcal{Z}, \mathcal{B})$ *be a measurable space and* $X : \Omega \to \mathcal{Z}$ *be a* $(\mathcal{F}|\mathcal{B})$*-measurable function. The law of a measurable function* X *is defined as the probability measure* $\mathcal{L}(X) : \mathcal{B} \to [0, 1]$*, given by* $(\mathcal{L}(X))(\beta) = \mathbb{P}(X^{-1}(\beta))$ *for all* $\beta \in \mathcal{B}$.

Definition B.5. *(Indicator function)*
Let $(\Omega, \mathcal{F})$ *and* $(\mathcal{Z}, \mathcal{B})$ *be two measurable spaces. Let* $\mathcal{Z}$ *be a set with a commutative binary operator,* $\circ : \mathcal{Z} \times \mathcal{Z} \to \mathcal{Z}$*, defined for its members and let there exist unique members in* $\mathcal{Z}$*,* $0_{\mathcal{Z}}$ *and* $1_{\mathcal{Z}}$*, such that for all* $z \in \mathcal{Z}$*,* $z \circ 0_{\mathcal{Z}} = 0_{\mathcal{Z}}$ *and* $z \circ 1_{\mathcal{Z}} = z$*. For any* $\vartheta \in \mathcal{F}$*, define the indicator function* $I_{\vartheta} : \Omega \to \mathcal{Z}$ *such that* $I_{\vartheta}(\omega) = \begin{cases} 1_{\mathcal{Z}} & \text{if } \omega \in \vartheta \\ 0_{\mathcal{Z}} & \text{otherwise} \end{cases}$

The following lemma shows that any indicator function defined for a measurable set $\vartheta \in \mathcal{F}$ is $(\mathcal{F}|\mathcal{B})$-measurable.

Lemma B.1. *(Measurability of indicator function)*
For all measurable spaces $(\Omega, \mathcal{F})$*,* $(\mathcal{Z}, \mathcal{B})$ *and* $\vartheta \in \mathcal{F}$*, the indicator function* $I_{\vartheta} : \Omega \to \mathcal{Z}$*, as given by Definition B.5, is a* $(\mathcal{F}|\mathcal{B})$*-measurable function.*

Proof: *Note that for any* $\beta \in \mathcal{B}$*,* $I_{\vartheta}^{-1}(\beta) = \begin{cases} \varnothing & \text{if } \beta \cap \{0_{\mathcal{Z}}, 1_{\mathcal{Z}}\} = \varnothing \\ \Omega & \text{if } 0_{\mathcal{Z}} \in \beta, 1_{\mathcal{Z}} \in \beta \\ \vartheta^c & \text{if } 0_{\mathcal{Z}} \in \beta, 1_{\mathcal{Z}} \notin \beta \\ \vartheta & \text{otherwise} \end{cases}$*. Since* $\varnothing, \Omega, \vartheta$ *and* ϑ^c *all belong to* $\mathcal{F}$*, for all* $\beta \in \mathcal{B}$*,* $I_{\vartheta}^{-1}(\beta) \in \mathcal{F}$*, implying* I_{ϑ} *is* $(\mathcal{F}|\mathcal{B})$*-measurable.* □

An indicator function I_{ϑ} is often used to restrict the support of other measurable functions X to a measurable set ϑ of interest by taking a product of the functions $X \cdot I_{\vartheta}$. The following lemma shows that a function defined through such a product is also $(\mathcal{F}|\mathcal{B})$-measurable.

Lemma B.2. *(Measurability of product with indicator function)*
Let $(\Omega, \mathcal{F}), (\mathcal{Z}, \mathcal{B})$ be as defined in Definition B.5. Let $X : \Omega \to \mathcal{Z}$ be a $(\mathcal{F}|\mathcal{B})$-measurable function. Then for every $\vartheta \in \mathcal{F}$, the indicator function $I_\vartheta : \Omega \to \mathcal{Z}$ is such that the product of functions $XI_\vartheta : \Omega \to \mathcal{Z}$, defined as $XI_\vartheta(\omega) = X(\omega) \circ I_\vartheta(\omega)$, is $(\mathcal{F}|\mathcal{B})$-measurable.

Proof: *Note that for any $\beta \in \mathcal{B}$, $(XI_\vartheta)^{-1}(\beta) = \begin{cases} [X^{-1}(\beta) \cap \vartheta] \cup \vartheta^c & \text{if } 0_\mathcal{Z} \in \beta \\ X^{-1}(\beta) \cap \vartheta & \text{otherwise} \end{cases}$. Since $X^{-1}(\beta) \in \mathcal{F}$ (by measurability of X) and $\vartheta, \vartheta^c \in \mathcal{F}$, both $X^{-1}(\beta) \cap \vartheta$ and $[X^{-1}(\beta) \cap \vartheta] \cup \vartheta^c$ belong to $\mathcal{F}$ (by union and intersection properties of measurable sets). Thus $XI_\vartheta : \Omega \to \mathcal{Z}$ is $(\mathcal{F}|\mathcal{B})$-measurable.* □

B.3 Integration of Measurable Functions

For the purpose of establishing an integral of measurable functions we need a fully ordered set $\mathcal{Z}$ with an attainable infimum element and let $\mathcal{Z}$ be a subset of a real vector space (such that multiplication by real scalars and addition is defined on $\mathcal{Z}$). Such a $\mathcal{Z}$ is isomorphic to $\mathbb{R}$ or a subset of $\mathbb{R}$. To begin with, we can simply consider, $\mathcal{Z} = [0, \infty) \subset \mathbb{R}$ which has these properties.

Definition B.6. *(Lebesgue integral for fully ordered, infimum attaining $\mathcal{Z}$)*
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a (probability) measure space and $(\mathcal{Z}, \mathcal{B})$ be a measurable space with a well defined addition operation and an attainable infimum in $\mathcal{Z}$ (e.g. $(\mathcal{Z}, \mathcal{B}) = ([0, \infty), \mathcal{B})$). Let $\vartheta \in \mathcal{F}$ be any measurable set and I_ϑ be the corresponding indicator function. Let $X : \Omega \to \mathcal{Z}$ be a non-negative $(\mathcal{F}|\mathcal{B})$-measurable function. Let $\phi = \{\vartheta_i \in \mathcal{F} : \cup\vartheta_i = \mathcal{Z}, \forall i \neq j, \vartheta_i \cap \vartheta_j = \varnothing\}$ be a finite decomposition of disjoint measurable sets for $\mathcal{Z}$ and $\Phi = \{\phi\}$ be a collection of all possible finite decompositions of $\mathcal{Z}$. Then an integral with respect to the measure $\mathbb{P}$ is defined on ϑ as

$$\int_\vartheta X d\mathbb{P} := \sup_{\phi \in \Phi} \sum_{\vartheta_i \in \phi} \left[\inf_{\omega \in \vartheta_i} XI_\vartheta(\omega) \right] \mathbb{P}(\vartheta_i) \tag{B.1}$$

Since the infimum is always attainable in $\mathcal{Z}$, the infimum in (B.1) is well defined. Further since the supremum is considered over all possible measurable set covers for ϑ, the integral value evaluated is unique and the integral is thus well defined. Note that the integral can still evaluate to ∞ and is still considered as well defined. Further if $\mathcal{B}$ is the Borel σ-algebra then the definition in (B.1) is consistent with the dual definition of the integral obtained by interchanging the infimum and supremum (i.e. $\int_\vartheta X d\mathbb{P} := \sup_{\phi \in \Phi} \sum_{\vartheta_i \in \phi} [\inf_{\omega \in \vartheta_i} XI_{\vartheta_i}(\omega)] \mathbb{P}(\vartheta_i) = \inf_{\phi \in \Phi} \sum_{\vartheta_i \in \phi} \left[\sup_{\omega \in \vartheta_i} XI_{\vartheta_i}(\omega)\right] \mathbb{P}(\vartheta_i)$, see [100, Exercise 15.2]).

For a more general $\mathcal{Z}$ where an infimum is not always attainable (e.g. $\mathcal{Z} = \mathbb{R}$), but $\mathcal{Z}$ fully ordered, we split the set $\mathcal{Z}$ using some $z_0 \in \mathcal{Z}$ into subsets $\mathcal{Z}^+ := \{z \in \mathcal{Z} : z \geq z_0\}$ and $\mathcal{Z}^- := \{z \in \mathcal{Z} : z \leq z_0\}$ for which an infimum and supremum respectively are attainable. Let $\mathcal{B}$ be such that $\mathcal{Z}^+$ and $\mathcal{Z}^-$ are measurable sets and X be $(\mathcal{F}|\mathcal{B})$-measurable. Then consider the $(\mathcal{F}|\mathcal{B})$-measurable functions $X^+ := XI_{X^{-1}(\mathcal{Z}^+)}$ and $X^- := -(XI_{X^{-1}(\mathcal{Z}^-)})$ and define the integral $\int_\vartheta X d\mathbb{P} := \int_\vartheta X^+ d\mathbb{P} - \int_\vartheta X^- d\mathbb{P}$. For a consistent definition of the integral such that it does not depend on the value of z_0 that we use to split $\mathcal{Z}$, we must ensure firstly that $\mathcal{Z}^+$ and $\mathcal{Z}^-$ are measurable sets in $\mathcal{B}$ for any $z_0 \in \mathcal{Z}$ and secondly that the value of $\int_\vartheta X^+ d\mathbb{P} - \int_\vartheta X^- d\mathbb{P}$ is independent of z_0. Theorem B.1 below shows that $\mathcal{B}$ being the Borel σ-algebra on $\mathcal{Z}$ is necessary and sufficient to ensure these two requirements for arbitrary measure spaces $(\Omega, \mathcal{F}, \mathbb{P})$. Thus a for a general integral definition on fully ordered $\mathcal{Z}$ we must restrict ourselves to Borel measurable functions X.

Definition B.7. *(Integration of Borel measurable functions)*
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a (probability) measure space and $(\mathcal{Z}, \mathcal{B})$ be a measurable space with $\mathcal{Z}$ fully ordered and $\mathcal{B}$ the Borel σ-algebra on $\mathcal{Z}$. Let $X : \Omega \to \mathcal{Z}$ be $(\mathcal{F}|\mathcal{B})$-measurable. For a fixed $z_0 \in \mathcal{Z}$, let $\mathcal{Z}^+ := \{z \in \mathcal{Z} : z \geq z_0\}$ and $\mathcal{Z}^- := \{z \in \mathcal{Z} : z \leq z_0\}$ be the measurable sets in $\mathcal{B}$ and $X^+ := XI_{X^{-1}(\mathcal{Z}^+)}$ and $X^- := -(XI_{X^{-1}(\mathcal{Z}^-)})$ be the corresponding $(\mathcal{F}|\mathcal{B})$-measurable functions with range $\{z \in \mathcal{Z} : z \geq z_0\}$. Then using Definition B.6,

$$\int_\vartheta X d\mathbb{P} = \int_\vartheta X^+ d\mathbb{P} - \int_\vartheta X^- d\mathbb{P} \tag{B.2}$$

The above integral is well defined as long as both $\int_\vartheta X^+ d\mathbb{P}$ and $\int_\vartheta X^- d\mathbb{P}$ are not ∞ at the same time. Without loss of generality, for $\mathcal{Z} = \mathbb{R}$, z_0 can be taken to be 0 (by Theorem B.1).

Theorem B.1. *(Borel measurability for Integral consistency)*
For Lebesgue integral as defined by (B.2), the integral value is independent of the choice of z_0 if and only if $\mathcal{B}$ is a Borel σ-algebra generated by the sets $\{\{z \in \mathcal{Z} : z > z_0\} : z_0 \in \mathcal{Z}\}$

Proof: *Note that if $\mathcal{B}$ is the Borel σ-algebra on $\mathcal{Z}$, then Z^+ and Z^- are measurable sets in $\mathcal{Z}$ for any z_0 and thus $X^{-1}(\mathcal{Z}^+)$ and $X^{-1}(\mathcal{Z}^-)$ are measurable sets in $\mathcal{F}$ ($\because X$ is $(\mathcal{F}|\mathcal{B})$-measurable). Then by Lemma B.2, $X^+ := XI_{X^{-1}(\mathcal{Z}^+)}$ and $X^- := -(XI_{X^{-1}(\mathcal{Z}^-)})$ are $(\mathcal{F}|\mathcal{B})$-measurable and the integral for them can be defined as given by (B.1).*

The uniqueness of the integral value can be seen by considering splitting around two values $z_0, z_0' \in \mathcal{Z}$ and without loss of generality $z_0 < z_0'$. Let $\mathcal{Z}^+_{z_0} := \{z \in \mathcal{Z} : z \geq z_0\}$, $\mathcal{Z}^-_{z_0} := \{z \in \mathcal{Z} : z \leq z_0\}$ and $\mathcal{Z}^+_{z_0'} := \{z \in \mathcal{Z} : z \geq z_0'\}$, $\mathcal{Z}^-_{z_0'} := \{z \in \mathcal{Z} : z \leq z_0'\}$. Similarly let $X^+_{z_0} = XI_{X^{-1}(\mathcal{Z}^+_{z_0})}$, $X^-_{z_0} = -XI_{X^{-1}(\mathcal{Z}^-_{z_0})}$ and $X^+_{z_0'} = XI_{X^{-1}(\mathcal{Z}^+_{z_0'})}$, $X^-_{z_0'} = -XI_{X^{-1}(\mathcal{Z}^-_{z_0'})}$. Let $\int_\vartheta X d\mathbb{P} = \int_\vartheta X^+_{z_0} d\mathbb{P} - \int_\vartheta X^-_{z_0} d\mathbb{P}$ and note that $\int_\vartheta X d\mathbb{P} = \int_\vartheta X^+_{z_0} d\mathbb{P} - \int_\vartheta X^-_{z_0} d\mathbb{P} = \int_\vartheta X^+_{z_0'} d\mathbb{P} - \int_\vartheta -XI_{X^{-1}([z_0, z_0'])} d\mathbb{P} - \int_\vartheta X^-_{z_0} d\mathbb{P} = \int_\vartheta X^+_{z_0'} d\mathbb{P} - \int_\vartheta X^-_{z_0'} d\mathbb{P}$.

Thus $\mathcal{B}$ being the Borel σ-algebra is sufficient for the integral to be well defined and unique, independent of the choice of z_0.

Further if $\mathcal{B}$ is not the Borel σ-algebra on $\mathcal{Z}$, then there must exist a z_0 such that $Z^+ = \{z \in \mathcal{Z} : z \geq z_0\} \notin \mathcal{B}$ and thus X^+ is not $(\mathcal{F}|\mathcal{B})$-measurable. For such a z_0, the integral will not be well defined by B.1. Thus $\mathcal{B}$ being a Borel σ-algebra is necessary for (B.2) to be well defined for any z_0. □

The Lebesgue integral can be further extended to any general Banach space $\mathcal{Z}$ (e.g. $\mathbb{R}^n$ and $L^p(\mathcal{X}, \mathcal{G}, \mu)$ function spaces) using the notions of Pettis and Bochner integrals [101, Chapter 1].

B.3.1 Integration of Banach-valued functions

Let $\mathcal{Z}$ be a Banach space, $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z}$ and $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be a corresponding measurable space. Let $\mathcal{Z}^\star$ be the dual space containing all linear, continuous functionals on $\mathcal{Z}$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and let $F_{\Omega,\mathcal{Z}}$ be the space of all $(\mathcal{F}|\mathcal{B}(\mathcal{Z}))$-measurable functions. The notions of Pettis and Bochner integration for $X \in F_{\Omega,\mathcal{Z}}$ are defined as follows.

Definition B.8. *(Pettis Integral)*
$X \in F_{\Omega,\mathcal{Z}}$ is said to be Pettis integrable if for each $\vartheta \in \mathcal{F}$, there exists a $m_\vartheta \in \mathcal{Z}$ such that for all $h \in \mathcal{Z}^\star$, $h(m_\vartheta) = \int_\vartheta h(X(\omega))\mathbb{P}(d\omega)$ and we denote integration in the Pettis sense with $m_\vartheta = (P)\int_\vartheta X d\mathbb{P}$.

This is equivalent to the intuitively defined component wise integration for $\mathbb{R}^n$-valued functions as n functionals corresponding to the standard orthonormal basis for $\mathbb{R}^n$ span the whole dual space in $\mathbb{R}^n$.

Definition B.9. *(Bochner integration)*
$X \in F_{\Omega,\mathcal{Z}}$ is said to be Bochner integrable if for each $\vartheta \in \mathcal{F}$, there exists a sequence of simple functions $\{X_n \in F_{\Omega,\mathcal{Z}}\}$ such that $\lim_{n\to\infty} \int_\vartheta ||X_n(\omega) - X(\omega)||_{\mathcal{Z}}\mathbb{P}(d\omega) = 0$ and we denote the integration in the Bochner sense with $(B)\int_\vartheta X d\mathbb{P} = \lim_{n\to\infty}(L)\int_\vartheta X_n d\mathbb{P}$.

The necessary and sufficient condition for $X \in F_{\Omega,\mathcal{Z}}$ to be Bochner integrable is that $\int_\Omega ||X||_{\mathcal{Z}} d\mathbb{P} < \infty$ [101, Theorem 1.8]. Furthermore it can be shown that every Bochner integrable function is also Pettis integrable and that the integrals have the same value [101]. Thus from here on we will consider integrability for Banach-valued functions in the sense of Bochner. Pettis integrability follows automatically if the function is Bochner integrable and will be used when convenient with the understanding that the value computed is the same as that for the Bochner integral. Also we will simply, denote the Bochner integral as $\int_\vartheta$ instead of $(B)\int_\vartheta$.

B.4 Expectation and moments of measurable functions

Definition B.10. *(Expectation)*
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and $X : \Omega \to \mathcal{Z}$ be a $(\mathcal{F}|\mathcal{B})$-measurable function. The expectation of X, denoted $\mathbb{E}[X]$, is defined as $\mathbb{E}[X] = \int_\Omega X d\mathbb{P}$.

The set of $(\mathcal{F}|\mathcal{B})$-measurable functions satisfying $\mathbb{E}[\,||X||_{\mathcal{Z}}\,] < \infty$ is denoted $L^1(\Omega, \mathcal{F}, \mathbb{P})$. In general, $L^p(\Omega, \mathcal{F}, \mathbb{P}) = \{X \in F_{\Omega,\mathcal{Z}} : \mathbb{E}[||X||^p_{\mathcal{Z}}] < \infty\}$.

For $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, X is Bochner and thus Pettis integrable. The expectation of X can also thus be regarded as a map from $\mathcal{Z}^\star \to \mathbb{R}$ as $\mathbb{E}[X] : \mathcal{Z}^\star \to \mathbb{R}$ given by $\mathbb{E}[X](h) = \int_\Omega h(X(\omega))\mathbb{P}(d\omega)$ and is called the first moment of X.

For $k \in \mathbb{N}$, we define the k^{th} moment and central moments as a tensor, $M_k : F^k_{\Omega,\mathcal{Z}} \times \mathcal{Z}^{\star k} \to \mathbb{R}$ as follows.

Definition B.11. *(Moments)*
For $k \in \mathbb{N}$, $X_1, \ldots, X_k \in F_{\Omega,\mathcal{Z}}$ and $h_1, \ldots, h_k \in \mathcal{Z}^\star$ we define the k^{th} moment tensor $M_k((X_1, \ldots, X_k), (h_1, \ldots, h_k)) = \int_\Omega h_1(X_1(\omega)) \cdot h_2(X_2(\omega)) \cdot \ldots h_k(X_k(\omega))\mathbb{P}(d\omega)$.

Note thus that $M_1(X, h) = \mathbb{E}[X](h) = h(\mathbb{E}[X])$. Further for a fixed $(X_1, \ldots, X_k)$, we denote by $M_k(X_1, \ldots, X_k) : \mathcal{Z}^k \to \mathbb{R}$, the section $M_k(X_1, \ldots, X_k)(h_1, \ldots, h_k) = M_k((X_1, \ldots, X_k), (h_1, \ldots, h_k))$.

Definition B.12. *(Central Moments)*
For $k \in \mathbb{N}$, $X_1, \dots, X_k \in F_{\Omega,\mathcal{Z}}$ and $h_1, \dots, h_k \in \mathcal{Z}^\star$ we define the k^{th} central moment tensor $C_k((X_1, \dots, X_k), (h_1, \dots, h_k)) = \int_\Omega h_1(X_1(\omega) - \mathbb{E}[X_1]) \cdot h_2(X_2(\omega) - \mathbb{E}[X_2]) \cdot \dots h_k(X_k(\omega) - \mathbb{E}[X_k])\mathbb{P}(d\omega)$.

Similar to the moments by $C_k(X_1, \dots, X_k) : \mathcal{Z}^{\star k} \to \mathbb{R}$ we denote the section $C_k(X_1, \dots, X_k)(h_1, \dots, h_k) = C_k((X_1, \dots, X_k), (h_1, \dots, h_k))$. In particular $C_2((X_1, X_2)) : \mathcal{Z}^{\star 2} \to \mathbb{R}$ is called the covariance between X_1 and X_2.

Definition B.13. *(Covariance)*
For $X_1, X_2 \in F_{\Omega,\mathcal{Z}}$ we define the covariance of X_1, X_2 as the central moment tensor $C_2(X_1, X_2)$.

B.5 Probability density

Definition B.14. *(Probability density function)*
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space. If there exists a $(\mathcal{F}|\mathcal{B}(\mathbb{R}))$-measurable function $\delta : \Omega \to [0, \infty)$ and a measure $\mu : \mathcal{F} \to \mathbb{R}$ such that for all $A \in \mathcal{F}$, $\mathbb{P}(A) = \int_A \delta d\mu$, then δ is said to be the density of $\mathbb{P}$ with respect to the measure μ.

The question of when such a density function δ exists for a measure $\mathbb{P}$ is addressed by the Radon-Nikodym theorem presented in Section B.6.

Quite commonly with $\Omega = \mathbb{R}^n$, μ is taken to be the Lebesgue measure on $\mathbb{R}^n$. Then the density with respect to this Lebesgue measure such that $\mathbb{P}(A) = \int_A \delta dx$ refers to the familiar notions of probability density functions on $\mathbb{R}^n$ (e.g. the Gaussian density function).

B.6 Radon-Nikodym Theorem

Let $\nu, \mu \in \mathcal{M}_\sigma(\Omega, \mathcal{F})$ be two signed bounded measures on a σ-measurable space $(\Omega, \mathcal{F})$, one question motivated by the existence of density functions, is to ask when does such a $(\mathcal{F}|\mathcal{B}(\mathbb{R}))$-measurable density function $\delta : \Omega \to [0, \infty)$ exist, such that $\forall A \in \mathcal{F}$, $\nu(A) = \int_A \delta d\mu$ (see for example the existence question for probability density functions in Section B.5).

The Radon-Nikodym theorem attempts to answer this question. In order to present the theorem though the following terms need to be defined first.

Definition B.15. *(Mutually singular measures)*
Let μ, ν be signed measures on a measurable space $(\Omega, \mathcal{F})$. μ and ν are said to be mutually singular, denoted $\mu \perp \nu$, if there exist measurable sets $S_\mu, S_\nu \in \mathcal{F}$ such that

$$\mu(S_\mu) = 0, \quad \nu(S_\nu) = 0, \quad S_\mu \cap S_\nu = \varnothing \quad \textit{and} \quad S_\mu \cup S_\nu = \Omega \tag{B.3}$$

Definition B.16. *(Absolute continuity of measures)*
Let μ, ν be signed measures on a measurable space $(\Omega, \mathcal{F})$. ν is said to be absolutely continuous with respect to μ, denoted $\nu << \mu$, if for all $\vartheta \in \mathcal{F}$, $\mu(\vartheta) = 0$ implies $\nu(\vartheta) = 0$.

Theorem B.2. *(Radon-Nikodym Theorem)*
For any signed measures μ, ν on $(\Omega, \mathcal{F})$ with $\nu << \mu$, there exists a density function $\delta \in L^1(\Omega, \mathcal{F}, \mu)$ such that for all $A \in \mathcal{F}$, $\nu(A) = \int_A \delta d\mu$

For proof, we refer the reader to the proof for [102, Theorem 28.1].

Further it is known that for any signed measure $\mu \in \mathcal{M}_\sigma$, the band $B_\mu = \{\nu \in \mathcal{M}_\sigma : \nu(A) = \int_A f_\nu d\mu, f_\nu \in L^1(\Omega, \mathcal{F}, \mu)\}$ is such that $B_\mu \oplus (B_\mu)^d = \mathcal{M}_\sigma$ (by Exercise 27.5 and Theorem 12.2 in [102]). Thus any measure $\lambda \in \mathcal{M}_\sigma$ in terms of $\mu \in \mathcal{M}_\sigma$ can be written as $\lambda = \nu + \delta$ with $\nu \in B_\mu$ and $\delta \in (B_\mu)^d$, i.e., $\nu << \mu$ and $\delta \perp \mu$.

B.6.1 Conditional Expectation

Definition B.17. *(Conditional Expectation)*
For a probability measure space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ-algebra. Given a $(\mathcal{F}|\mathcal{B})$-measurable function $X : \Omega \to \mathcal{Z}$ the conditional expectation, denoted $\mathbb{E}[X||\mathcal{G}]$, is a $(\mathcal{G}|\mathcal{B})$-measurable function satisfying,

$$\forall \vartheta \in \mathcal{G} \qquad \int_\vartheta (\mathbb{E}[X|\mathcal{G}] - X) d\mathbb{P} \tag{B.4}$$

The notion of conditional expectation represents a fair policy under partial information contained in $\mathcal{G}$, i.e., if one were to get a random return of X with probability $\mathbb{P}$ without knowing completely which $\nu \in \mathcal{F}$ occurred, but only a partial observation on which $\vartheta \supset \nu$ in $\mathcal{G}$ occurred, is available, a fair price (zero expected loss) is given by $\mathbb{E}[X|\mathcal{G}]$.

The $(\mathcal{G}|\mathcal{B})$-measurable function $\mathbb{E}[X|\mathcal{G}]$ is not unique, however any two versions of $\mathbb{E}[X|\mathcal{G}]$ are $\mathbb{P} - a.s.$ equal ([100, Section 34]). Further it is easy to see that $\mathbb{E}[X|\{\varnothing, \Omega\}] = \mathbb{E}[X]$ and $\mathbb{E}[X|\mathcal{F}] = X$.

B.7 Gaussian Measures

The presentation here is restricted to Gaussian measures on separable Banach spaces, for a more detailed presentation on locally convex spaces, the reader is referred to [26, Chapter 2].

We first introduce the Gaussian measure and density on $\mathbb{R}$ and use that notion to define a Gaussian measure on any (possibly infinite dimensional) separable Banach space.

Definition B.18. *(Gaussian density on $\mathbb{R}$)*
For known scalar constants $m \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$, let $p_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ be the probability density with respect to the Lebesgue measure on $\mathbb{R}$ for a measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, 1]$ such that $\forall A \in \mathcal{B}(\mathbb{R})$, $\mu(A) = \int_A p_{m,\sigma}(x)dx$. μ is called a non-degenerate Gaussian measure on $\mathbb{R}$ and $p_{m,\sigma}$ the Gaussian density function.

However, not all Gaussian measures are absolutely continuous with respect to the Lebesgue measure and have a Gaussian density function, if we are to consider the useful notion of Dirac measures as degenerate Gaussian measures. Thus in general we define the Gaussian measure (including the degenerate case) as follows,

Definition B.19. *(Gaussian measure on $\mathbb{R}$)*
A probability measure $\mu : \mathcal{B}(\mathbb{R}) \to [0,1]$ is called a Gaussian measure on $\mathbb{R}$, if it has a Gaussian density $p_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$, for some known scalar constants $m \in \mathbb{R}$ and $\sigma^2 \in (0,\infty)$, or if it is a Dirac measure δ_m, centered at some $m \in \mathbb{R}$.

Definition B.20. *(Gaussian measure on Banach spaces)*
Let $\mathcal{Z}$ be a separable Banach space and $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z}$. Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be the measurable space on the real line and let $\mathcal{Z}^\star$ be the dual space to $\mathcal{Z}$. A probability measure $\mu : \mathcal{B}(\mathcal{Z}) \to [0,1]$ is said to be Gaussian if for all $h \in \mathcal{Z}^\star$ the law $\mu \circ h^{-1} : \mathcal{B}(\mathbb{R}) \to [0,1]$ is a Gaussian measure on $\mathbb{R}$.

Definition B.21. *(Mean and Covariance functions)*
Let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be a measurable Banach space. Given a measure $\mu : \mathcal{B}(\mathcal{Z}) \to [0,1]$ for any $h \in \mathcal{Z}^\star$, the law $\mu \circ h^{-1} : \mathcal{B}(\mathbb{R}) \to [0,1]$ is said to have a mean function $m_\mu : \mathcal{Z}^\star \to \mathbb{R}$ defined as $m_\mu(h) = \mathbb{E}[h] = \int_{\mathcal{Z}} h d\mu$ and covariance function $\Sigma_\mu : \mathcal{Z}^\star \times \mathcal{Z}^\star \to \mathbb{R}$ defined as $\Sigma_\mu(h,g) = \mathbb{E}[hg] - m_\mu(h)m_\mu(g) = \int_{\mathcal{Z}} (h(\zeta) - m_\mu(h))(g(\zeta) - m_\mu(g))\mu(d\zeta)$.

Theorem B.3. *(Fernique theorem - integrability of Gaussian measure)*
Let μ be a Gaussian measure on a separable Banach space $\mathcal{Z}$. Then there exists a real positive scalar $\alpha > 0$ such that

$$\int_{\mathcal{Z}} e^{\alpha ||x||^2_{\mathcal{Z}}} \mu(dx) < \infty \tag{B.5}$$

We refer the reader to [103, Theorem 2.3.1, Corollary 3.3.2] for the proof. As corollaries to the Fernique theorem we get the boundedness of all moments for a Gaussian measure and continuity of the mean and covariance functions.

Corollary B.1. *(Bounded moments)*
Let μ be a Gaussian measure on a separable Banach space $\mathcal{Z}$. Then for every $1 \leq p < \infty$,

$$\int_{\mathcal{Z}} ||x||^p_{\mathcal{Z}} \mu(dx) < \infty \tag{B.6}$$

Proof: *Since $||x||^p_{\mathcal{Z}} \leq c_{\alpha,p} e^{\alpha ||x||^2_{\mathcal{Z}}}$ for some finite constant $c_{\alpha,p}$, we have $\int_{\mathcal{Z}} ||x||^p_{\mathcal{Z}} \mu(dx) \leq c_{\alpha,p} \int_{\mathcal{Z}} e^{\alpha ||x||^2_{\mathcal{Z}}} \mu(dx) < \infty$.* □

Corollary B.2. *(Integrability of $\mathcal{Z}^\star$)*
Let μ be a Gaussian measure on a separable Banach space $\mathcal{Z}$. Then for every $1 \leq p < \infty$ and for all $h \in \mathcal{Z}^\star$,

$$\int_{\mathcal{Z}} |h(x)|^p \mu(dx) < \infty \tag{B.7}$$

Proof: *Note that by continuity of $h \in \mathcal{Z}^\star$, $|h(x)| \leq ||h||_{\mathcal{Z}^\star} ||x||_{\mathcal{Z}}$ and thus $\int_{\mathcal{Z}} |h(x)|^p \mu(dx) \leq ||h||^p_{\mathcal{Z}^\star} \int_{\mathcal{Z}} ||x||^p_{\mathcal{Z}} \mu(dx) < \infty$.* □

Corollary B.3. *(Continuous embedding of $\mathcal{Z}^\star$)*
There exists a continuous inclusion $i : \mathcal{Z}^\star \to L^2(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ such that $\forall h \in \mathcal{Z}^\star$, $i(h) = h$ and there exists a constant $\kappa \in (0, \infty)$ such that $\forall h \in \mathcal{Z}^\star$, $i(h) = h$ and $||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)} \leq \kappa ||h||_{\mathcal{Z}^\star}$.

Proof: *Note that $||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)} = \left(\int_\mathcal{Z} |h(\zeta)|^2 d\mu(\zeta)\right)^{1/2}$ and by continuity of $h \in \mathcal{Z}^\star$, $|h(\zeta)| \leq ||h||_{\mathcal{Z}^\star} ||\zeta||_\mathcal{Z}$. Thus $||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)} = \left(\int_\mathcal{Z} |h(\zeta)|^2 d\mu(\zeta)\right)^{1/2} \leq \left(\int_\mathcal{Z} ||h||^2_{\mathcal{Z}^\star} ||\zeta||^2_\mathcal{Z} d\mu(\zeta)\right)^{1/2} = ||h||_{\mathcal{Z}^\star} \left(\int_\mathcal{Z} ||\zeta||^2_\mathcal{Z} d\mu(\zeta)\right)^{1/2}$. By Corollary B.1, $\left(\int_\mathcal{Z} ||\zeta||^2_\mathcal{Z} d\mu(\zeta)\right)^{1/2} < \infty$ and thus there exists a $\kappa = \left(\int_\mathcal{Z} ||\zeta||^2_\mathcal{Z} d\mu(\zeta)\right)^{1/2} \in (0, \infty)$ such that $||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)} \leq \kappa ||h||_{\mathcal{Z}^\star}$.* □

Corollary B.4. *(Continuity of mean and covariance functions)*
Let μ be a Gaussian measure on a separable Banach space $\mathcal{Z}$. The mean function $m_\mu : \mathcal{Z}^\star \to \mathbb{R}$ is continuous (i.e. there exists a finite constant $\kappa_m \in (0, \infty)$ such that for all $h \in \mathcal{Z}^\star$, $|m_\mu(h)| \leq \kappa ||h||_{\mathcal{Z}^\star}$). The covariance function $\Sigma_\mu : \mathcal{Z}^\star \times \mathcal{Z}^\star \to \mathbb{R}$ is continuous (i.e. there exists a finite constant $\kappa_\sigma \in (0, \infty)$ such that for all $h, g \in \mathcal{Z}^\star$, $|\Sigma_\mu(h, g)| \leq \kappa_\sigma ||h||_{\mathcal{Z}^\star} ||g||_{\mathcal{Z}^\star}$)

Proof: *Note that $|m_\mu(h)| = |\int_\mathcal{Z} h(x)\mu(dx)| \leq \int_\mathcal{Z} |h(x)|\mu(dx) \leq ||h||_{\mathcal{Z}^\star} \int_\mathcal{Z} ||x||_\mathcal{Z} \mu(dx)$ (by continuity of h). Since $\int_\mathcal{Z} ||x||_\mathcal{Z} \mu(dx) < \infty$, there exists a constant $\kappa_m = \int_\mathcal{Z} ||x||_\mathcal{Z} \mu(dx) \in (0, \infty)$ such that $m_\mu(h) \leq \kappa_m ||h||_{\mathcal{Z}^\star}$. Similarly $|\Sigma_\mu(h, g)| = |\int_\mathcal{Z} h(x)g(x)\mu(dx) - m_\mu(h)m_\mu(g)| \leq \int_\mathcal{Z} |h(x)||g(x)|\mu(dx) + |m_\mu(h)||m_\mu(g)| \leq ||h||_{\mathcal{Z}^\star} ||g||_{\mathcal{Z}^\star} \int_\mathcal{Z} ||x||^2_\mathcal{Z} \mu(dx) + \kappa_m^2 ||h||_{\mathcal{Z}^\star} ||g||_{\mathcal{Z}^\star}$. Since $\int_\mathcal{Z} ||x||^2_\mathcal{Z} \mu(dx) < \infty$, there exists a constant $\kappa_\sigma = (\int_\mathcal{Z} ||x||^2_\mathcal{Z} \mu(dx) + \kappa_m^2) \in (0, \infty)$ such that $|\Sigma_\mu(h, g)| \leq \kappa_\sigma ||h||_{\mathcal{Z}^\star} ||g||_{\mathcal{Z}^\star}$.* □

Theorem B.4. *(Representer for the mean function)*
Let μ be a Gaussian measure on a separable Banach space $\mathcal{Z}$. There exists a representer $m \in \mathcal{Z}$ for the mean function $m_\mu : \mathcal{Z}^\star \to \mathbb{R}$ such that for all $h \in \mathcal{Z}^\star$, $h(m) = m_\mu(h)$.

Proof: *By [104, Proposition 3.14], there exists a $m \in \mathcal{Z}$ such that for all $h \in \mathcal{Z}^\star$, $m_\mu(h) = h(m)$ if $m_\mu : \mathcal{Z}^\star \to \mathbb{R}$ is continuous with respect to the weak* topology $\sigma(\mathcal{Z}^\star, \mathcal{Z})$. Further by [104, Theorem 3.28] for separable Banach spaces, weak* continuity is equivalent to continuity along weak* convergent sequences. For a sequence $f_n \in \mathcal{Z}^\star$ converging in the weak* topology to $f \in \mathcal{Z}^\star$, there exists a finite constant $M \in (0, \infty)$ such that for all n, $||f_n||_{\mathcal{Z}^\star} < M$ [104, Proposition 3.13(iii)]. Then by the Lebesgue dominated convergence theorem, $\lim_{n\to\infty} m_\mu(f_n) = \lim_{n\to\infty} \int_\mathcal{Z} f_n d\mu = \int_\mathcal{Z} f d\mu = m_\mu(f)$, implying m_μ is continuous along each weak* convergent sequence and thus by [104, Proposition 3.14], there exists a $m \in \mathcal{Z}$ such that for all $h \in \mathcal{Z}^\star$, $m_\mu(h) = h(m)$.* □

Appendix C

Function spaces

C.1 Banach space of continuous, bounded functions $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$

We refer the reader to [105, Section IV.6.1] for the proofs of theorems presented below and to [106] for a summarized version of the results.

Let $\mathcal{X}$ be a normal (T_4-separable) topological space, that is, the points in $\mathcal{X}$ are all closed and two disjoint closed sets can be separated by open neighborhoods. All metric spaces are known to be normal. Let $\mathcal{Y}$ be a Banach space.

Theorem C.1. *The space of continuous, bounded functions $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$ from a normal topological space $\mathcal{X}$ to a Banach space $\mathcal{Y}$ with the norm $||f||_{\mathcal{C}_b(\mathcal{X})} = \sup\{||f(x)||_{\mathcal{Y}} : x \in \mathcal{X}\}$ is a Banach space.*

Proof: *By Lemma [105, Lemma 1.4.18], $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$ is a vector space on the real field and by [105, Corollary 1.7.7], the vector space is closed, implying that $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$ is a Banach space.* □

Theorem C.2. *Let $\mathcal{X}$ be a normal topological space. The dual space to $\mathcal{C}_b(\mathcal{X})$, $\mathcal{C}_b(\mathcal{X})^\star$ is isometric to the space of regular bounded and finitely additive measure on $\mathcal{B}(\mathcal{X})$. For any measure $\nu \in \mathcal{C}_b(\mathcal{X})^\star$, $f \in \mathcal{C}_b(\mathcal{X})$, $\nu(f) = \int_{\mathcal{X}} f(x) d\nu(x)$, $||\nu||_{\mathcal{C}_b(\mathcal{X})^\star} = \sup\{|\nu(f)| : ||f||_{\mathcal{C}_b(\mathcal{X})} \leq 1\}$ and $|\nu(f)| \leq ||\nu||_{\mathcal{C}_b(\mathcal{X})^\star} ||f||_{\mathcal{C}_b(\mathcal{X})}$.*

The above theorem is a restatement of [105, Theorem 2, IV.6.2]

Definition C.1. *(Regular bounded and finitely additive measure)*
On the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ a real valued measure ν is called,

1. *bounded, if for all $E \in \mathcal{B}(\mathcal{X})$, $|\nu(E)| < \infty$. Thus every finite additive measure is bounded.*

2. *finitely additive, if for any finite collection of disjoint sets $E_1, \ldots, E_n \in \mathcal{B}(\mathcal{X})$, $\nu(\cup E_i) = \sum_i E_i$. Thus any finite additive measure is bounded and finitely additive.*

3. *regular, if for every $\epsilon > 0$ and $E \in \mathcal{B}(\mathcal{X})$ there exists a closed set F and an open set G such that $F \subset E \subset G$ and for every $C \subset G \backslash F$, $C \in \mathcal{B}(\mathcal{X})$, $\mu(C) < \epsilon$. Thus all radon measures are regular.*

C.2 $L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ spaces

We refer the reader to [107, Chapter 9] for a more comprehensive overview of L^p spaces and the proofs of theorems below.

Definition C.2. *($L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ space)*
Let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be a measurable space and $\mathcal{Y}$ be a Banach space. For $p \in [1, \infty] \subset \mathbb{R}$ and a σ-additive nonnegative measure $\mu : \mathcal{B}(\mathcal{Z}) \to [0, \infty]$, a normed space of measurable functions from $\mathcal{Z}$ to $\mathcal{Y}$, called a $L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ space, is given by

$$L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu) = \{f : \mathcal{Z} \to \mathcal{Y} : \int_{\mathcal{Z}} ||f(\zeta)||^p_{\mathcal{Y}} d\mu(\zeta) < \infty\}$$

and

$$||f||_{L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)} = \left(\int_{\mathcal{Z}} ||f(\zeta)||^p_{\mathcal{Y}} d\mu(\zeta) \right)^{1/p}$$

Theorem C.3. *($L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ is a Banach space)*
For any nonnegative σ-additive measure μ, $(L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu), ||\cdot||_{L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)})$ is a complete, normed space, i.e. Banach space.

Proof: *See proof for [107, Theorem 9.6].* □

Note that the case for $p = \infty$ is excluded in the above theorem and the dual index for $p = \infty$, $p' = 1$ gives only a subset of the dual space in the case for $p = \infty$.

Theorem C.4. *(Dual to $L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$)*
For a σ-finite measure $\mu : \mathcal{B}(\mathcal{Z}) \to [0, \infty]$ and $p \in [1, \infty)$, the dual space of continuous linear functionals on $L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ is given by $L^{p'}(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ where p' is the dual index (see definition below) for p.

Proof: *See [107, Theorem 9.19]* □

Definition C.3. *(Dual index)*
The dual index p' for $p \in [1, \infty]$ is defined as the number $p' \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$. The dual index for $p = 1$ is $p' = \infty$ and vice versa.

Theorem C.5. *(Holder's inequality)*
For any measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$, $p \in [1, \infty]$ and p' the dual index to p,

$$\forall h \in L^{p'}(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu), f \in L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu), \qquad ||hf||_{L^1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)} \leq ||h||_{L^{p'}(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)} ||f||_{L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)}$$

Theorem C.6. *For a positive finite measure and $1 \leq p \leq q < \infty$, $L^q(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu) \subseteq L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ and*

$$\forall f \in L^q(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu), \qquad ||f||_{L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)} \leq (\mu(\mathcal{Z}))^{1/p - 1/q} ||f||_{L^q(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)}$$

Corollary C.1. *For a probability measure μ and $1 \leq p \leq q < \infty$,*

$$\forall f \in L^q(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu), \qquad ||f||_{L^p(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)} \leq ||f||_{L^q(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)}$$

C.3 Reproducing Kernel Hilbert Spaces

Let $\mathcal{Z}$ be a separable Banach space, $\mathcal{B}(\mathcal{Z})$ the Borel σ-algebra and μ be a Gaussian measure on $\mathcal{Z}$.

Definition C.4. *(Reproducing Kernel Hilbert Spaces (RKHS))*
Let $\mathcal{Z}$ be a separable Banach space and μ be a Gaussian measure on $\mathcal{Z}$. Let $J : \mathcal{Z}^\star \to \mathcal{Z}$ be the positive semidefinite linear operator given by

$$J(h) = \int_{\mathcal{Z}} \zeta h(\zeta) d\mu(\zeta) \tag{C.1}$$

and let $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$ be the Hilbert space induced by the semidefinite operator (as defined in Section A.6.1). The Hilbert space induced by J is called a Reproducing Kernel Hilbert Space (RKHS) or sometimes as the Cameron-Martin space for the Gaussian measure μ.

Note that the RKHS is often also defined as the Hilbert space on which the evaluation operator is continuous (i.e. bounded) and has a "reproducing property", [18]. The approach here follows from the literature on stochastic processes and Gaussian measures as presented in [26] or [108] and further properties such as the bounded evaluation operators are shown to be a consequence of the above definition.

Theorem C.7. *(Existence and uniqueness of RKHS in the Gaussian measure space)*
Let $\mathcal{Z}$ be a Banach space, $\mathcal{B}(\mathcal{Z})$ the Borel σ-algebra and μ be a Gaussian measure on $\mathcal{Z}$. Then there exists a unique positive semi-definite, self adjoint operator $J : \mathcal{Z}^\star \to \mathcal{Z}$ as defined in (C.1), such that $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$ is a continuously embedded Hilbert subspace of $\mathcal{Z}$ and $\forall h, g \in \mathcal{Z}^\star$, $\mathbb{E}[hg] = \langle Jh, Jg\rangle_{\mathcal{H}}$.

Proof: *First we show that the operator J defined by (C.1) is a bounded, self adjoint, positive semidefinite linear operator, i.e., $J \in \mathcal{S}^+(\mathcal{Z})$ and thus by Theorem A.4 it induces a Hilbert space $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$ with the inner product, $\langle x, y\rangle_{\mathcal{H}} = (J^{-1}x)(y) = (J^{-1}y)(x)$, and $\mathcal{H}$ is continuously embedded in $\mathcal{Z}$.*

Consider the map $J : \mathcal{Z}^\star \to \mathcal{Z}$ given by $J(h) = \int_{\mathcal{Z}} \zeta h(\zeta) d\mu(\zeta)$. It is easy to verify that J is linear, i.e., $J(h + g) = J(h) + J(g)$ (by linearity of integration).

To see that J is bounded, note that, $||J(h)||_{\mathcal{Z}} = ||\int_{\mathcal{Z}} \zeta h(\zeta)\mu(d\zeta)||_{\mathcal{Z}} \leq \int_{\mathcal{Z}} ||\zeta h(\zeta)||_{\mathcal{Z}}\mu(d\zeta) = \int_{\mathcal{Z}} ||\zeta||_{\mathcal{Z}}|h(\zeta)|\mu(d\zeta) \leq \int_{\mathcal{Z}} ||h||_{\mathcal{Z}^\star}||\zeta||^2_{\mathcal{Z}}\mu(d\zeta) = \left(\int_{\mathcal{Z}} ||\zeta||^2_{\mathcal{Z}}\mu(d\zeta)\right) ||h||_{\mathcal{Z}^\star}$. Note that by Corollary B.1 $\int_{\mathcal{Z}} ||\zeta||^2_{\mathcal{Z}}\mu(d\zeta) < \infty$. Thus there exists a constant $\kappa = \int_{\mathcal{Z}} ||\zeta||^2_{\mathcal{Z}}\mu(d\zeta)$ in $(0, \infty)$ such that $\forall h \in \mathcal{Z}^\star$, $||J(h)||_{\mathcal{Z}} \leq \kappa||h||_{\mathcal{Z}^\star}$ implying J is bounded.

Further now since $||\int_{\mathcal{Z}} \zeta h(\zeta)\mu(d\zeta)||_{\mathcal{Z}} < \infty$, by Fubini's theorem we have for any $g \in \mathcal{Z}^\star$, $g(Jh) = g(\int_{\mathcal{Z}} \zeta h(\zeta)d\mu(\zeta)) = \int_{\mathcal{Z}} g(\zeta)h(\zeta)d\mu(\zeta) = h(Jg)$. Thus J is self adjoint and we have for all $h \in \mathcal{Z}^\star$, $h(Jh) = \int_{\mathcal{Z}} h(\zeta)^2 d\mu(\zeta) \geq 0$ implying $J \geq 0$.

Thus we have shown that J as defined by (C.1) belongs to $\mathcal{S}^+(\mathcal{Z})$ and by Theorems A.3 and A.4, the Hilbert space $\mathcal{H}$ induced by J is a continuously embedded Hilbert subspace of $\mathcal{Z}$. Further, by Theorem A.5, $\mathbb{E}[hg] = \int_{\mathcal{Z}} h(\zeta)g(\zeta)d\mu(\zeta) = g(Jh) = \langle Jh, Jg\rangle_{\mathcal{H}}$.

To show that J is the unique operator satisfying $\forall h, g \in \mathcal{Z}^\star$, $\mathbb{E}[hg] = \langle Jh, Jg\rangle_{\mathcal{H}}$, we proceed by contradiction. Let there exist another operator $J' \in \mathcal{S}^+(\mathcal{Z})$ such that $\mathbb{E}[hg] = \langle J'h, J'g\rangle_{\mathcal{H}}$, then we have $\mathbb{E}[hg] = h(Jg) = h(J'g)$. Thus $\forall h, g \in \mathcal{Z}^\star$ $h((J - J')g) = 0$ implying $\forall g \in \mathcal{Z}^\star$,

$(J - J')g = 0$ implying $J = J'$. Thus J given by (C.1) is the unique operator in $\mathcal{S}^+(\mathcal{Z})$ satisfying $\mathbb{E}[hg] = \langle Jh, Jg \rangle_{\mathcal{H}}$. □

Lemma C.1. *The following properties hold for $J(h) = \int_{\mathcal{Z}} \zeta h(\zeta) d\mu(\zeta)$,*

1. $Jh = 0 \iff h \overset{\mu-a.e.}{=} 0$
2. $f_1 = f_2 \iff f_1 = Jh_1,\ f_2 = Jh_2$ *and* $h_1 \overset{\mu-a.e.}{=} h_2$
3. $||Jh||_{\mathcal{H}} = ||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)}$
4. $||Jh||_{\mathcal{Z}} \leq ||J||_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||h||_{\mathcal{Z}^\star}$ *with* $||J||_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} = \int_{\mathcal{Z}} ||\zeta||^2_{\mathcal{Z}} d\mu(\zeta)$
5. $||J||^{-1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)} = ||J||^{-1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||Jh||_{\mathcal{H}} \leq ||h||_{\mathcal{Z}^\star} \leq ||Jh||_{\mathcal{H}} = ||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)}$

Proof: *For the first statement note that $h \overset{\mu-a.e.}{=} 0$ implies $||f||_{\mathcal{H}} = ||Jh||_{\mathcal{H}} = h(Jh)^{1/2} = (\int_{\mathcal{Z}} h(\zeta)^2 d\mu(\zeta))^{1/2} = 0 \implies Jh = 0$. On the other hand, $J(h) = 0$ implies $||Jh||_{\mathcal{H}} = 0$, i.e., $h(Jh) = \int_{\mathcal{Z}} h(\zeta)^2 d\mu(\zeta) = 0 \implies h \overset{\mu-a.e.}{=} 0$. The second statement follows from the first, by considering $\tilde{h} = h_1 - h_2$ and noting that $\tilde{h} \overset{\mu-a.e.}{=} 0$ implies $h_1 \overset{\mu-a.e.}{=} h_2$. The third statement follows from the inner product definition, $||Jh||_{\mathcal{H}} = (h(Jh))^{1/2} = \left(\int_{\mathcal{Z}} h(\zeta)^2 d\mu(\zeta)\right)^{1/2} = ||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)}$. The fourth statement was already shown in Theorem C.7, where J was shown to be bounded.*

For the final statement, using Cauchy-Schwarz inequality, note that $|h(\zeta)| \leq ||h||_{\mathcal{Z}^\star} ||\zeta||_{\mathcal{Z}}$. Then $||Jh||_{\mathcal{H}} = h(Jh)^{1/2} = \left(\int_{\mathcal{Z}} h(\zeta)^2 d\mu(\zeta)\right)^{1/2} \leq \left(\int_{\mathcal{Z}} ||h||^2_{\mathcal{Z}^\star} ||\zeta||^2_{\mathcal{Z}} d\mu(\zeta)\right)^{1/2} = \left(\int_{\mathcal{Z}} ||\zeta||^2_{\mathcal{Z}} d\mu(\zeta)\right)^{1/2} ||h||_{\mathcal{Z}^\star} = ||J||^{1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||h||_{\mathcal{Z}^\star}$, i.e. $||J||^{-1/2}_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}} ||Jh||_{\mathcal{H}} \leq ||h||_{\mathcal{Z}^\star}$.

For the other side of the inequality, note by Holder's inequality that $||Jh||_{\mathcal{H}} = \left(\int_{\mathcal{Z}} h(\zeta)^2 d\mu(\zeta)\right)^{1/2} = \left(||h^2||_{L^1(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)}\right)^{1/2} \leq \left(||h||^2_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)}\right)^{1/2} = ||h||_{L^2(\mathcal{Z},\mathcal{B}(\mathcal{Z}),\mu)} = ||Jh||_{\mathcal{H}}$. □

The reproducing kernel Hilbert space $\mathcal{H}$ is a continuously embedded subspace in $\mathcal{Z}$. When the subspace $\mathcal{H}$ is dense in $\mathcal{Z}$, for the natural inclusion $i : \mathcal{H} \to \mathcal{Z}$, $i(f) = f$ along with the spaces, $(i, \mathcal{H}, \mathcal{Z})$ forms an abstract Wiener space [26, Theorem 3.9.6]. [109, Theorem 7] shows that an RKHS space induced on a separable Banach space by a non-degenerate Gaussian measure forms an Abstract Wiener space.

C.3.1 Kernels, Adjoints and Covariance operators in RKHS

Finally we would like to establish the relation of the RKHS as defined above with the commonly understood notion of RKHS in terms of positive definite kernel functions as defined in [18] and clarify the relation of J and the kernel to the covariance operator of the measure μ.

For the relation of J to the kernel functions we first establish the relation of adjoints for linear operators acting on the RKHS above to J.

Theorem C.8. *(Adjoints for Hilbert space-valued operators from J)*
Let $\mathcal{Y}$ be a separable Hilbert space, $\mathcal{Z}$ a separable Banach space. Let $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$ be a dense RKHS subspace of $\mathcal{Z}$, induced by $J \in \mathcal{S}^+(\mathcal{Z})$ and $L : \mathcal{H} \to \mathcal{Y}$ be a bounded linear operator with $\mathrm{dom}(L) = \mathcal{H}$*. Let $L' : \mathcal{Z} \to \mathcal{Y}$ be the unique, bounded linear extension of L to $\mathcal{Z}$ as justified by [?, Theorem 3]. For every $y \in \mathcal{Y}$, let $M_y : \mathcal{Z} \to \mathbb{R}$ denote the bounded linear functional given by $M_y(\zeta) = \langle L'\zeta, y\rangle_{\mathcal{Y}}$. Then the adjoint $L^\star$ is given by*

$$\forall y \in \mathcal{Y}, \quad L^\star y = J(M_y) = \int_{\mathcal{Z}} \zeta M_y(\zeta) d\mu(\zeta) \tag{C.2}$$

Proof: *By definition of the adjoint we have for all $y \in \mathcal{Y}$ and $f \in \mathrm{dom}(L) = \mathcal{H}$, $M_y(f) = \langle Lf, y\rangle_{\mathcal{Y}} = \langle f, L^\star y\rangle_{\mathcal{H}}$. Since $M_y \in \mathcal{Z}^\star$, by Theorem A.5, note that for any $f \in \mathcal{H}$, $M_y(f) = \langle JM_y, f\rangle_{\mathcal{H}}$. Thus we have $\forall f \in \mathcal{H}$, $\langle L^\star y - J(M_y), f\rangle_{\mathcal{H}} = 0$, implying $L^\star y = J(M_y) = \int_{\mathcal{Z}} \zeta M_y(\zeta) d\mu(\zeta) = \int_{\mathcal{Z}} \zeta \langle L'\zeta, y\rangle_{\mathcal{Y}} d\mu(\zeta)$.* □

Below we show the application of Theorem C.8 for a few commonly used dense embeddings of RKHS in Banach spaces.

C.3.2 Kernel for evaluation operator L_x on RKHS embedding in $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$

Theorem C.9. *(RKHS Kernel for $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$)*
Let $\mathcal{Y}$ be a Hilbert space, $\mathcal{Z} = \mathcal{C}_b(\mathcal{X}, \mathcal{Y})$ be the Banach space of $\mathcal{Y}$-valued continuous and bounded functions with a metric space domain $\mathcal{X}$. Let μ be a non-degenerate Gaussian measure on $\mathcal{B}(\mathcal{Z})$. Let $J \in \mathcal{S}^+(\mathcal{Z})$ be the operator given by (C.1) inducing an RKHS, $\mathcal{H} = \overline{J(\mathcal{Z}^\star)}$. Let $\{L_x : \mathcal{Z} \to \mathcal{Y} : x \in \mathcal{X}\}$ be the collection of bounded evaluation operators such that $L_x f = f(x)$. Then the following properties hold,

1. *$L_x|\mathcal{H} : \mathcal{H} \to \mathcal{Y}$ is a bounded linear evaluation operator with domain restricted to $\mathcal{H}$ such that $\forall f \in \mathcal{H}$, $(L_x|\mathcal{H})f = L_x f = f(x)$.*
2. *There exists an unique operator-valued function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Y},\mathcal{Y}}$ such that $(L_x|\mathcal{H})^\star = K(x, \cdot)$, i.e., $\forall y \in \mathcal{Y}$, $(L_x|\mathcal{H})^\star y = K(x, \cdot)y$.*

Proof: *Note that for all $f \in \mathcal{H} \subseteq \mathcal{C}_b(\mathcal{X}, \mathcal{Y})$, we have $||(L_x|\mathcal{H})f||_{\mathcal{Y}} = ||L_x f||_{\mathcal{Y}} \leq ||f||_{\mathcal{Z}}$ and from Lemma C.1-4 and 5, $||f||_{\mathcal{Z}} \leq (||J||_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}})||f||_{\mathcal{H}}$. Thus for all $f \in \mathcal{H}$, $||(L_x|\mathcal{H})f||_{\mathcal{Y}} \leq (||J||_{\mathcal{L}_{\mathcal{Z}^\star,\mathcal{Z}}})||f||_{\mathcal{H}}$, implying the linear operator $L_x|\mathcal{H}$ is bounded on $\mathcal{H}$.*

Now for $(L_x|\mathcal{H})^\star$ from (C.2), note that $(L_x|\mathcal{H})^\star y = J(M_y) = \int_{\mathcal{Z}} \zeta \langle L_x\zeta, y\rangle_{\mathcal{Y}} d\mu(\zeta)$ is a function in $\mathcal{H}$, then for any $x, s \in \mathcal{X}$, we can define $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Y},\mathcal{Y}}$ as $K(x, s) = \int_{\mathcal{Z}} L_s\zeta \langle L_x\zeta, \cdot\rangle_{\mathcal{Y}} d\mu(\zeta)$. □

The following properties for the kernel function K are then straightforward to verify from the definition of K,

Lemma C.2. *(Properties of the evaluation kernel on $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$)*

1. *$K(x, x) > 0$ (i.e. $K(x, x)$ is a positive definite operator)*

2. *$K(x,s) = K(s,x)^\star$*
3. *$\int_\mathcal{X}\int_\mathcal{X}\langle f(s), K(x,s)f(x)\rangle_\mathcal{Y} d\nu(x)d\nu(s) \geq 0$ for all $f \in L^2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$ (Mercer's property)*
4. *$\langle K(x,s)y, z\rangle_\mathcal{Y} = \langle K(x,\cdot)y, K(s,\cdot)z\rangle_\mathcal{H}$*
5. *$||(L_x|\mathcal{H})^\star||_{\mathcal{L}_{\mathcal{Y},\mathcal{H}}} = ||K(x,\cdot)||_{\mathcal{L}_{\mathcal{Y},\mathcal{H}}} = ||K(x,x)||^{1/2}_{\mathcal{L}_{\mathcal{Y},\mathcal{Y}}}$*
6. *$||K(x,s)||_{\mathcal{L}_{\mathcal{Y},\mathcal{Y}}} \leq ||K(x,x)||^{1/2}_{\mathcal{L}_{\mathcal{Y},\mathcal{Y}}}||K(s,s)||^{1/2}_{\mathcal{L}_{\mathcal{Y},\mathcal{Y}}}$*
7. *$\langle L_x f, z\rangle_\mathcal{Y} = \langle f, K(x,\cdot)z\rangle_\mathcal{H}$ (The reproducing property)*

Proof: *For the first and second statement, note that, for any $x, s \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$ we have $\langle y_1, K(x,s)y_2\rangle_\mathcal{Y} = \langle y_1, \int_\mathcal{Z} L_s\zeta\langle L_x\zeta, y_2\rangle_\mathcal{Y} d\mu(\zeta)\rangle_\mathcal{Y} = \int_\mathcal{Z}\langle L_s\zeta, y_1\rangle_\mathcal{Y}\langle L_x\zeta, y_2\rangle_\mathcal{Y} d\mu(\zeta) = \langle y_2, \int_\mathcal{Z} L_x\zeta\langle L_s\zeta, y_1\rangle_\mathcal{Y} d\mu(\zeta)\rangle_\mathcal{Y} = \langle y_2, K(s,x)y_1\rangle_\mathcal{Y}$. This shows that $K(x,s)^\star = K(s,x)$ (second statement) and that for any $y \neq 0 \in \mathcal{Y}$, $\langle y, K(x,x)y\rangle_\mathcal{Y} = \int_\mathcal{Z}\langle L_x\zeta, y\rangle^2_\mathcal{Y} d\mu(\zeta) > 0$ (first statement).*

For the third statement (Mercer's property), $\int_\mathcal{X}\int_\mathcal{X}\langle f(s), K(x,s)f(x)\rangle_\mathcal{Y} d\nu(x)d\nu(s) = \int_\mathcal{X}\int_\mathcal{X}\int_\mathcal{Z}\langle L_s\zeta, f(s)\rangle_\mathcal{Y}\langle L_x\zeta, f(x)\rangle_\mathcal{Y} d\mu(\zeta)d\nu(x)d\nu(s) = \int_\mathcal{Z}\left(\int_\mathcal{X}\langle L_x\zeta, f(x)\rangle_\mathcal{Y} d\nu(x)\right)^2 d\mu(\zeta) \geq 0$

The fourth to seventh statement follows directly from the definition of the kernel section $K(x,\cdot)$ being the adjoint for the evaluation operator L_x. □

Another question that arises is for what condition does a function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Y},\mathcal{Y}}$ correspond to kernel for an evaluation operator densely defined in an RKHS. This was addressed by Mercer's theorem [110] for real valued function spaces and by an extension to $\mathbb{C}^n$-valued functions in [111].

Theorem C.10. *(Second moment (Covariance) operator J)*
$h(Jg) = \mathbb{E}[hg]$ computes the second moment for any measure μ. For evaluation operators L_x, L_s, $\mathbb{E}[L_x f, L_s f] = K(x,s)$ gives the second moment for vectors in $\mathcal{Y}$. For zero mean Gaussian measures this is the same as the covariance operator and thus the kernel $K(x,s)$ gives the covariance for the evaluated vectors in $\mathcal{Y}$ for a zero mean Gaussian process.

C.4 The space of signed measures: $\mathcal{M}_\sigma(\mathcal{X}, \mathcal{B}_\mathcal{X})$

Below is a summary of material presented in [30, Chapter 14]. We will call the space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ a σ-measurable space if $\mathcal{B}_\mathcal{X}$ is a σ-algebra on $\mathcal{X}$.

Definition C.5. *(Signed measure)*
On a σ-measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, a mapping $\nu : \mathcal{B}_\mathcal{X} \to [-\infty, \infty]$ that satisfies the additivity property for any countable collection of disjoint measurable sets $\{A_n \in \mathcal{B}_\mathcal{X} : n \in \mathbb{N}, \forall i \neq j, A_i \cap A_j = \varnothing\}$, i.e., $\nu(\cup_n A_n) = \sum_n \nu(A_n)$ and $\nu(\varnothing) = 0$ is a called σ-additive signed measure, or simply as a signed measure. ($-\infty$ and ∞ are included in the range). The space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \nu)$ is called the signed measure space.

Definition C.6. *(ν-positive and negative subsets)*
On a signed measure space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \nu)$, a measurable set $A \in \mathcal{B}_\mathcal{X}$ is called ν-positive (negative) if for all $\mathcal{B}_\mathcal{X}$-measurable subsets $B \subseteq A$, $\nu(B) \geq 0$ ($\nu(B) \leq 0$). The set A is called strongly ν-positive (negative) if for all measurable subsets $B \subseteq A$, $\nu(B) > 0$ ($\nu(B) < 0$).

For a signed measure space the [102, Lemma 27.1] tells us that for any $A \in \mathcal{B}_\mathcal{X}$ such that $\nu(A) > 0$ ($\nu(A) < 0$) there must exist a subset $B \in \mathcal{B}_\mathcal{X}$, $B \subseteq A$ such that B is strongly ν-positive (negative). From this lemma the Hahn decomposition theorem [102, Theorem 27.2] follows.

Theorem C.11. *(Hahn decomposition)*
In a signed measure space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \nu)$ there exist disjoint ν-positive and ν-negative sets $\mathcal{X}^+$ and $\mathcal{X}^-$ such that $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^-$.

Proof: *For $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \nu)$ such that for all $A \in \mathcal{B}_\mathcal{X}$, $\nu(A) = 0$, the result holds trivially since any subset $\mathcal{X}^+ \subset \mathcal{X}$ and $\mathcal{X}^- = \mathcal{X} \backslash \mathcal{X}^+$ gives such a decomposition. Thus without loss of generality we assume that there exists at-least one $A \in \mathcal{B}_\mathcal{X}$ such that $\nu(A) \geq 0$ or $\nu(A) \leq 0$. Further if for all $A \in \mathcal{B}_\mathcal{X}$, $\nu(A) \geq 0$ (or $\forall A \in \mathcal{B}_\mathcal{X}$, $\nu(A) \leq 0$) the result is trivial $\mathcal{X}^+ = \mathcal{X}$ and $\mathcal{X}^- = \varnothing$ (or $\mathcal{X}^- = \mathcal{X}$ and $\mathcal{X}^+ = \varnothing$).*

Thus we assume there exist some $A, B \in \mathcal{B}_\mathcal{X}$ such that $\nu(A) > 0$ and $\nu(B) < 0$. Then by [102, Lemma 27.1] we know there must exists at-least one strongly ν-positive $A^+ \subseteq A$. Let $P = \{A \in \mathcal{B}_\mathcal{X} : A \text{ is strongly } \nu\text{-positive}\}$ and $l = \sup\{\nu(A) : A \in P\}$. For any countable sequence $\{A_n \in P : n \in \mathbb{N}\}$ such that $\lim_{n\to\infty} \nu(A_n) = l$, we have $\mathcal{X}^+ = \cup_{n\in\mathbb{N}} A_n$ which is strongly ν-positive. Then $\mathcal{X}^- = \mathcal{X} \backslash \mathcal{X}^+$ must be ν-negative. To see this note that if $\mathcal{X}^-$ is not ν-negative then there would exist a subset $B \in \mathcal{X}^-$ such that $\nu(B) > 0$, but then $\nu(\mathcal{X}^+ \cup B) = \nu(\mathcal{X}^+) + \nu(B) > l$, but this contradicts the requirement that l is the supremum. Thus $\mathcal{X}^-$ and $\mathcal{X}^+$ gives the required decomposition. □

As a corollary, for a signed measure space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \nu)$ such that there exist $A, B \in \mathcal{B}_\mathcal{X}$ with $\nu(A) > 0$ and $\nu(B) < 0$, we have a unique decomposition $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^- \cup \mathcal{X}^0$ with $\mathcal{X}^+$ strongly ν-positive, $\mathcal{X}^-$ strongly ν-negative and $\mathcal{X}^0$ a ν-null set (i.e., $\nu(\mathcal{X}^0) = 0$). From the Hahn decomposition, the Jordan decomposition of measures follows.

Lemma C.3. *(Jordan decomposition of measures)*
For a signed measure ν on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ there exists a decomposition $\nu = \nu^+ - \nu^-$ where both ν^+ and ν^- are nonnegative σ-additive measures and at-least one of them is a finite measure.

Proof: *Let $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^- \cup \mathcal{X}^0$ be the unique Hahn decomposition for $\mathcal{X}$. Define the measures $\nu^+(A) = \nu(A \cap \mathcal{X}^+)$, $\nu^-(A) = \nu(A \cap \mathcal{X}^-)$. Noting that $\nu(A \cap \mathcal{X}^0) = 0$, we have $\nu^+(A) + \nu^-(A) = \nu(A \cap \mathcal{X}^+) + \nu(A \cap \mathcal{X}^-) + \nu(A \cap \mathcal{X}^0) = \nu((A \cap \mathcal{X}^+) \cup (A \cap \mathcal{X}^-) \cup (A \cap \mathcal{X}^0)) = \nu(A)$.* □

A norm on $M_\sigma(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is given by $||\nu|| = \sup\{|\nu|(A) : A \in \mathcal{B}_\mathcal{X}\}$.

Vector valued analogues of the same can be found in [32]

Appendix D

Stochastic processes

Let $\mathcal{Z}$ be a separable Banach space of functions from a set $\mathcal{X}$ to a Banach space $\mathcal{Y}$. Let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \nu)$ be a probability measure on $\mathcal{Z}$. We will call a $(\mathcal{B}(\mathcal{Z})|\mathcal{B}(\mathcal{Z}))$-measurable function $X : \mathcal{Z} \to \mathcal{Z}$ a stochastic process. Such a stochastic process is also referred to as a random field sometimes in literature. In particular the identity mapping $X_\nu(\zeta) = \zeta$ is a stochastic process characterized by the measure ν. For any stochastic process $X : \mathcal{Z} \to \mathcal{Z}$ in general, the measure characterizing X is the push forward measure $\nu \circ X^{-1} : \mathcal{B}(\mathcal{Z}) \to [0, 1]$.

In particular this can be applied to generate a space of Gaussian processes when ν is a Gaussian measure on the space of functions $\mathcal{Z}$ and X is restricted to a space of bounded affine functions, as shown below.

D.1 A Banach space of Gaussian processes: $\mathcal{Z}_\mu$

Definition D.1. *(Gaussian process)*
Let $\mathcal{Z}$ be a Banach space of functions from $\mathcal{X}$ to $\mathcal{Y}$ and $\mu : \mathcal{B}(\mathcal{Z}) \to [0, 1]$ be a Gaussian probability measure on $\mathcal{Z}$. A $(\mathcal{B}(\mathcal{Z})|\mathcal{B}(\mathcal{Z}))$-measurable function $X : \mathcal{Z} \to \mathcal{Z}$ is called a Gaussian process if the push forward measure $\mu \circ X^{-1} : \mathcal{B}(\mathcal{Z}) \to [0, 1]$ is a Gaussian measure on $\mathcal{Z}$.

Lemma D.1. *(Affine functions on $\mathcal{Z}$ are Gaussian processes)*
Let $\mathcal{Z}$ be a Banach space of functions from $\mathcal{X}$ to $\mathcal{Y}$ and $\mu : \mathcal{B}(\mathcal{Z}) \to [0, 1]$ be a Gaussian probability measure on $\mathcal{Z}$. Let $A \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ be a bounded linear operator and $b \in \mathcal{Z}$ be a given vector in $\mathcal{Z}$. Then the measurable function $X(\zeta) = A\zeta + b$ is a Gaussian process.

Proof: *The pre-image $X^{-1}(z) = A^{-1}z - b$ is given by the preimage map A^{-1} (A is not necessarily invertible, the notation is used for a preimage map here). By definition of Gaussian measures on $\mathcal{Z}$, $\mu \circ X^{-1}$ is a Gaussian measure if, and only if, for any functional $h \in \mathcal{Z}^\star$, $\mu \circ X^{-1} \circ h^{-1}$ is a Gaussian measure on $\mathbb{R}$. Further for any $\vartheta \in \mathcal{B}(\mathcal{Z})$, $\vartheta_b = \vartheta - b$ is a Borel measurable set, i.e., $\vartheta_b \in \mathcal{B}(\mathcal{Z})$. Under a change of variable $\mathcal{Z}_b = \mathcal{Z} - b$, if μ is Gaussian on $\mathcal{Z}$, $\mu_b(\cdot) = \mu(\cdot - b)$ is a Gaussian measure on $\mathcal{Z}_b$. For each $h \in \mathcal{Z}^\star$, $h_b = h \circ X - h(b) = h \circ A$ belongs to $\mathcal{Z}_b^\star$ and $h_b^{-1}z = A^{-1}h^{-1}(z)$.*

Now, note that $\mu \circ X^{-1} \circ h^{-1}(z) = \mu(A^{-1}h^{-1}z - b) = \mu_b(A^{-1}h^{-1}z) = \mu_b(h_b^{-1}z)$. Thus $\mu \circ X^{-1} \circ h^{-1}(z) = \mu_b(h_b^{-1}(z))$ is a Gaussian measure on $\mathbb{R}$, since μ_b is a Gaussian measure on $\mathcal{Z}_b$ and

$h_b \in \mathcal{Z}_b^\star$, implying X is a Gaussian process. □

The Feldman-Hajek theorem [26, Chapter 6] states that two Gaussian measures on $\mathcal{Z}$ can either be equivalent or mutually singular and nothing else. Not all Gaussian measures are equivalent on infinite dimensional $\mathcal{Z}$. Thus affine transformations as used above do not in general lead to equivalent Gaussian measures, except under certain conditions on A (see [26, Chapter 6]). So no claim on equivalence of the push forward Gaussian measures are made, in general.

The general Banach space of measurable functions $X : \mathcal{Z} \to \mathcal{Z}$ is given by the notion of a Lebesgue-Bochner space. We show below that a set of bounded affine transformations on $\mathcal{Z}$ provides an example of an L^1 Lebesgue-Bochner space of Gaussian processes.

Lemma D.2. *(A vector space of Gaussian processes)*
Let $A_1, A_2 \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$, $b_1, b_2 \in \mathcal{Z}$ and $X_1(\zeta) = A_1\zeta + b_1$, $X_2(\zeta) = A_2\zeta + b_2$ be two Gaussian processes on $\mathcal{Z}$. Then for any scalar $\lambda_1, \lambda_2 \in \mathbb{R}$, $\lambda_1 X_1 + \lambda_2 X_2$ is a Gaussian process on $\mathcal{Z}$. The set $\mathcal{Z}_\mu = \{A\zeta + b : A \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}, b \in \mathcal{Z}\}$ is a vector space of Gaussian processes on $\mathcal{Z}$.

Proof: *For any $\lambda_1, \lambda_2 \in \mathbb{R}$, $A_1, , A_2 \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ and $b_1, b_2 \in \mathcal{Z}$, $A = \lambda_1 A_1 + \lambda_2 A_2 \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ and $b = \lambda_1 b_1 + \lambda_2 b_2 \in \mathcal{Z}$. Thus $X = \lambda_1 X_1 + \lambda_2 X_2 = A\zeta + b$ is a Gaussian process by Lemma D.1.* □

Lemma D.3. *(A norm on $\mathcal{Z}_\mu$)*
Let $\mathcal{Z}_\mu = \{A\zeta + b : A \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}, b \in \mathcal{Z}\}$ is a vector space of Gaussian processes on the Gaussian probability measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$. A norm on $\mathcal{Z}_\mu$ is given as $||X||_{\mathcal{Z}_\mu} = \int_{\mathcal{Z}} ||X(\zeta)||_{\mathcal{Z}} d\mu(\zeta)$.

Proof: *The properties of the norm for $||\cdot||_{\mathcal{Z}_\mu}$ can be verified as follows. For any $X_1, X_2 \in \mathcal{Z}_\mu$, $||X_1 + X_2||_{\mathcal{Z}_\mu} = \int_{\mathcal{Z}} ||X_1(\zeta) + X_2(\zeta)||_{\mathcal{Z}} d\mu(\zeta) \leq \int_{\mathcal{Z}} ||X_1(\zeta)||_{\mathcal{Z}} d\mu(\zeta) + \int_{\mathcal{Z}} ||X_2(\zeta)||_{\mathcal{Z}} d\mu(\zeta) = ||X_1||_{\mathcal{Z}_\mu} + ||X_2||_{\mathcal{Z}_\mu}$. For any $a \in \mathbb{R}$, $||aX||_{\mathcal{Z}_\mu} = a||X||_{\mathcal{Z}_\mu}$. Further $||X||_{\mathcal{Z}_\mu} = 0$ implies $\int_{\mathcal{Z}} ||X(\zeta)||_{\mathcal{Z}} d\mu(\zeta) = 0$ implying $X(\zeta) = 0$, μ-almost everywhere.* □

Lemma D.4. *(A Banach space of Gaussian processes: $\mathcal{Z}_\mu$)*
Let $\mathcal{Z}$ be a separable Banach space and $\mu : \mathcal{B}(\mathcal{Z}) \to [0, 1]$ be a Gaussian probability measure on $\mathcal{Z}$. Then the space $(\mathcal{Z}_\mu, ||\cdot||_{\mathcal{Z}_\mu})$ is a Banach space.

Proof: *From Lemma D.2 and D.3, $(\mathcal{Z}_\mu, ||\cdot||_{\mathcal{Z}_\mu})$ is a normed vector space. To check completeness of the normed space, for any Cauchy sequence $\{X_n : n \in \mathbb{N}\}$ in $\mathcal{Z}_\mu$, $||X_n - X_m||_{\mathcal{Z}_\mu} \to 0$ implies $\int_{\mathcal{Z}} ||(A_n - A_m)\zeta + b_n - b_m||_{\mathcal{Z}} d\mu(\zeta) \to 0$ implies $||(A_n - A_m)\zeta + b_n - b_m||_{\mathcal{Z}} \to 0$ for , μ-almost all ζ. Also $\lim_{n,m\to\infty} ||(A_n - A_m)\zeta + b_n - b_m||_{\mathcal{Z}} = \lim_{n,m\to\infty} || - (A_n - A_m)\zeta + b_n - b_m||_{\mathcal{Z}}$ and $|| - (A_n - A_m)\zeta + b_n - b_m||_{\mathcal{Z}} \geq \Bigg| ||(A_n - A_m)\zeta||_{\mathcal{Z}} - ||(b_n - b_m)||_{\mathcal{Z}} \Bigg| \to 0$ for μ-almost all ζ, implying $||A_n - A_m||_{\mathcal{L}_{\mathcal{Z},\mathcal{Z}}} \to 0$ and $||b_n - b_m||_{\mathcal{Z}} \to 0$. Thus for every Cauchy sequence in $\mathcal{Z}_\mu$ there must exist corresponding Cauchy sequences $\{A_n \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}\}$ and $\{b_n \in \mathcal{Z}\}$. Since $\mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ and $\mathcal{Z}$ are complete spaces, X_n converges to a Gaussian process $X(\zeta) = A\zeta + b$ with $A = \lim A_n \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ and $b = \lim b_n \in \mathcal{Z}$, i.e. $X \in \mathcal{Z}_\mu$. Thus $(\mathcal{Z}_\mu, ||\cdot||_{\mathcal{Z}_\mu})$ is a complete normed vector space, i.e. Banach space.* □

Theorem D.1. *(Dual to $\mathcal{Z}_\mu$)*
The dual space to $\mathcal{Z}_\mu$ is given by the space $\mathcal{Z}_\mu^\star = \{h : \mathcal{Z} \to \mathcal{Z}^\star : \forall X \in \mathcal{Z}_\mu, \int_{\mathcal{Z}} |h(\zeta)X(\zeta)| d\mu(\zeta) < \infty\}$, with the dual action given as $\mathbb{E}_\mu[hX] = \int_{\mathcal{Z}} h(\zeta)X(\zeta) d\mu(\zeta)$. The dual norm is given as $||h||_{\mathcal{Z}_\mu} = \sup\{|\mathbb{E}_\mu[hX]| : ||X||_{\mathcal{Z}_\mu} = 1\}$.

Proof: *See proof for [112, Theorem 42].*

□

D.2 An RKHS of Gaussian processes: $\mathcal{H}_{\mu,\nu}$

Consider a Banach space of Gaussian processes $\mathcal{Z}_\mu$ as given by Section D.1. For a Gaussian measure $\nu : \mathcal{B}(\mathcal{Z}_\mu) \to [0,1]$, we can then define an RKHS $\mathcal{H}_{\mu,\nu} = \overline{J(\mathcal{Z}_\mu)}$ with $J : \mathcal{Z}_\mu^\star \to \mathcal{Z}_\mu$ given as $J(h) = \int_{\mathcal{Z}_\mu} \zeta h(\zeta) d\nu(\zeta)$ as was done for a general Banach space in Theorem C.7.

D.2.1 Kernels for RKHS of bounded, continuous stochastic processes

Let $\mathcal{Z} = \mathcal{C}_b(\mathcal{X})$ and $\mathcal{Z}_\mu$ be the Banach space of Gaussian processes defined on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ as per Lemma D.4. For a Gaussian measure ν on $\mathcal{Z}_\mu$, let $\mathcal{H}_{\mu,\nu}$ be the RKHS of Gaussian processes in $\mathcal{Z}_\mu$.

The evaluation operator $L_x : \mathcal{Z} \to \mathcal{Y}$, $L_x f = f(x)$ is a bounded linear operator for $\mathcal{Z} = \mathcal{C}_b(\mathcal{X})$. The push forward measure $\mu \circ L_x^{-1} : \mathcal{B}(\mathcal{Y}) \to [0,1]$ is a Gaussian measure on $\mathcal{Y}$ and $\mathcal{Y}_{\mu \circ L_x^{-1}}$ (denoted as $\mathcal{Y}_\mu$ for notational convenience) is the corresponding Banach space of $\mathcal{Y}$-valued Gaussian measurable functions. $L_x : \mathcal{Z}_\mu \to \mathcal{Y}_\mu$, given as $L_x X(\zeta) = X(L_x \zeta)$ is then a bounded linear operator as shown below.

Lemma D.5. *($L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_\mu$ is a bounded linear operator)*
Let $\mathcal{H}$ be the RKHS induced by μ on $\mathcal{Z}$. $||L_x X||_{\mathcal{Y}_\mu} \le ||L_x||_{\mathcal{L}_{\mathcal{H}_{\mu,\nu},\mathcal{Y}_\mu}} ||X||_{\mathcal{Z}_\mu}$ and $||L_x||_{\mathcal{L}_{\mathcal{H}_{\mu,\nu},\mathcal{Y}_\mu}} = ||L_x||_{\mathcal{L}_{\mathcal{H},\mathcal{Y}}}$, i.e. $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_\mu$ is a bounded linear operator.

Proof: $||L_x X||_{\mathcal{Y}_\mu} = \int_{\mathcal{Y}} ||X(y)||_{\mathcal{Y}} d(\mu \circ L_x^{-1}(y)) = \int_{\mathcal{Z}} ||L_x X(\zeta)||_{\mathcal{Y}} d\mu(\zeta) \le ||L_x||_{\mathcal{L}_{\mathcal{H},\mathcal{Y}}} \int_{\mathcal{Z}} ||X(\zeta)||_{\mathcal{Z}} d\mu(\zeta) = ||L_x||_{\mathcal{L}_{\mathcal{H},\mathcal{Y}}} ||X||_{\mathcal{Z}_\mu}$. *From Theorem C.9, $||L_x||_{\mathcal{L}_{\mathcal{H},\mathcal{Y}}} < \infty$ and thus we have $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_\mu$ to be bounded.* □

Theorem D.2. *(Kernel for $\mathcal{H}_{\nu,\mu}$)*
For a Hilbert space $\mathcal{Y}$, there exists a positive semidefinite kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Y}_\mu,\mathcal{Y}_\mu}$ such that $K(x, \cdot)y = L_x^\star y$ for all $y \in \mathcal{Y}_\mu$

Proof: *Since $L_x : \mathcal{Z}_\mu \to \mathcal{Y}_\mu$ is a bounded linear operator on the Banach space $\mathcal{Z}_\mu$ for which the RKHS $\mathcal{H}_{\mu,\nu}$ is defined, by Theorem C.9, there exists a positive semidefinite kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Y}_\mu,\mathcal{Y}_\mu}$ such that $K(x, \cdot)y = L_x^\star y$ for all $y \in \mathcal{Y}_\mu$.* □

D.3 Stochastic integral and differential equations

D.3.1 Wiener Process

Let T be some constant in $[0, \infty]$, ($T = \infty$ is admissible) and $\mathcal{H} = L_0^{2,1}([0,T], \mathbb{R}^n)$ be the separable Hilbert space of once differentiable functions with square integrable derivatives and boundary condition $f(0) = 0$ for all $f \in \mathcal{Z}$ and inner product $\langle f_1, f_2 \rangle_{\mathcal{H}} = \int_0^T \langle Df_1(s), Df_2(s) \rangle_{\mathbb{R}^n} ds$ with $Df = \partial f / \partial t$.

Let $\mathcal{Z} = \mathcal{C}_{b,0}([0,T], \mathbb{R}^n)$ be the Banach space of continuous and bounded functions satisfying the boundary condition $f(0) = 0$ for all $f \in \mathcal{Z}$.

It is known then that with the natural inclusion $i : \mathcal{H} \to \mathcal{Z}$ given as $i(f) = f$ for all $f \in \mathcal{H}$, $\mathcal{H}$ is dense in $\mathcal{Z}$ under the inner product norm. Thus the closure of $\mathcal{H}$ under the inner product norm is $\mathcal{Z}$ and $(i, \mathcal{H}, \mathcal{Z})$ forms an abstract Wiener space. For details we refer the reader to [113, Chapter 1]. It is also known that the Hilbert space in any abstract Wiener space coincides with the RKHS space induced by a Gaussian measure on the separable Banach space [26, Theorem 3.9.5].

Thus for any Gaussian measure $\mu : \mathcal{B}(\mathcal{Z}) \to [0, 1]$, the induced RKHS $\mathcal{H}$ is dense in $\mathcal{Z}$. The classical Wiener measure is a Gaussian measure $\mu_{Wiener} : \mathcal{B}(\mathcal{Z}) \to [0, 1]$ corresponding to a zero mean and covariance $C : \mathcal{Z}^\star \times \mathcal{Z}^\star \to \mathbb{R}$ such that for all linear operators $L_{t,s}$, defined below, the push forward Gaussian measure $L_{t,s_\#}\mu_{Wiener}$ has the covariance matrix $C(t, s) = |t - s|I_n$ and for any finite collection of times $t_1 \geq t_2 \cdots \geq t_n$, the push forward probability measures $L_{t_1,0_\#}\mu_{Wiener}, L_{t_2,t_{1\#}}\mu_{Wiener}, \ldots, L_{t_{n-1},t_{n\#}}\mu_{Wiener}$ are independent Gaussian measures. The identity map $W : \mathcal{Z} \to \mathcal{Z}$ defined as $W(\zeta) = \zeta$ on the probability measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu_{Wiener})$ is then a stochastic Markov process, classically known as the Wiener process.

As is the case for any abstract Wiener space $(i, \mathcal{H}, \mathcal{Z})$ and the corresponding Gaussian measure μ, $\mu(\mathcal{H}) = 0$. Thus for the case of the Wiener process, $\mu_{Wiener}(L_0^{2,1}([0, T], \mathbb{R})^n) = 0$, i.e., the set of differentiable paths of the Wiener process is a zero measure set w.r.t. μ_{Wiener}. Also $\mu_{Wiener}(\mathcal{C}_{b,0}([0, T], \mathbb{R}^n)) = 1$, thus the paths of the Wiener process are continuous and bounded, μ_{Wiener}-almost surely.

Let $P : \mathcal{Z} \to \mathbb{R}^n$ be any bounded linear function, then the push forward probability measure space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_\#\mu_{Wiener})$ is given by the push forward measure $P_\#\mu_{Wiener} = \mu_{Wiener} \circ P^{-1}$. For any $t \in [0, T]$, let $L_t : \mathcal{Z} \to \mathbb{R}^n$ denote the bounded linear evaluation operator $L_t f = f(t)$. Then the push forward probability measure $L_{t\#}\mu_{Wiener}$ defines probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and we denote by W_t or $W(t)$ the random vector $W_t : \mathbb{R}^n \to \mathbb{R}^n$ given by $W_t(x) = x$. W_t can be interpreted as the random vector denoting the values taken by the paths of the Wiener process W at time t. As mentioned earlier the measure μ_{Wiener} is defined such that for all times $t \geq s \geq 0$, $L_{t,s} : \mathcal{Z} \to \mathbb{R}^n := L_t - L_s$ is such that $L_{t,s_\#}\mu_{Wiener}$ is a Gaussian measure with mean zero and covariance $(t - s)I_n$ (I_n being the $n \times n$ identity matrix).

D.3.2 Ito Integral

Let $\{\mathcal{G}_t\}$ be the natural time filtration generated by possible paths of a process W up-to time t, i.e., $\mathcal{G}_t = \sigma(\{W_s : s \in [0, t]\})$. Let W be a process adapted to this filtration.

The Ito Integral is essentially a generalized Lebesgue-Stieltjes integral (clarified below) with the integrands and integrator functions being stochastic processes and refer the reader to [114] for a comprehensive treatment.

The Lebesgue-Stieltjes can be written analogous to the Lebesgue integral defined with respect to a probability measure in Section B.3 as follows,

Definition D.2. *(Lebesgue-Stieltjes integral for monotone integrator functions)* *Let $[a, b] \subseteq \mathbb{R}$ be some interval in $\mathbb{R}$, $\mathcal{X}$ be Banach space and $f : [a, b] \to \mathcal{X}$ be a $(\mathcal{B}(\mathbb{R})|\mathcal{B}(\mathcal{X}))$-measurable function.*

Let $g : [a, b] \to \mathbb{R}$ be a monotone non-decreasing, right continuous function and let $\mu_g : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ be the non-negative Borel measure such that for any interval $[a, b]$, $\mu_g([a, b]) = g(b) - g(a)$ (the Caratheodory's extension theorem is used to show that defining μ_g over every interval in $\mathbb{R}$ as

$\mu_g([a,b]) = g(b) - g(a)$ is sufficient to define an unique Borel measure μ_g over $\mathcal{B}(\mathbb{R})$ that agrees on every interval). The Lebesgue-Stieltjes integral of f with respect to the function g is defined as the Lebesgue integral with respect to the measure μ_g, i.e.,

$$\int_a^b f dg = \int_{[a,b]} f d\mu_g \tag{D.1}$$

Definition D.3. *(Lebesgue-Stieltjes integral for non-monotone integrator functions)*
Let $[a,b] \subseteq \mathbb{R}$ be some interval in $\mathbb{R}$, $\mathcal{X}$ be Banach space and $f : [a,b] \to \mathcal{X}$ be a $(\mathcal{B}(\mathbb{R})|\mathcal{B}(\mathcal{X}))$-measurable function. Let $g : [a,b] \to \mathbb{R}$ be a function of bounded variation, i.e., for any interval $[a,b] \subseteq \mathbb{R}$, let $P_{[a,b]}$ be the set of all finite disjoint interval partitions of $[a,b]$ of the form $\phi = \{a = a_0 < a_1 < \cdots < a_{n-1} < a_n = b\}$ (for some $n \in \mathbb{N}$) and the total variation defined as $V([a,b],g) := \left(\sup_{P_{[a,b]}} \sum_{i=0}^{n-1} |g(a_{i+1}) - g(a_i)|\right)$ is bounded (i.e. $V([a,b],g) < \infty$). To define the Lebesgue-Stieltjes integral with respect to such a g over an interval $[a,b] \subseteq \mathbb{R}$, let $g^+(x) = V([x,a],g)$ and $g^-(x) = g^+(x) - g(x)$ be the two monotonically non-decreasing functions. Using Definition D.2 for integrals with respect to monotone functions, the integral with respect to g is then defined as

$$\int_a^b f dg = \int_a^b f dg^+ - \int_a^b f dg^- \tag{D.2}$$

The Lebesgue-Stieltjes can be extended further to any Banach valued integrator function g as done by [115].

Let $\mathcal{X}, \mathcal{Z}$ be some Banach spaces, $g : [a,b] \to \mathcal{Z}$ be a Banach-valued function on the interval $[a,b] \subseteq \mathbb{R}$. The total variation for a Banach valued function on any interval $[a,b]$ is given by $V([a,b],g) = \sup_{P_{[a,b]}} \sum_{i=0}^{n-1} ||g(a_{i+1}) - g(a_i)||_{\mathcal{Z}}$. Given a third Banach space $\mathcal{Y}$ as a bounded bilinear form $B : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$, a notion of bounded semi-variation is given by $V_B([a,b],g) := \sup_{\phi \in P_{[a,b]}, ||x_i||_{\mathcal{X}} \leq 1} || \sum_{i=0}^{n-1} B(g(a_{i+1}) - g(a_i), x_i)||_{\mathcal{Y}}$. For a function g of bounded semi-variation over the interval $[a,b]$, the Lebesgue-Stieltjes integral with respect to g can be defined as

Definition D.4. *(Lebesgue-Stieltjes integral for Banach valued integrator functions)*
Let $[a,b] \subseteq \mathbb{R}$ be some interval in $\mathbb{R}$, $\mathcal{X}, \mathcal{Z}, \mathcal{Y}$ be three Banach spaces, $f : [a,b] \to \mathcal{X}$ be a $(\mathcal{B}(\mathbb{R})|\mathcal{B}(\mathcal{X}))$-measurable function and $g : [a,b] \to \mathcal{Z}$ be a function of bounded semi-variation with respect to a bounded bilinear map $B : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$. For some partition $\phi \in P_{[a,b]}$, let $S(\phi, f, g) = \sum_{i=0}^{n-1} B(g(a_{i+1}) - g(a_i), f(a_i))$ and let $\delta(\phi) = \sup_{a_i \in \phi} |a_{i+1} - a_i|$ be the maximum interval length in the partition ϕ. If there exists a vector $l \in \mathcal{Y}$ such that for every $\epsilon > 0$ if there exists a $\phi \in P_{[a,b]}$ of maximum length $\delta(\phi)$ satisfying $||l - S(\phi, f, g)||_{\mathcal{Y}} < \epsilon$, then f is said to be Lebesgue-Stieltjes integrable with respect to g on the interval $[a,b]$ and the integral value is l.

$$\forall \epsilon > 0, \quad \exists \phi \in P_{[a,b]} : \left|\left| \int_a^b f dg - S(\phi, f, g) \right|\right|_{\mathcal{Y}} < \epsilon \tag{D.3}$$

In the context of a $\mathbb{R}^n$-valued classical Wiener process the Ito integral can be defined as a special case of the above notion of a Lebesgue-Stieltjes integral for Banach valued integrator functions. Let μ be a Gaussian measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and $\mathcal{Z}$ be a Banach space of measurable functions from $\mathbb{R}^n$

to $\mathbb{R}^n$. Let $F : \mathcal{Z} \to \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ be a Banach valued function $\mathcal{L}_{\mathcal{Z}_\mu,\mathcal{Z}_\mu}$ being the Banach space of bounded linear operators from $\mathcal{Z}$ to itself with the standard induced operator norm. Let $W : [0,T] \to \mathcal{Z}$ be the standard Wiener process defined on the interval $[0,T]$ with $W(0) = 0$, μ_{Wiener}-almost surely. Let $B : \mathcal{L}_{\mathcal{Z},\mathcal{Z}} \times \mathcal{Z} \to \mathcal{Z}$ be the bounded bilinear map given by $B(A,z) = Az$. Then it is easy to verify that W has bounded semi-variation with respect to B (since for any $A \in \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$, $||A(W(t_{i+1}) - W(t_i))||_{\mathcal{Z}} = \mathbb{E}[||A(W(t_{i+1}) - W(t_i))||_{\mathbb{R}^n}] \leq \mathbb{E}[||A||_{\mathbb{R}^{n\times n}}||(W(t_{i+1}) - W(t_i))||_{\mathbb{R}^n}] = \mathbb{E}[||A||_{\mathbb{R}^{n\times n}}]\mathbb{E}[||(W(t_{i+1}) - W(t_i))||_{\mathbb{R}^n}] = 0$ as $\mathbb{E}[||(W(t_{i+1}) - W(t_i))||_{\mathbb{R}^n}] = 0$ for the Wiener process). For any partition $\phi \in P_{[0,T]}$, $S(\phi, F, W) = \sum_{i=0}^{n-1} F(t_i)(W(t_{i+1}) - W(t_i))$ and the Ito integral $\int_0^T F dW$ is the Lebesgue-Stieltjes integral with respect to W as defined in D.4.

Let $X : [0,T] \to \mathcal{Z}$ be a stochastic process of bounded semi-variation with respect to the bilinear map B defined above. Let $W : [0,T] \to \mathcal{Z}$ be the standard Wiener process. Let $\alpha : \mathbb{R}^n \to \mathbb{R}^n$ and $\Sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ be two Borel measurable functions. Then $\alpha \circ X : [0,T] \to \mathcal{Z}$ and $\Sigma \circ X : [0,T] \to \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ are functions of bounded variations under some appropriate restrictions on α, Σ. X is said to be driven by the Wiener process W if it satisfies for all intervals $[0,T]$,

$$\int_0^T dX = \int_0^T \alpha(X)dt + \int_0^T \Sigma(X)dW \qquad (\mu_{Wiener} - a.s.) \tag{D.4}$$

This is written in its differential form as

$$dX = \alpha(X)dt + \Sigma(X)dW \tag{D.5}$$

and X is said to be a solution to the differential equation (D.5).

D.3.3 Fokker-Planck Equation

The Fokker-Planck equation over a Hilbert space $\mathcal{X}$ can be written as the PDE describing a function $\rho : [0,\infty) \times \mathcal{X} \to \mathbb{R}$ satisfying,

$$\frac{\partial \rho}{\partial t} = -\text{div}_{\mathcal{X}}(\rho\alpha) + \text{trace}(\nabla_x^2 \rho \Sigma) \tag{D.6}$$

for given functions $\alpha : \mathcal{X} \to \mathcal{X}$ (called the drift function) and $\Sigma : \mathcal{X} \to \mathcal{L}_{\mathcal{X},\mathcal{X}}$ (called the diffusion function). The gradient operator $\nabla_{\mathcal{X}} : \mathcal{C}^\infty(\mathcal{X}) \to \mathcal{L}_{\mathcal{X},\mathbb{R}}(\mathcal{X})$ maps a function $\rho \in \mathcal{C}^\infty(\mathcal{X})$ to a linear functional field $\nabla_{\mathcal{X}}\rho : \mathcal{X} \to \mathcal{L}_{\mathcal{X},\mathbb{R}}$ such that $\lim_{h\to 0} ||\rho(x+h) - \rho(x) - (\nabla_x \rho)h||_{\mathcal{X}}/||h||_{\mathcal{X}} = 0$ for all $x \in \mathcal{X}$. The divergence operator $\text{div} : \mathcal{C}^\infty(\mathcal{X},\mathcal{X}) \to \mathcal{C}^\infty(\mathcal{X},\mathbb{R})$ is defined as $\text{div}(f) = \sum_{i=1}^n \nabla_{\mathcal{X}}(\langle f, e_i\rangle_{\mathcal{X}})$. The Hessian operator is the tensor field valued operator $\nabla^2_{\mathcal{X}} : \mathcal{C}^\infty(\mathcal{X}) \to \mathcal{L}_{\mathcal{X},\mathcal{X}}(\mathcal{X})$ such that for any $h_1, h_2 \in \mathcal{X}$, $\langle h_1, (\nabla^2_{\mathcal{X}}\rho)h_2\rangle_{\mathcal{X}} = \nabla_{\mathcal{X}}((\nabla_{\mathcal{X}}\rho)h_2)h_1$. $\text{trace} : \mathcal{L}_{\mathcal{X},\mathcal{X}} \to \mathbb{R}$ in a Hilbert space is defined as $\text{trace}(L) = \sum_{i=1}^n \langle e_i, Le_i\rangle_{\mathcal{X}}$ for an orthonormal basis $\{e_i : i = 1, \ldots, n\}$ for a n-dimensional Hilbert space $\mathcal{X}$ (in general n can be ∞ and the same definition is admissible taking $n = \infty$).

Let $\mathcal{P}_{\mathcal{X}}$ be the set of all probability density functions with respect to some measure on $\mathcal{X}$. With initial condition of $\rho(0,\cdot) \in \mathcal{P}_{\mathcal{X}}$ and $\Sigma \in \mathcal{S}^+(\mathcal{X})$ being a positive semidefinite operator in $\mathcal{L}_{\mathcal{X},\mathcal{X}}$, the solution ρ to (D.6) is such that for all time $t \in [0,\infty)$, $\rho(t,\cdot) \in \mathcal{P}_{\mathcal{X}}$. As a result the equation is often used to describe the evolution of probability density functions in time.

In particular, it is known that (D.6) describes the density evolution for the push forward measure at time t for the stochastic differential equation,

$$dX_t = \alpha(X_t)dt + \sqrt{2\Sigma(X_t)}dW_t \tag{D.7}$$

for the standard Wiener process W_t.

Bibliography

[1] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc. 68 (1950), 337-404*, 1950.

[2] G. Wahba, *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, 1990. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9781611970128

[3] J. A. Suykens, C. Alzate, and K. Pelckmans, "Primal and dual model representations in kernel-based learning," *Statist. Surv.*, vol. 4, pp. 148–183, 2010. [Online]. Available: https://doi.org/10.1214/09-SS052

[4] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Math. Dokl.*, vol. 4, pp. 1035–1038, 1963.

[5] M. Unser, J. Fageot, and H. Gupta, "Representer theorems for sparsity-promoting ℓ_1 regularization," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5167–5180, 2016.

[6] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, ser. COLT '01/EuroCOLT '01. London, UK, UK: Springer-Verlag, 2001, pp. 416–426. [Online]. Available: http://dl.acm.org/citation.cfm?id=648300.755324

[7] A. Argyriou, C. A. Micchelli, and M. Pontil, "When is there a representer theorem? vector versus matrix regularizers," *Journal of Machine Learning Research*, vol. 10, no. Nov, pp. 2507–2529, 2009.

[8] F. Dinuzzo and B. Schölkopf, "The representer theorem for hilbert spaces: a necessary and sufficient condition," in *Advances in Neural Information Processing Systems 25.* Curran Associates Inc., 2012, pp. 189–196.

[9] A. Argyriou and F. Dinuzzo, "A unifying view of representer theorems," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–748–II–756. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3044976

[10] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[11] C. K. I. W. Rasmussen, Carl Edward, *Gaussian processes for machine learning.* MIT Press, 2006.

[12] Y. Cho and L. K. Saul, "Kernel methods for deep learning," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 342–350. [Online]. Available: http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf

[13] I. Rebai, Y. BenAyed, and W. Mahdi, "Deep multilayer multiple kernel learning," *Neural Computing and Applications*, vol. 27, no. 8, pp. 2305–2314, Nov 2016. [Online]. Available: https://doi.org/10.1007/s00521-015-2066-x

[14] A. C. Damianou and N. D. Lawrence, "Deep gaussian processes," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, 2013, pp. 207–215. [Online]. Available: http://jmlr.org/proceedings/papers/v31/damianou13a.html

[15] J. A. Suykens, "Deep restricted kernel machines using conjugate feature duality," *Neural computation*, vol. 29, no. 8, pp. 2123–2163, 2017.

[16] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Advances in neural information processing systems*, 2014, pp. 2627–2635.

[17] J. Conway, *A Course in Abstract Analysis*, ser. Graduate studies in mathematics. American Mathematical Soc. [Online]. Available: https://books.google.be/books?id=GD7QxvMOFUcC

[18] C. A. Micchelli and M. A. Pontil, "On learning vector-valued functions," *Neural Comput.*, vol. 17, no. 1, pp. 177–204, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1162/0899766052530802

[19] H. Q. Minh and V. Sindhwani, "Vector-valued manifold regularization," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. USA: Omnipress, 2011, pp. 57–64. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104482.3104490

[20] H. Q. Minh, L. Bazzani, and V. Murino, "A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning," *Journal of Machine Learning Research*, vol. 17, no. 25, pp. 1–72, 2016. [Online]. Available: http://jmlr.org/papers/v17/14-036.html

[21] S. Kulkarni and M. Nair, "A characterization of closed range operators," *Indian Journal of Pure and Applied Mathematics*, vol. 31, no. 4, pp. 353–362, 2000.

[22] J. Conway, *A Course in Functional Analysis*, ser. Graduate Texts in Mathematics. Springer New York, 1994.

[23] A. Sandovici, “Von neumann’s theorem for linear relations,” *Linear and Multilinear Algebra*, vol. 66, no. 9, pp. 1750–1756, 2018. [Online]. Available: https://doi.org/10.1080/03081087.2017.1369930

[24] K. Yoshida, *Functional Analysis*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013. [Online]. Available: https://books.google.ch/books?id=2Mb3CAAAQBAJ

[25] S. Kulkarni, M. Nair, and G. Ramesh, “Some properties of unbounded operators with closed range,” *Proceedings Mathematical Sciences*, vol. 118, no. 4, pp. 613–625, 2008.

[26] V. Bogachev, *Gaussian Measures*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2015. [Online]. Available: https://books.google.ch/books?id=UtufBwAAQBAJ

[27] A. Sudan, O. van Gaans, and P. Spreij, “Infinite order sobolev spaces and the schwartz space,” 2012.

[28] L. Rosasco, M. Belkin, and E. D. Vito, “On learning with integral operators,” *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 905–934, 2010.

[29] G. Gnecco, M. Gori, S. Melacci, and M. Sanguineti, “Learning as constraint reactions,” in *Artificial Neural Networks*. Springer, 2015, pp. 245–270.

[30] A. Zaanen, *Introduction to Operator Theory in Riesz Spaces*. Springer Berlin Heidelberg, 2012. [Online]. Available: https://books.google.ch/books?id=cgvpCAAAQBAJ

[31] I. Dobrakov, “On integration in banach spaces, vii,” *Czechoslovak Mathematical Journal*, vol. 38, no. 3, pp. 434–449, 1988.

[32] G. Schwarz, “Variations on vector measures,” *Pacific Journal of Mathematics*, vol. 23, no. 2, pp. 373–375, 1967.

[33] W. Rudin, *Principles of mathematical analysis*, ser. International series in pure and applied mathematics. McGraw-Hill, 1964. [Online]. Available: https://books.google.ch/books?id=iifvAAAAMAAJ

[34] A. Liniger, A. Domahidi, and M. Morari, “Optimization-Based Autonomous Racing of 1:43 Scale RC Cars,” *Optimal Control Applications and Methods*, vol. 36, no. 5, p. 628 – 647, Jul. 2014.

[35] A. T. Rashid, A. A. Ali, M. Frasca, and L. Fortuna, “Path planning with obstacle avoidance based on visibility binary tree algorithm,” *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1440 – 1449, 2013.

[36] D. Connell and H. M. La, “Dynamic path planning and replanning for mobile robots using rrt,” *arXiv preprint arXiv:1704.04585*, 2017.

[37] J. Liu, P. Jayakumar, J. L. Stein, and T. Ersal, “A nonlinear model predictive control algorithm for obstacle avoidance in autonomous ground vehicles within unknown environments,” Army Tank Automotive Research Development and Engineering Center Warren MI, Tech. Rep., 2015.

[38] M. G. Plessen, D. Bernardini, H. Esen, and A. Bemporad, “Spatial-based predictive control and geometric corridor planning for adaptive cruise control coupled with obstacle avoidance,” *IEEE Trans. on Control Systems Technology*, vol. 26, no. 1, pp. 38–50, Jan 2018.

[39] J. V. Frasch, A. Gray, M. Zanon, H. J. Ferreau, S. Sager, F. Borrelli, and M. Diehl, “An auto-generated nonlinear mpc algorithm for real-time obstacle avoidance of ground vehicles,” in *2013 European Control Conf. (ECC)*, July 2013, pp. 4136–4141.

[40] T. Mercy, W. V. Loock, and G. Pipeleers, “Real-time motion planning in the presence of moving obstacles,” in *2016 European Control Conf. (ECC)*, June 2016, pp. 1586–1591.

[41] Salmah, Sutrisno, E. Joelianto, A. Budiyono, I. E. Wijayanti, and N. Y. Megawati, “Model predictive control for obstacle avoidance as hybrid systems of small scale helicopter,” in *2013 3rd International Conf. on Instrumentation Control and Automation (ICA)*, Aug 2013, pp. 127–132.

[42] S. M. Erlien, S. Fujita, and J. C. Gerdes, “Shared steering control using safe envelopes for obstacle avoidance and vehicle stability,” *IEEE Trans. on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 441–451, Feb 2016.

[43] A. Bemporad and C. Rocchi, “Decentralized hybrid model predictive control of a formation of unmanned aerial vehicles,” *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 11 900 – 11 906, 2011, 18th IFAC World Congress.

[44] A. Bemporad, C. Pascucci, and C. Rocchi, “Hierarchical and hybrid model predictive control of quadcopter air vehicles,” *IFAC Proceedings Volumes*, vol. 42, no. 17, pp. 14 – 19, 2009, 3rd IFAC Conf. on Analysis and Design of Hybrid Systems.

[45] J.-H. Chuang, “Potential-based modeling of three-dimensional workspace for obstacle avoidance,” *IEEE Trans. on Robotics and Automation*, vol. 14, no. 5, pp. 778–785, Oct 1998.

[46] T. Paul, T. R. Krogstad, and J. T. Gravdahl, “Uav formation flight using 3d potential field,” in *2008 16th Mediterranean Conf. on Control and Automation*, June 2008, pp. 1240–1245.

[47] T. Schouwenaars, B. D. Moor, E. Feron, and J. How, “Mixed integer programming for multi-vehicle path planning,” in *2001 European Control Conf. (ECC)*, Sept 2001, pp. 2603–2608.

[48] J. Miura, “Support vector path planning,” in *2006 IEEE/RSJ International Conf. on Intelligent Robots and Systems*, Oct 2006, pp. 2894–2899.

[49] N. Morales, J. Toledo, and L. Acosta, “Path planning using a multiclass support vector machine,” *Applied Soft Computing*, vol. 43, pp. 498 – 509, 2016.

[50] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," *CoRR*, vol. abs/1612.01079, 2016.

[51] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016.

[52] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Simultaneous perception and path generation using fully convolutional neural networks," *CoRR*, vol. abs/1703.08987, 2017.

[53] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *CoRR*, vol. abs/1704.07911, 2017.

[54] X. Huo, X. S. Ni, and A. K. Smith, "A survey of manifold-based learning methods," *Recent advances in data mining of enterprise data*, pp. 691–745, 2007.

[55] A. Domokos, J. M. Ingram, and M. M. Marsh, "Projections onto closed convex sets in hilbert spaces," *Acta Mathematica Hungarica*, vol. 152, no. 1, pp. 114–129, Jun 2017.

[56] R. Luchsinger, "Pumping cycle kite power," in *Airborne Wind Energy*, ser. Green Energy and Technology, U. Ahrens, M. Diehl, and R. Schmehl, Eds. Springer Berlin Heidelberg, 2013, pp. 47–64.

[57] M. L. Loyd, "Crosswind kite power," *Journal of Energy*, vol. 4, pp. 106–111, Jun. 1980.

[58] B. Houska and M. Diehl, "Optimal control for power generating kites," in *In Proceedings of the 9th European Control Conference, Kos, Greece*, 2007, p. 3560–3567.

[59] ——, "Robustness and stability optimization of power generating kite systems in a periodic pumping mode," in *Control Applications (CCA), 2010 IEEE International Conference on*, Sept 2010, pp. 2172–2177.

[60] M. Erhard, G. Horn, and M. Diehl, "A quaternion-based model for optimal control of the SkySails airborne wind energy system," *ArXiv e-prints*, Aug. 2015.

[61] A. Ilzhöfer, B. Houska, and M. Diehl, "Nonlinear mpc of kites under varying wind conditions for a new class of large-scale wind power generators," *International Journal of Robust and Nonlinear Control*, vol. 17, no. 17, pp. 1590–1599, 2007. [Online]. Available: http://dx.doi.org/10.1002/rnc.1210

[62] S. Gros, M. Zanon, and M. Diehl, "Control of airborne wind energy systems based on nonlinear model predictive control amp; moving horizon estimation," in *Control Conference (ECC), 2013 European*, July 2013, pp. 1017–1022.

[63] M. Erhard and H. Strauch, "Flight control of tethered kites in autonomous pumping cycles for airborne wind energy," *ArXiv e-prints*, Sep. 2014.

[64] L. Fagiano, A. Zgraggen, M. Morari, and M. Khammash, “Automatic crosswind flight of tethered wings for airborne wind energy: Modeling, control design, and experimental results,” *Control Systems Technology, IEEE Transactions on*, vol. 22, no. 4, pp. 1433–1447, July 2014.

[65] S. Diwale, A. Alessandretti, I. Lymperopoulos, and C. N. Jones, “A nonlinear adaptive controller for airborne wind energy systems,” in *American Control Conference (ACC), 2016.* American Automatic Control Council (AACC), 2016, pp. 4101–4106.

[66] T. Faulwasser and R. Findeisen, “Nonlinear model predictive control for constrained output path following,” *IEEE Trans. Automat. Contr.*, vol. 61, no. 4, pp. 1026–1039, April 2016.

[67] D. Panagou, H. G. Tanner, and K. J. Kyriakopoulos, “Control of nonholonomic systems using reference vector fields,” in *2011 50th IEEE Conference on Decision and Control and European Control Conference.* IEEE, 2011, pp. 2831–2836.

[68] V. Topogonov, *Differential Geometry of Curves and Surfaces - A Concise Guide.* Birkhäuser, Boston, 2006.

[69] J. Andersson, “A General-Purpose Software Framework for Dynamic Optimization,” PhD thesis, Arenberg Doctoral School, KU Leuven, Department of Electrical Engineering (ESAT/SCD) and Optimization in Engineering Center, Kasteelpark Arenberg 10, 3001-Heverlee, Belgium, October 2013.

[70] A. Wächter and L. T. Biegler, “On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming,” *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10107-004-0559-y

[71] Skysails propulsion for cargo ships. [Online]. Available: http://www.skysails.info/english/skysails-marine/

[72] S. S. Diwale, I. Lymperopoulos, and C. N. Jones, “Optimization of an airborne wind energy system using constrained gaussian processes,” in *2014 IEEE Conference on Control Applications (CCA).* IEEE, 2014, pp. 1394–1399.

[73] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.* Cambridge, MA: MIT Press, 2006.

[74] D. Jones, “A taxonomy of global optimization methods based on response surfaces,” *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[75] E. Brochu, V. M. Cora, and N. de Freitas, “A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning,” arXiv.org, eprint arXiv:1012.2599, December 2010.

[76] M. Schonlau, W. J. Welch, and D. R. Jones, “Global versus local search in constrained optimization of computer models,” in *New Developments and Applications in Experimental Design: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference*, vol. 34, 1998, p. 11.

[77] B. J. Williams, T. J. Santner, W. I. Notz, and J. S. Lehman, "Sequential design of computer experiments for constrained optimization," in *Statistical Modelling and Regression Structures*, T. Kneib and L. Fahrmeir, Eds. Springer - Verlag, 2010, pp. 449–472.

[78] R. B. Gramacy, G. A. Gray, S. L. Digabel, H. K. Lee, P. Ranjan, G. Wells, and S. M. Wild, "Modeling an augmented lagrangian for improved blackbox constrained optimization," Sandia National Laboratories, Livermore, CA, Tech. Rep. arXiv:1403.4890, Mar 2014.

[79] F. Berkenkamp, A. P. Schoellig, and A. Krause, "Safe controller optimization for quadrotors with gaussian processes," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 491–496.

[80] L. Fagiano and M. Milanese, "Airborne wind energy: An overview," in *American Control Conference (ACC 2012)*, Montreal, QC, Jun. 2012, pp. 3132–3143.

[81] M. Erhard and H. Strauch, "Control of towing kites for seagoing vessels," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 5, pp. 1629–1640, 2013.

[82] ——, "Sensors and navigation algorithms for flight control of tethered kites," in *Control Conference (ECC), 2013 European*. IEEE, 2013, pp. 998–1003.

[83] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[84] D. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[85] P. Boyle and M. Frean, "Dependent gaussian processes," in *Advances in neural information processing systems*, 2005, pp. 217–224.

[86] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[87] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2879–2904, 2011.

[88] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, pp. 3250–3265, 2012. [Online]. Available: http://infoscience.epfl.ch/record/177246

[89] M. A. Patterson and A. V. Rao, "GPOPS- II: A matlab software for solving multiple-phase optimal control problems using hp–adaptive gaussian quadrature collocation methods and sparse nonlinear programming," *ACM Transactions on Mathematical Software*, vol. 39, no. 3, 2013.

[90] J. H. Baayen and W. J. Ockels, "Tracking control with adaption of kites," *ArXiv e-prints*, Nov. 2010.

[91] A. Alessandretti, A. Aguiar, and C. Jones, "Trajectory-tracking and path-following controllers for constrained underactuated vehicles using model predictive control," in *Control Conference (ECC), 2013 European*, July 2013, pp. 1371–1376.

[92] R. Carona, A. P. Aguiar, and J. Gaspar, "Control of unicycle type robots tracking, path following and point stabilization."

[93] D.-X. Zhou, "Derivative reproducing properties for kernel methods in learning theory," *Journal of computational and Applied Mathematics*, vol. 220, no. 1-2, pp. 456–463, 2008.

[94] C. S. Ong, X. Mary, S. Canu, and A. J. Smola, "Learning with non-positive kernels," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 81.

[95] H. Zhang and J. Zhang, "Vector-valued reproducing kernel banach spaces with applications to multi-task learning," *Journal of Complexity*, vol. 29, no. 2, pp. 195 – 215, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885064X12000817

[96] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with gaussian processes," in *International Conference on Machine Learning*, 2015, pp. 997–1005.

[97] C. Zimmer, M. Meister, and D. Nguyen-Tuong, "Safe active learning for time-series modeling with gaussian processes," in *Advances in Neural Information Processing Systems*, 2018, pp. 2730–2739.

[98] R. Arens and J. Dugundji, "Topologies for function spaces." *Pacific J. Math.*, vol. 1, no. 1, pp. 5–31, 1951. [Online]. Available: https://projecteuclid.org:443/euclid.pjm/1102613148

[99] R. Engelking, *General topology*, ser. Sigma series in pure mathematics. Heldermann Verlag, 1989. [Online]. Available: https://books.google.ch/books?id=K3spAQAAMAAJ

[100] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. Wiley, 1995. [Online]. Available: https://books.google.ch/books?id=z39jQgAACAAJ

[101] "Chapter 1 probability theory in banach spaces: An introductory survey," in *Random Integral Equations*, ser. Mathematics in Science and Engineering, A. Bharucha-Reid, Ed. Elsevier, 1972, vol. 96, pp. 7 – 63. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0076539208608048

[102] A. C. Zaanen, *Signed Measures and the Radon-Nikodym Theorem*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 183–192. [Online]. Available: https://doi.org/10.1007/978-3-642-60637-3_14

[103] A. Lunardi, M. Miranda, and D. Pallara, *Infinite Dimensional Analysis*. 19th Internet Seminar 2015/2016: Infinite Dimensional Analysis, Dept. of Math. and Comp. Sci., University of Ferrara. [Online]. Available: http://dmi.unife.it/it/ricerca-dmi/seminari/isem19/lectures/lecture-notes/view

[104] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, ser. Universitext. Springer New York, 2010. [Online]. Available: https://books.google.ch/books?id=GAA2XqOIIGoC

[105] N. Dunford and J. Schwartz, *Linear Operators: General theory*, ser. Pure and applied mathematics. Interscience Publishers, 1958. [Online]. Available: https://books.google.ch/books?id=DuJQAAAAMAAJ

[106] D. Lorenz. Dual spaces of continuous functions. [Online]. Available: https://regularize.wordpress.com/2011/11/11/dual-spaces-of-continuous-functions/

[107] A. Knapp, *Basic Real Analysis*. Birkhäuser Boston, 2005. [Online]. Available: https://doi.org/10.1007/0-8176-4441-5_9

[108] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, 2nd ed., ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2014.

[109] G. Kallianpur, "Abstract wiener processes and their reproducing kernel hilbert spaces," *Probability Theory and Related Fields*, vol. 17, no. 2, pp. 113–123, 1971.

[110] J. Mercer and A. R. Forsyth, "Xvi. functions of positive and negative type, and their connection the theory of integral equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 209, no. 441-458, pp. 415–446, 1909. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1909.0016

[111] E. D. Vito, V. Umanità, and S. Villa, "An extension of mercer theorem to matrix-valued measurable kernels," *Applied and Computational Harmonic Analysis*, vol. 34, no. 3, pp. 339 – 351, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1063520312001029

[112] T. Kalmes, A. Pichler *et al.*, "On banach spaces of vector-valued random variables and their duals motivated by risk measures," *Banach Journal of Mathematical Analysis*, vol. 12, no. 4, pp. 773–807, 2018.

[113] D. Bell, *The Malliavin Calculus*, ser. Dover Books on Mathematics. Dover Publications, 2006. [Online]. Available: https://books.google.ch/books?id=e331DAAAQBAJ

[114] G. Shilov, B. Gurevich, and R. Silverman, *Integral, Measure, and Derivative: A Unified Approach*, ser. Dover books on advanced mathematics. Dover Publications, 1966. [Online]. Available: https://books.google.ch/books?id=R3j7tASHOOgC

[115] Š. Schwabik, "Abstract perron-stieltjes integral," *Mathematica Bohemica*, vol. 121, no. 4, pp. 425–447, 1996.

Ce document a été imprimé au Centre d’impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.