2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Non-Convex Learning with Guaranteed Convergence: Perspectives on Stochastic Optimal Control

Alessio Moreschini[1], Matteo Scandella[2], Thomas Parisini[3]

Abstract— **We present and analyze a novel functional learning paradigm that operates on Reproducing Kernel Hilbert Spaces (RKHSs) without relying on the Representer Theorem, demonstrating its potential to learn stochastic optimal control policies in feedback form. Our methodology, based on the newly introduced concept of the Fréchet discrete derivative on RKHS, ensures that sequences generated by the iterative method remain within the intersection of all sublevel sets of the cost function, regardless of the chosen learning rate. In this way, we guarantee a consistent decrease in the cost function evaluation with each iteration until convergence, which is a significant finding in machine learning. By further decomposing the overall functional optimization problem into a suitable sequence of sub-problems, we create a cascade iterative method that computes each function in a domino effect. We briefly address the Witsenhausen counterexample problem, validating our convergence to a local minimum. Although the numerical solution obtained for the Witsenhausen counterexample problem is not yet on par with established ad-hoc methods, the preliminary results are promising and offer a fresh perspective in the field of stochastic optimal control.**

Index Terms— **Functional learning; Non-convex optimization; Team theory; Stochastic optimal control**

I. Introduction

In several major areas of technological interest, it is key to solve infinite-dimensional optimization problems, also called functional optimization problems [1]. Specifically, a functional has to be minimized (or maximized) with respect to admissible policies belonging to infinite-dimensional spaces of functions [2]. Use-cases include large-scale communication and traffic networks, stochastic optimal control in the feedback form of nonlinear dynamic systems, optimal management of complex team organizations, freeway traffic congestion control, and so on. Solving such functional optimization problems in closed form is impossible except in very specific cases and learning the optimal policies is still an open challenge that is worth investigating.

In this initial article, we present a novel and promising functional learning approach that operates on Reproducing Kernel Hilbert Spaces (RKHSs) without relying on the Representer Theorem. The Representer Theorem [3] is a fundamental result in the theory of RKHS. It states that, under some assumptions, the minimizer of a cost function defined over an RKHS can be written as a linear combination of linear functional of the kernel function used to define the RKHS. Nonetheless, this theorem is irrelevant in many functional optimization problems, particularly those involving non-convex optimization [4], such as differential dynamic programming [5,6], and control frameworks such as stochastic optimal control in the feedback form (see Example 1) and team optimal control problems (see Example 2). In these cases, ad hoc numerical approximate solutions are sought that indeed also suffer from the well-known Bellman's *curse of dimensionality* and that are based on iterative methods such as gradient descent. The proposed approach relies on the rationale presented in [7] that solves the problem guaranteeing convergence to a stationary point of the cost function. To shed light on the optimization context that underpins the proposed functional learning methodology, we borrow from [1] two notable examples in stochastic feedback optimal control. These examples serve to illustrate the methodological and practical implications of our methodology and underscore its potential to enable an effective and general policy iteration approach.

Example 1 (Stochastic Optimal Control with Imperfect State Information)**.** We consider a Stochastic Optimal Control problem in which the decision maker (DM) has access to a vector y_t of measurements that only contains imperfect information on the system state x_t. More specifically, we assume that y_t is a p-dimensional vector resulting from an *observation* or *measurement channel* of the form

$$\forall t \in \{1, \dots, T-1\}, \qquad y_t = g_t(x_t, \eta_t), \tag{1}$$

where g_t is a known function and η_t is a random vector. As can be easily understood, the unavailability of perfect information on the system state makes the resulting stochastic optimal control problems much more difficult than problems in which there are perfect measures on x_t.

The DM has then to retain in its memory the information collected up to the stage t in order to compute the optimal decision u_t° by minimizing a cost functional. The information is gathered in a vector that is called the *information vector*. Then, this vector is made of all the measures acquired up to the stage t and all the controls generated up to the stage $t-1$, *i.e.* $I_0 \coloneqq y_0$ and

$$\forall t \in \{1, \dots, T-1\}, \qquad I_t \coloneqq \operatorname{col}\bigl(y_0^t, u_0^{t-1}\bigr),$$

This work has been partially supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 739551 (KIOS CoE).

[1] Alessio Moreschini is with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. Email: a.moreschini@imperial.ac.uk

[2] Matteo Scandella is with the Department of Management, Information and Production Engineering, University of Bergamo, via Marconi 5, 24044, Dalmine (BG), Italy. Email: matteo.scandella@unibg.it

[3] Thomas Parisini is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK, with the Department of Electronic Systems, Aalborg University, Denmark, and with the Department of Engineering and Architecture, University of Trieste, Italy. Email: t.parisini@imperial.ac.uk

where we used the notation $y_0^t = \mathrm{col}\big(y_0, \dots, y_t\big)$ and $u_0^{t-1} = \mathrm{col}\big(u_0, \dots, u_{t-1}\big)$. It is worth noting that the transition from I_t to I_{t+1} is described by a relationship that plays the role of a state equation for the information vector.

We consider a discrete-time dynamic system whose state equation is given by

$$\forall t \in \{1, \dots, T-1\}, \qquad x_{t+1} = f_t(x_t, u_t, \xi_t). \tag{2}$$

Since the state is measurable through (1), we do not know the initial state x_0. We assume that x_0, $\xi_0, \dots, \xi_{T-1}$, $\eta_0, \dots, \eta_{T-1}$ are mutually independent random vectors with known probability density functions and that these functions do not depend on other variables. The optimization problem is

$$\underset{u_t}{\mathrm{argmin}} \sum_{t=0}^{T-1} h_t(x_t, u_t, \xi_t) + h_T(x_T)\,.$$

As stated previously, since the DM cannot measure the state x_t exactly, it has to retain the information state vector I_t in its memory. It follows that the control functions take on the closed-loop form

$$\forall t \in \{1, \dots, T-1\}, \qquad u_t = \mu_t(I_t).$$

We define the sequence of *control functions* as a *control law*. We choose the minimization of the expected value of the cost function as a suitable criterion to design the optimal control law. Thus, the functional to be minimized is given by

$$\underset{\mu_1, \dots, \mu_{t-1}}{\mathrm{argmin}} \underset{x_0, \xi_0^{T-1}, \eta_0^{T-1}}{\mathbb{E}} \left[\sum_{t=0}^{T-1} h_t[x_t, \mu_t(I_t), \xi_t] + h_T(x_T) \right]. \tag{3}$$

Therefore, we can state the following functional optimization problem.

Functional Optimization Problem 1. Find the optimal control functions $\mu_0^\circ, \dots, \mu_{T-1}^\circ$ that solves (3) subject to (1) and (2). ◁

Example 2 (Team Optimal Control)**.** There are many situations in which a process is influenced by several decision-makers DMs. We consider the case where various DMs share different information, but they make decisions aimed at accomplishing a common goal, *i.e.*, minimizing the same cost functional. Such an organization can be mathematically described within the framework of Marschak and Radner's *team theory*. All the a priori information is assumed to be shared by the DMs. This information is given by the cost functional, the probability densities of the random variables, the models of the DMs' observation channels, and, in the dynamic case, of the system controlled in common.

Let us consider a set $\{\mathsf{DM}_1, \dots, \mathsf{DM}_M\}$ of M DMs (or agents). Each of them, on the basis of its own *information vector* $I_i \in \mathbb{R}^{q_i}$, must make its decisions $u_i \in \mathbb{R}^{m_i}$. The information vector I_i of DM_i includes all the information useful for making decisions, that is, information on the decisions made by the other agents and on a vector $z \in \mathbb{R}^r$ that describes a stochastic environment, *i.e.*, the uncertainties in the external world that are not influenced by any DM. Then, each information vector can be described by the function,

$$\forall i \in \{1, \dots, M\}, \quad I_i = g_i(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_M, z),$$

where z is a random vector with a known probability density function. The functions g_i and the connections among them play a central role in team theory.

Definition 1. *The set of functions $g := \{g_1, \dots, g_M\}$ is called the "information structure" of the team.*

A *causality condition* in the teams has to be verified. This means that, if the control action u_i of DM_i affects the information vector I_j of DM_j, then u_j does not affect I_i. By "affecting" we mean that the decisions of a DM modify the information vector of another. The following definition distinguishes two different behaviors in a team:

Definition 2. *A team is said to be static if the functions g_i with $i \in \{1, \dots, M\}$ are independent of the DMs' control actions,* i.e. *$I_i = g_i(z)$. A team that is not static is said to be dynamic.*

Based on what we previously said, the decision or control function of each decision maker DM_i takes the form

$$\forall i \in \{1, \dots, M\}, \qquad u_t = \mu_t(I_t).$$

We assume $\mu_i \in \mathcal{S}_i$, where $\mathcal{S}_i$ is the set of all admissible control functions for the decision maker DM_i. The decision or control law of the team is defined as $\mu := \mathrm{col}(\mu_1, \dots, \mu_M)$. The agents $\mathsf{DM}_1, \dots, \mathsf{DM}_M$ cooperate on the minimization of the mean value of the common cost function

$$J(\mu_1, \dots, \mu_M, z)\,. \tag{4}$$

Hence, we have a *team optimal control problem*. As we said previously, we assume that all the possible a priori information, *i.e.* $g, J, \mathcal{S}_1, \dots, \mathcal{S}_M$ and the probability density function of z, is known to all the DMs. Therefore, we can state the following functional optimization problem.

Functional Optimization Problem 2. Find the optimal control functions $\mu_0^\circ, \dots, \mu_{T-1}^\circ$ that minimize the expected value of (4), *i.e.* $\underset{z}{\mathbb{E}}\big[J(\mu_1(I_1), \dots, \mu_M(I_M), z\big]$ for all $\mu_i \in \mathcal{S}_i$. ◁

To tackle these examples with a more abstract functional learning approach, in Section III-B we propose a novel iterative method that relies on the concept of the Fréchet discrete derivative on RKHS introduced in Section III-A. In particular, the method is designed to ensure that sequences generated by the iterative optimization method remain within the intersection of all sublevel sets of the cost function, regardless of the selected learning rate, ensuring robustness with respect to the selection of the learning rate. Indeed, we have demonstrated that this approach guarantees a monotonic decrease in the cost function evaluation for each iteration until convergence. Building on this property, in Section III-C we devised a cascade technique that simplifies the minimization problem of each function independently. Specifically, we segmented the overarching optimization problem into a sequence of smaller problems. This led to the development of a cascade iterative method that computes each function in a domino effect (as in a decision-making problem), utilizing only the available information at hand.

Finally, to numerically validate the method, in Section III-D we have initiated a preliminary discussion on the potential of the proposed functional method for addressing the Witsenhausen counterexample problem.

II. Preliminaries and problem statement

Throughout the article, we denote by $\mathbb{R}$ and $\mathbb{N}$ the fields of real and natural numbers, respectively ($0 \in \mathbb{N}$). The set of vectors having n rows with real-valued entries is denoted by $\mathbb{R}^n$, and the set of matrices having n rows and m columns with real-valued entries is denoted by $\mathbb{R}^{n\times m}$. Given $n \in \mathbb{N}$ and a vector $x \in \mathbb{R}^n$, $|x|$ is the Euclidean norm of x. Given a random variable $x \in \mathbb{R}$ distributed according to the Normal Distribution with 0-mean, we write $x \sim \mathcal{N}(0,\sigma)$ where $\sigma = \mathbb{E}\big[x^2\big]$. We consider Reproducing Kernel Hilbert Spaces (RKHSs) over $\mathbb{R}$ of functions that map $\mathcal{Q}$ into $\mathbb{R}$, where $\mathcal{Q}$ is an arbitrarily selected set. To define RKHSs in a formal way, we introduce the following statements [8,9].

Definition 3. *A Hilbert space $\mathcal{H}$ over $\mathbb{R}$ of functions $h : \mathcal{Q} \to \mathbb{R}$ is called an RKHS if, for every $q \in \mathcal{Q}$, the evaluation functional $E_{[q]} : \mathcal{H} \to \mathbb{R}$,* i.e. *the functional such that $E_{[q]}(v) = v(q)$ for each $v \in \mathcal{H}$, is linear and bounded.*

In the following, we denote with $\mathcal{H}^*$ is the dual space of the Hilbert space $\mathcal{H}$, *i.e.* the Hilbert space containing all the linear bounded functionals that map $\mathcal{H}$ to $\mathbb{R}$. Then, to better understand this definition, we recall the well-known Riesz Representation Theorem [10, Thm. 3.4], which is a powerful result in the theory of Hilbert spaces which classifies the elements of $\mathcal{H}^*$ in terms of the inner product $\langle\,\cdot\,,\,\cdot\,\rangle$ of $\mathcal{H}$.

Theorem 1. *For each $T \in \mathcal{H}^*$ there exists a unique $g \in \mathcal{H}$, called Riesz representation of T, such that for every $h \in \mathcal{H}$ we have $T(h) = \langle h, g\rangle$.*

From Definition 3 and Theorem 1, we derive that, for each $q \in \mathcal{Q}$, there exist a function $k_q \in \mathcal{H}^*$ such that $v(q) = E_{[q]}(v) = \langle k_q, v\rangle$ for every $v \in \mathcal{H}$. Therefore, for each RKHS, we can define the reproducing kernel function K such that $K(u,v) = \langle k_u, k_v\rangle$. It also turns out that there exist a bijection between reproducing kernels functions and RKHS. This statement is formalized in the following definition and theorem.

Definition 4. *A function $K : \mathcal{Q}\times\mathcal{Q} \to \mathbb{R}$ is called a reproducing kernel if it has the following properties:* ***(i)*** *given $a \in \mathcal{Q}$ and $b \in \mathcal{Q}$, $K(a,b) = K(b,a)$;* ***(ii)*** *given $w \in \mathbb{N}$, $\{a_j : 1 \le j \le w\} \subset \mathcal{Q}$, and $\{c_j : 1 \le j \le w\} \subset \mathbb{R}$,*

$$\sum_{i=1}^{w}\sum_{j=1}^{w} c_i c_j K(a_i, a_j) \ge 0.$$

Theorem 2. *For every RKHS there exists a unique reproducing kernel $K : \mathcal{Q}\times\mathcal{Q} \to \mathbb{R}$. Conversely, given a reproducing kernel $K : \mathcal{Q}\times\mathcal{Q} \to \mathbb{R}$ there exists a unique RKHS of real-valued functions on $\mathcal{Q}$ with K as its reproducing kernel.*

Corollary 2.1. *Given $a \in \mathcal{Q}$, the Riesz representation of $E_{[a]} \in \mathcal{H}^*$ is $K(a,\cdot) \in \mathcal{H}$,* i.e. *$E_{[a]}(v) = \langle v, K(a,\cdot)\rangle_K$ for every $v \in \mathcal{H}$.*

In the following derivations, we denote by $\mathcal{H}_K$ the RKHS associated with the reproducing kernel K, while $\langle\,\cdot\,,\,\cdot\,\rangle_K$ and $|\cdot|_K$ denote the inner product and its induced norm on $\mathcal{H}_K$ respectively.

Throughout this article, we deal with the (possibly non-convex) functional optimization problem

$$\underset{v\in\mathcal{H}_K}{\operatorname{argmin}}\ J(v), \tag{5}$$

where $J : \mathcal{H}_K \to \mathbb{R}$ is an operator that maps functions in $\mathcal{H}_K$ to $\mathbb{R}$ (see Example 1 and 2 previously illustrated). We require the following assumption to guarantee a solution for the minimization problem (5).

Assumption 1. *The function J is bounded from below. Moreover, there exists $h_\star \in \mathcal{H}_K$ such that $J(h_\star) \le J(v)$ for every $v \in \mathcal{H}_K$.*

Optimization problems of the form (5) are widely used in learning problems [11] because, in certain situations, (5) boils down to a finite-dimensional problem for every $\mathcal{H}_K$. In particular, consider the case of

$$J(v) = g\big(L_1(v), L_2(v), \cdots, L_n(v), |v|_K\big),$$

where $L_i \in \mathcal{H}_K^*$, for all $i \in \{1,\dots,n\}$, and g is a function that satisfies specific assumptions[1] which guarantees, among other properties, that the solution of (5) is unique. Then, there exists $c_1,\dots,c_n \in \mathbb{R}$ such that

$$h_\star = \sum_{i=1}^{n} c_i \bar{L}_i,$$

where $\bar{L}_i \in \mathcal{H}_K$ is the Riesz representation of L_i. This result is called Representer Theorem [3], and it allows reducing (5) to an optimization problem in the coefficients $c_1,\dots,c_n$. The Representer Theorem provides a solution in many learning applications in both machine learning [12,13] and system identification [11] for both linear [14]–[17] and nonlinear systems [18,19]. However, if the assumptions of the Representer Theorem are not satisfied, (5) needs to be solved numerically. This is the case for all non-convex optimization problems such as Example 1 and Example 2. A problem belonging to this category is the celebrated Witsenhausen counterexample [20,21] that we consider in Section III-D.

Yet, despite the shortcomings of the Representer Theorem, a numerical solution (5) can be still retrieved using the functional gradient descend method. However, before defining it, we recall the definition of Fréchet Derivative on an RKHS.

[1] The assumptions are omitted here because they are not necessary in this paper. However, details on necessary and sufficient conditions can be found in [3].

Definition 5. *A function $f : \mathcal{H}_K \to \mathbb{R}$ is called Fréchet differentiable at $v \in \mathcal{H}_K$ if there exists $df(v) \in \mathcal{H}_K^*$ such that*

$$\lim_{|h|_K \to 0^+} \frac{|f(v+h) - f(v) - df(v)(h)|}{|h|_K} = 0.$$

The operator $df(v)$ is called Fréchet derivative of f at $v \in \mathcal{H}_K$. The function f is said to be Fréchet differentiable in $\mathcal{H}_K$ if it is Fréchet differentiable at every $v \in \mathcal{H}_K$ and the operator $df(v)$ is simply called Fréchet derivative of f. Since $df(v) \in \mathcal{H}_K^*$, thanks to Theorem 1, there exists a unique $\mathrm{D}f(v) \in \mathcal{H}$ such that $\mathrm{D}f(v)(h) = \langle \mathrm{D}f(v), h \rangle_K$, for every $h \in \mathcal{H}_K$. Then, assuming that J is Fréchet differentiable, the gradient descend method can be defined as an iterative method such that

$$\forall n \in \mathbb{N}, \qquad h_{n+1} = h_n - \delta_n \mathrm{D}J(h_n)$$

where $\{\delta_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ is a bounded sequence the method start from a function $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_K$ is a sequence of functions. However, this method suffers from many well-known shortcomings. When optimizing a model parameterized by a function, gradient descent aims to minimize an objective functional. The update rule for gradient descent involves adjusting the parameters in the opposite direction of the gradient of the objective functional. The values of the sequence δ_n determines the converging rate of the method, and it usually referred as the learning rate of the method. If the learning rate is too small, progress is slow, potentially leading to long search times. Conversely, if it is too large, the optimization process may oscillate until diverging and missing the optimal values. Indeed, the difference between continuous-time gradient dynamics and its discretized version, derived from integrating continuous trajectories, is often substantial and tied to the integration time selection, see *e.g.* [22]–[26].

In the next section, we propose an algorithm that is based on the discrete Fréchet derivative [7] of the objective function which effectively mitigates this diverging phenomenon by allowing for learning rates arbitrarily large.

Before we dive into the main result, we introduce additional notations. Given $\mathcal{A} \subset \mathcal{H}_K$, $s \in \mathcal{H}_K$, and a function $f : \mathcal{H}_K \to \mathbb{R}$ we define the sets

$$\begin{aligned} \mathbb{L}_s(f) &:= \{h \in \mathcal{H}_K \,:\, f(h) \le f(s)\}, \\ S_f(\mathcal{A}) &:= \{v \in \mathcal{A} \,:\, f(v) \ge f(s), \forall s \in \mathcal{A}\}. \end{aligned}$$

Finally, since $s \in S_f(\mathbb{L}_s(f))$ we note that $S_f(\mathbb{L}_s(f)) \neq \varnothing$.

III. Main Result

A. Definition of discrete Fréchet derivative on $\mathcal{H}_K$

The notion of discrete gradient emerged as a powerful tool in several applications of control theory and optimization, see [27]–[30]. The definition of discrete Fréchet derivative was given by [7] referring to Fréchet differentiable functional with domain the set of all continuous functions that map $[a,b]$ to $\mathbb{R}^n$. Here, we introduce the notion of discrete Fréchet derivative for functions that map $\mathcal{H}_K$ to $\mathbb{R}$.

Definition 6. *A function $f : \mathcal{H}_K \to \mathbb{R}$ is called discrete Fréchet differentiable at $v \in \mathcal{H}_K$ if there exists $\overline{\mathrm{D}f}(v,\cdot) : \mathcal{H}_K \to \mathcal{H}_K$ such that for every $s \in \mathcal{H}_K$*

$$\left\langle s - v, \overline{\mathrm{D}f}(v,s) \right\rangle_K = f(s) - f(v), \tag{6a}$$

$$\lim_{s \to v} \overline{\mathrm{D}f}(v,s) = \mathrm{D}f(v). \tag{6b}$$

The operator $\overline{\mathrm{D}f}(v,\cdot)$ is called discrete Fréchet derivative of f at $v \in \mathcal{H}_K$. The function f is said to be discrete Fréchet differentiable in $\mathcal{H}_K$ if it is Fréchet differentiable at every $v \in \mathcal{H}_K$ and the operator $\overline{\mathrm{D}f}(\cdot,\cdot) : \mathcal{H}_K \times \mathcal{H}_K \to \mathcal{H}_K$ is called discrete Fréchet derivative of f.

Example 3. Consider $q \in \mathcal{Q}$ and its evaluation functional $E_{[q]}$. Using Corollary 2.1 and for the linearity of the inner product, for every $s, v \in \mathcal{H}_K$, we have

$$\begin{aligned} E_{[q]}(s) - E_{[q]}(v) &= \langle s, K(q,\cdot) \rangle_K - \langle v, K(q,\cdot) \rangle_K \\ &= \langle s - v, K(q,\cdot) \rangle_K \end{aligned}$$

Using the property (6a), the difference $E_{[q]}(s) - E_{[q]}(h)$ implies that $\langle s - v, K(q,\cdot) \rangle_K = \left\langle s - v, \overline{\mathrm{D}E_{[q]}}(v,s) \right\rangle_K$. Thus, we have that $\overline{\mathrm{D}E_{[q]}}(v,s) = K(v,\cdot)$ for every $q \in \mathcal{Q}$. ◁

Example 4. Consider the norm functional $L(\cdot) = |\cdot|_K^2$ and two functions $v, s \in \mathcal{H}_K$. Using the linearity property of the inner product, we have

$$\begin{aligned} L(s) - L(v) &= |s|_K^2 - |v|_K^2 \\ &= \langle s, s \rangle_K - \langle v, v \rangle_K \\ &= \langle s - v, s \rangle_K + \langle v, s - v \rangle_K \\ &= \langle s - v, s + v \rangle_K. \end{aligned}$$

Finally, using (6a) the difference $L(s) - L(v)$ implies that $\langle s - v, s + v \rangle_K = \left\langle s - v, \overline{\mathrm{D}L}(v,s) \right\rangle_K$. Thus, we have that $\overline{\mathrm{D}L}(v,s) = s + v$. ◁

It is worth bearing in mind that the discrete Fréchet derivative obtained from equation (6) is usually non-unique. This highlights the need for careful consideration and analysis of discrete Fréchet derivative on $\mathcal{H}_K$. It must be emphasized that the forthcoming derivations are universally applicable to any derivative satisfying (6), and are not limited to any particular one.

B. Discrete Fréchet gradient method on $\mathcal{H}_K$

Following the functional learning method discussed in [7], we introduce a discrete Fréchet gradient iterative method which is given by the (implicit) difference equation

$$\forall n \in \mathbb{N}, \qquad h_{n+1} = h_n - \delta_n \overline{\mathrm{D}J}(h_n, h_{n+1}), \tag{7}$$

where $J : \mathcal{H}_K \to \mathbb{R}$ is an operator that maps functions in $\mathcal{H}_K$ to $\mathbb{R}$, $\{\delta_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ is a bounded sequence of learning rates, and $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_K$ is a sequence of functions. Here, we assume that J is a discrete Fréchet differentiable functional. Hence, to find a local minimum for J, the idea is to apply an iterative method for every k constructing a sequence $\{h_n\}_{n\in\mathbb{N}}$ starting from h_0 such that

h_{n+1} is a zero of the implicit equation (7). The method (7) is called the discrete Fréchet gradient method on $\mathcal{H}_k$.

The following statement is a preliminary result of the discrete Fréchet gradient method on $\mathcal{H}_K$ which holds regardless of the sequence $\{\delta_n\}_{n\in\mathbb{N}}$ and the shape of J.

Lemma 1. *Let $J:\mathcal{H}_K\to\mathbb{R}$ be any discrete Fréchet differentiable function, and $\{\delta_n\}_{n\in\mathbb{N}}\subset(0,\infty)$ and $\{h_n\}_{n\in\mathbb{N}}\subset\mathcal{H}_K$ such that* (7) *holds. Then:*

$$\forall n\in\mathbb{N},\quad J(h_{n+1})=J(h_n)\iff h_{n+1}=h_n; \tag{8a}$$

$$\forall n\in\mathbb{N},\quad h_{n+1}\in\bigcap_{i=0}^{n}\mathbb{L}_{h_i}(J). \tag{8b}$$

Proof of (8a). The left implication ($\Leftarrow$) is trivial. The right implication ($\Rightarrow$) holds from the assumption that J is a discrete Fréchet differentiable function associated with (7). Indeed, using (6a) and (7) we have for every $n\in\mathbb{N}$,

$$\begin{aligned}J(h_{n+1})-J(h_n)&=\left\langle h_{n+1}-h_n,\overline{\mathrm{D}J}(h_n,h_{n+1})\right\rangle_K\\&=-\delta_n^{-1}|h_{n+1}-h_n|_K^2\le 0.\end{aligned}\tag{9}$$

Thus, if $J(h_{n+1})=J(h_n)$, we have $\delta_n^{-1}|h_{n+1}-h_n|_K^2=0$. Finally, since δ_n is bounded for all n by assumption, we conclude that $|h_{n+1}-h_n|_K^2=0\implies h_{n+1}=h_n$ for every $n\in\mathbb{N}$. □

Proof of (8b). Since (9) holds for every $n\in\mathbb{N}$ and every δ_n, and since $|h_{n+1}-h_n|_K^2\ge 0$ for every $h_n\in\mathcal{H}_K$ and $h_{n+1}\in\mathcal{H}_K$, we have that

$$\begin{aligned}(9)&\implies J(h_{n+1})\le J(h_n)\\&\implies J(h_{n+1})\le J(h_n)\le\cdots\le J(h_1)\le J(h_0)\\&\implies \mathbb{L}_{h_n}(J)\subseteq\cdots\subseteq\mathbb{L}_{h_1}(J)\subseteq\mathbb{L}_{h_0}(J).\end{aligned}$$

From the implications above and the definition of the sublevel set of J, we conclude that

$$h_{n+1}\in\mathbb{L}_{h_n}(J)\subseteq\cdots\subseteq\mathbb{L}_{h_0}(J)\implies h_{n+1}\in\bigcap_{i=0}^{n}\mathbb{L}_{h_i}(J).$$

□

Lemma 1 ensures that any sequence $\{h_n\}_{n\in\mathbb{N}}$ generated by the iterative method (7) is contained in the intersection of all the sublevel sets at each $n\in\mathbb{N}$, for every sequence $\{\delta_n\}$ and any fixed initial function $h_0\in\mathcal{H}_K$. In addition, the iterative method (7) also ensures that each sublevel sets of $\mathbb{L}_{h_n}(J)$ never increase with n, see Figure 1. Combining (8a) and (8b), we note that if $h_{n+1}\ne h_n$ then $J(h_{n+1})<J(h_n)$. Thus, the method guarantees that the cost function evaluation decreases monotonically with n unless the algorithm reaches convergence where $h_{n+1}=h_n$. Yet, without any further assumption on J, it turns out that convergence is not guaranteed because, for each $n\in\mathbb{N}$, $\mathbb{L}_{h_n}(J)$ is closed but not necessarily bounded from below. However, if Assumption 1 is satisfied, we can prove the following result.

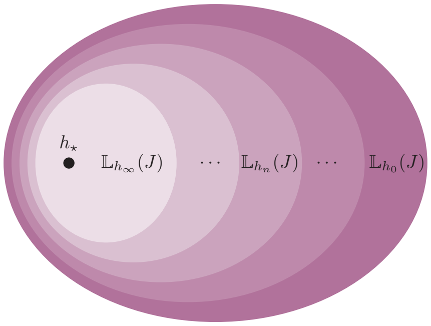

Fig. 1. Sublevel sets of J for every sequence $\{h_n\}_{n\in\mathbb{N}}$ generated by (7).

Theorem 3. *Let $J:\mathcal{H}_K\to\mathbb{R}$ be a discrete Fréchet differentiable function which satisfies Assumption 1, and $\{\delta_n\}_{n\in\mathbb{N}}\subset(0,\infty)$ and $\{h_n\}_{n\in\mathbb{N}}\subset\mathcal{H}_K$ such that* (7) *holds. Then, $\{h_n\}_{n\in\mathbb{N}}$ is bounded, and there exists $h_\infty\in\mathcal{H}_K$ such that*

$$h_n\xrightarrow[n\to\infty]{}h_\infty\in S_J\left(\bigcap_{i=0}^{\infty}\mathbb{L}_{h_i}(J)\right).\tag{10}$$

Proof. From Assumption 1, for every $n\in\mathbb{N}$ we have

$$(9)\Rightarrow J(h_\star)\le J(h_{n+1})\le J(h_n)\le\cdots\le J(h_1)\le J(h_0).$$

Accordingly, there exists (at least) a function $h_\infty\in\mathcal{H}_K$ such that $J(h_\star)\le J(h_\infty)\le J(h_0)$ with the property that $J(h_n)\to J(h_\infty)$ as $n\to\infty$. For every $n\in\mathbb{N}$, using (8b) and the trivial fact $h_{n+1}\in\mathbb{L}_{h_{n+1}}(J)$, we have

$$h_{n+1}\in\bigcap_{i=0}^{n+1}\mathbb{L}_{h_i}(J)\implies h_\infty\in S_J\left(\bigcap_{i=0}^{\infty}\mathbb{L}_{h_i}(J)\right).$$

Finally, we note that $\mathbb{L}_{h_n}(J)\subset\mathcal{H}_K$ can be equivalently defined as the preimage $J^{-1}([J(h_\star),J(h_n)])$ for every h_n. Then every sublevel set $\mathbb{L}_{h_n}(J)$ is compact since $[J(h_\star),J(h_n)]$ is a compact subset of $\mathbb{R}$ for every h_n. Therefore, by induction, we finally conclude that $\bigcap_{i=0}^{\infty}\mathbb{L}_{h_i}(J)\subset\mathcal{H}_K$ is compact and $\{h_n\}_{n\in\mathbb{N}}$ is bounded in $\mathcal{H}_K$. □

Example 5 (Application to the one dimensional non-convex case)**.** Consider the reproducing kernel $K(a,b)=\exp(-|a-b|^2)$, for all $a,b\in\mathcal{Q}=[0,5]\subset\mathbb{R}$ and the RKHS that it defines $\mathcal{H}_K$. We aim to solve the non-convex problem (5) with $J:\mathcal{H}_K\to\mathbb{R}$ such that

$$J(h)=|h|_K^2+\sum_{i=0}^{10}\Big(y_i-\sin\big(h(x_i)\big)\Big)^2,$$

where, for every $i\in\{0,\dots,10\}$, $x_i=\frac{i}{2}$ and $y_i=\frac{\log|x_i+1|}{\cos x_i}$. For this cost function, the Representer Theorem does not hold [3] and needs to be solved numerically. Yet, using the chain rule and the discrete Fréchet derivatives obtained in Examples 3 and 4, we have that

$$\begin{aligned}\overline{\mathrm{D}J}(h_n,h_{n+1})&=h_{n+1}+h_n+\sum_{i=0}^{10}\overline{\mathrm{D}E_{[x_i]}}(h_n,h_{n+1})w(x_i)\\&=h_{n+1}+h_n+\sum_{i=0}^{10}K(x_i,\cdot)w_n(x_i),\end{aligned}$$

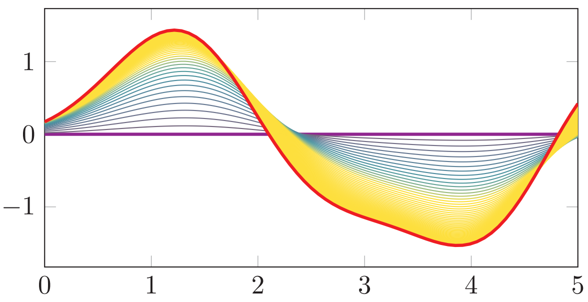

Fig. 2. The first 50 functions obtained from Example 5. The thick purple line is the starting function h_0 and the red think line is the function of the last iteration h_n.

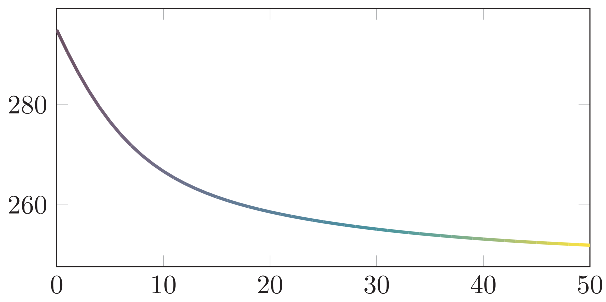

Fig. 3. Values of the cost function J for the first 50 functions h_n obtained from Example 5.

in which

$$w_n(x_i) := (\sin(h_{n+1}(x_i)) + \sin(h_n(x_i)) - y_i)p_n(x_i),$$
$$p_n(x_i) := \frac{\sin(h_{n+1}(x_i)) - \sin(h_n(x_i))}{h_{n+1}(x_i) - h_n(x_i)}.$$

The proposed method requires the solution of an implicit equation in a potentially infinite dimensional functional space. Thus, to tackle the problem numerically, we need to rely on a finite-dimensional approximation of the function h_n. Here, since $h_n \in \mathcal{H}_K$ for every $n \in \mathcal{H}_K$, we can employ the approximation

$$\forall n \in \mathbb{N}, \qquad h_n \approx \sum_{j=0}^{m} c_{j,n} K(\bar{x}_j, \cdot) \in \mathcal{H}_K, \tag{11}$$

in which $c_{j,i} \in \mathbb{R}$, for all $j \in \{0, \dots, m\}$ and $i \in \mathbb{N}$ and $\bar{x}_i = \frac{5}{m} i$. Moreover, the implicit equation (7) can be evaluated in a finite amount of points of the domain $\mathcal{Q} = [0, 5]$. Hence, to make the problem feasible, we guarantee that the equality in (7) holds only when the functions are evaluated in $\bar{x}_i$ with $i \in \{0, \dots, m\}$. Finally, after these simplifications, the implicit equation (7) becomes

$$\mathbf{K}\mathbf{c}_{n+1} = \mathbf{K}\mathbf{c}_n - \delta_n(\mathbf{K}\mathbf{c}_{n+1} + \mathbf{K}\mathbf{c}_n + \mathbf{K}_s\mathbf{w}_n),$$

where $\mathbf{K} \in \mathbb{R}^{m+1 \times m+1}$ is the symmetric matrix whose (i, j)-th element is $K(\bar{x}_i, \bar{x}_j)$, $\mathbf{K}_s \in \mathbb{R}^{m+1 \times 11}$ is the matrix whose (i, j)-th element is $K(\bar{x}_i, x_j)$, $\mathbf{c}_i \in \mathbb{R}^{m+1}$ are the vector whose j-th element is $c_{j,i}$ (for every $i \in \mathbb{N}$) and $\mathbf{w}_i \in \mathbb{R}^{11}$ is the vector whose j-th element is $w_i(x_i)$ computed using the approximation (11) (for every $i \in \mathbb{N}$).

If m multiple of 10, then there exists a matrix $\mathbf{S} \in \mathbb{R}^{m \times 11}$ such that $\mathbf{K}_s = \mathbf{K}\mathbf{S}$. Thus, we finally obtain the equivalent implicit equation

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \delta_n(\mathbf{c}_{n+1} + \mathbf{c}_n + \mathbf{S}\mathbf{w}_n),$$

which can be solved to find $\mathbf{c}_{n+1}$ that defines the approximation of the function as in (11). The first 50 iterations obtained using $\delta_n = 0.005$, for all $n \in \mathbb{N}$, $m = 1000$, and starting from $\mathbf{c}_0 = \mathbf{0}_{m \times 1}$ are shown in Figure 2. Furthermore, it is readily seen in Figure 3 that the value of the cost function decreases monotonically at every iteration as proven in Lemma 1. $\triangleleft$

C. Cascade Discrete Fréchet Gradient Descent

In the realm of control theory, the backstepping approach is a powerful tool for stabilizing a specific class of nonlinear systems in a cascade fashion. The essence of backstepping lies in its recursive nature and the control design unfolds like a chain reaction. By breaking down the overall control problem into a series of sub-problems, backstepping ensures that each state is stabilized by a fictitious control which is the next state. Inspired by this cascade design, we redefine the discrete Fréchet gradient iterative method (7) as a recursive algorithm.

Consider $N \in \mathbb{N}$ reproducing kernels $K_1, \dots, K_N : \mathcal{Q} \to \mathbb{R}$ and the optimization problem

$$\underset{v_1 \in \mathcal{H}_{K_1}, \cdots, v_N \in \mathcal{H}_{K_N}}{\operatorname{argmin}} \qquad J(v_1, \dots, v_N),$$

where $J : \mathcal{H}_{1:N} \to \mathbb{R}$ where $\mathcal{H}_{1:N} := \mathcal{H}_{K_1} \times \cdots \times \mathcal{H}_{K_N}$. For simplicity, for every $(h^1, \dots, h^N) \in \mathcal{H}_{1:N}$ and $i \in \{1, \dots, N\}$, we define $h^{:i} := (h^1, \dots, h^{i-1})$ and $h^{i:} := (h^{i+1}, \dots, h^N)$. We also assume that the function $J\big(h^{:i}, \cdot, h^{i:}\big)$ is discrete Fréchet differentiable. Thus, we have that for every $h \in \mathcal{H}_{1:N}$, $i \in \{1, \dots, N\}$ and $s^i \in \mathcal{H}_{K_i}$,

$$\big\langle s^i - h^i, \overline{\mathrm{D}J}\big(s^i, h^{:i}, h^i, h^{i:}\big)\big\rangle_{K_i} = J\big(h^{:i}, s^i, h^{i:}\big) - J(h),$$

where $\overline{\mathrm{D}J}\big(\cdot, h^{:i}, h^i, h^{i:}\big) : \mathcal{H}_{K_i} \to \mathcal{H}_{K_i}$ is the (partial) discrete Fréchet derivative of f at $h_i \in \mathcal{H}_{K_i}$. With this in mind, we can characterize the discrete Fréchet gradient method (7) as a cascade discrete-time system described by the implicit difference equations

$$h^1_{n+1} = h^1_n - \delta_{1,n}\overline{\mathrm{D}J}\big(h^1_{n+1}, h^{:1}_n, h^1_n, h^{1:}_n\big), \tag{12a}$$
$$\vdots$$
$$h^N_{n+1} = h^N_n - \delta_{N,n}\overline{\mathrm{D}J}\big(h^N_{n+1}, h^{:N}_{n+1}, h^N_n, h^{N:}_n\big), \tag{12b}$$

in which, for every $i \in \{1, \dots, N\}$, $\{\delta_{i,n}\}_{n\in\mathbb{N}} \subset (0, \infty)$ is a bounded sequence and $\big\{h^i_n\big\}_{n\in\mathbb{N}} \subset \mathcal{H}_{K_i}$ is a sequence of functions.

We note that unlike the general method (7) the cascade descent algorithm (12) allows computing each function separately only using the available information. Moreover, it is possible to use different sequences of learning rates, $\{\delta_{i,n}\}_{n\in\mathbb{N}}$, for each sequence $\big\{h^i_n\big\}_{n\in\mathbb{N}}$. Indeed, we notice that any sequence $\{h_n\}_{n\in\mathbb{N}}$ generated by the method (12)

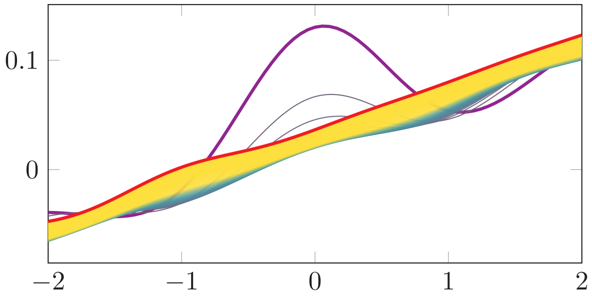

Fig. 4. The functions μ_1 obtained by applying the cascade method (12) to the cost function (14) at the $10i$ iteration with $i \in \{0, \dots, 50\}$. The thick purple line is the starting function $\mu_{1,0}$ and the red think line is the function of the last iteration $\mu_{1,n}$.

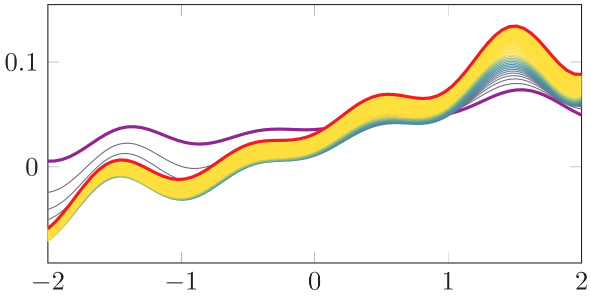

Fig. 6. The functions μ_2 obtained by applying the cascade method (12) to the cost function (14) at the $10i$ iteration with $i \in \{0, \dots, 50\}$. The thick purple line is the starting function $\mu_{2,0}$ and the red think line is the function of the last iteration $\mu_{2,n}$.

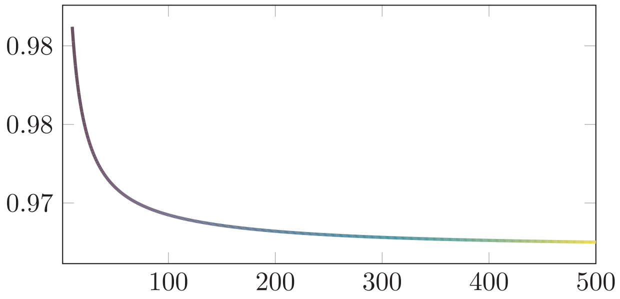

Fig. 5. The value of the cost function (14) obtained by applying the cascade method (12).

is again contained in the intersection of all the sublevel sets at each $n \in \mathbb{N}$, for every learning rate sequence and any fixed initial function $h_0 \in \mathcal{H}_K$. Following the argument of Lemma 1, the cascade method (12) yields for every $n \in \mathbb{N}$

$$J(h_{n+1}) - J(h_n) = -\sum_{i=1}^{N} \delta_{i,n}^{-1} \big| h_{n+1}^i - h_n^i \big|_{K_i}^2 \leq 0.$$

D. A Glimpse of the Witsenhausen Counterexample

Several problems in control theory, *e.g.* decentralized stochastic control problems, can be formulated as a chain of multi-stage decision-making problems [1]. A simple — yet still unsolved — decision-making problem is the *Witsenhausen counterexample*, formulated by Hans Witsenhausen in [20]. The counterexample was conceptualized to demonstrate how the decentralization of stochastic controls can break the certainty-equivalence property and how the choice of certain nonlinear functions may outperform all linear ones. Witsenhausen, however, did not demonstrate that the proposed nonlinear solution is a global one, and to date, neither analytical nor numerical methods have been found to determine the global solution to the problem. Although the global solution continues to be an unresolved problem, numerous notable efforts have been undertaken in recent years. For a summary of these numerical approaches, refer to [21] and [1, Sec 9.4].

The problem can be formulated as a two-stage decision-making problem of two Borel measurable functions $\mu_1 : \mathbb{R} \to \mathbb{R}$ and $\mu_2 : \mathbb{R} \to \mathbb{R}$. At the first decision stage, we assume observing a random variable $x \sim \mathcal{N}(0, \sigma^2)$. The choice of μ_1 is associated with the (first stage) cost function

$$J_1(\mu_1) = \mathbb{E}\Big[k^2 \big(x - \mu_1(x)\big)^2\Big],$$

for some $k \in (0, \infty)$. At the second decision stage, the knowledge of μ_1 is altered by an additive random variable $\eta \sim \mathcal{N}(0, 1)$ that is mutually independent with x. The choice of μ_2 is associated with the (second stage) cost function

$$J_2(\mu_1, \mu_2) = \mathbb{E}\bigg[\Big(\mu_1(x) - \mu_2\big(\mu_1(x) + \eta\big)\Big)^2\bigg].$$

The Witsenhausen counterexample requires finding a pair of functions (μ_1, μ_2) which solve the minimization problem

$$\operatorname*{argmin}_{\mu_1, \mu_2} \quad J_1(\mu_1) + J_2(\mu_1, \mu_2). \tag{13}$$

We have already seen in Section III-C that the cascade method (12) aligns seamlessly with this two-stage decision problem. Hence, to approach the minimization problem (13) through (12) we need to reformulate further the Witsenhausen counterexample in terms of functions (μ_1, μ_2) belonging to RKHSs. In this respect, considering two reproducing kernels $K_1, K_2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $\mu_1 \in \mathcal{H}_{K_1}$ and $\mu_2 \in \mathcal{H}_{K_2}$, we can redefine the Witsenhausen counterexample problem as such

$$\operatorname*{argmin}_{\mu_1 \in \mathcal{H}_{K_1}, \mu_2 \in \mathcal{H}_{K_2}} \quad J_1(\mu_1) + J_2(\mu_1, \mu_2), \tag{14}$$

where $J_1 : \mathcal{H}_{K_1} \to \mathbb{R}$ and $J_2 : \mathcal{H}_{K_1} \times \mathcal{H}_{K_2} \to \mathbb{R}$. To simulate the cascade method (12) we consider approximating the expected values through the Monte Carlo method for $n_{\mathrm{mc}} \in \mathbb{N}$ mutually independent samples. Thus, for all $i \in \{1, \dots, n_{\mathrm{MC}}\}$, let $x_i \sim \mathcal{N}(0, \sigma)$ and $\eta_i \sim \mathcal{N}(0, 1)$ be mutually independent random variables. Then, the cost function (14) can be approximated with

$$J_1(\mu_1) \approx \frac{k^2}{n_{\mathrm{mc}}} \sum_{i=1}^{n_{\mathrm{mc}}} \big(x_i - \mu_1(x_i)\big)^2,$$

$$J_2(\mu_1, \mu_2) \approx \frac{1}{n_{\mathrm{mc}}} \sum_{i=1}^{n_{\mathrm{mc}}} \Big(\mu_1(x_i) - \mu_2\big(\mu_1(x_i) + \eta_i\big)\Big)^2.$$

For the numerical validation of the method, we considered $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ associated with the reproducing kernels

$K_1(a,b) = K_2(a,b) = \exp(-|a-b|^2)$, for all $a, b \in \mathbb{R}$. The cost function (14) was implemented using the benchmark values $k = 0.2$, $\sigma = 5$, and $n_{\mathrm{mc}} = 3101$. The learning rate for both functions is a constant sequence of learning rate $\delta_n = 1$, for all $n \in \mathbb{N}$. A subset of the nonlinear function μ_1 and μ_2 obtained for the first 500 iterations are reported in Figure 4 and Figure 6 respectively. It is readily seen in Figure 5 that the functions μ_1 and μ_2 are approaching a (local) minimum[2] of the approximated cost function and the value of the cost function decreases monotonically at every iteration until convergence. This preliminary convergence outcome is encouraging, suggesting that with a meticulous tuning of this cascade method and a proper selection of the reproducing kernel, we may find a valuable numerical approach for seeking the (numerical) global minimum.

IV. Conclusions and Perspectives

We introduced a novel functional learning approach that operates on Reproducing Kernel Hilbert Spaces (RKHSs) without relying on the Representer Theorem. Our implicit method – which relies on the concept of the Fréchet discrete derivative on RKHS introduced here – is designed to ensure that sequences generated by the iterative method remain within the intersection of all sublevel sets of the (possibly non-convex) cost function, regardless of the selected learning rate. As a result, we have demonstrated that this approach guarantees a monotonic decrease in the cost function evaluation for each iteration until convergence. Based on this, we further developed a cascade method that facilitates the computation of each function in isolation. In particular, by breaking down the overall optimization problem into a series of sub-problems, we also developed a cascade iterative method that allows computing each function as a chain reaction only using the available information at hand.

The cascade structure seamlessly integrates into the Witsenhausen counterexample problem, allowing us to validate our method by demonstrating the algorithm's convergence to a local minimum. Yet, while the obtained numerical solution to the Witsenhausen counterexample problem is not yet comparable with the established methods of the literature, the preliminary results shown in this work are promising and provide a new perspective. Specifically, with a careful refinement of the cascade method and a strategic selection of the reproducing kernel, our method could emerge as a potent numerical strategy for pursuing the elusive global minimum in the Witsenhausen counterexample problem — a feat that is widely acknowledged as complex and challenging.

References

[1] R. Zoppoli, M. Sanguineti, G. Gnecco, and T. Parisini, *Neural Approximations for Optimal Control and Decision*. Springer Int. Publ., 2020.

[2] D. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, vol. 5. Athena Sci., 1996.

[3] F. Dinuzzo and B. Schölkopf, "The representer theorem for Hilbert spaces: a necessary and sufficient condition," in *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[4] P. Jain and P. Kar, "Non-convex optimization for machine learning," *Found. Trends Mach. Learn.*, vol. 10, no. 3-4, pp. 142–363, 2017.

[5] D. Q. Mayne, "A second-order gradient method for determining optimal trajectories of non-linear discrete-time systems," *Int. J. Control*, vol. 3, no. 1, pp. 85–95, 1966.

[6] D. Q. Mayne, "Differential dynamic programming – a unified approach to the optimization of dynamic systems," in *Control and Dynamic Systems*, vol. 10, pp. 179–254, 1973.

[7] A. Moreschini, G. Göksu, and T. Parisini, "Fréchet Discrete Gradient and Hessian Operators on Infinite-Dimensional Spaces," in *Proc. 7th IFAC Conf. Anal. Control Nonlinear Dyn. Chaos*, vol. 58, pp. 78–83, 2024.

[8] N. Aronszajn, "Theory of reproducing kernels," *Trans. Am. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[9] S. Saitoh and Y. Sawano, *Theory of Reproducing Kernels and Applications*. Springer Singap., 2016.

[10] J. B. Conway, *A Course in Functional Analysis*, vol. 96. Springer, 2019.

[11] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung, *Regularized System Identification: Learning Dynamic Models from Data*. Springer Int. Publ., 2022.

[12] J. A. K. Suykens, C. Alzate, and K. Pelckmans, "Primal and dual model representations in kernel-based learning," *Stat. Surv.*, vol. 4, no. none, pp. 148–183, 2010.

[13] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.

[14] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[15] M. Scandella, M. Mazzoleni, S. Formentin, and F. Previdi, "Kernel-based identification of asymptotically stable continuous-time linear dynamical systems," *Int. J. Control*, vol. 95, no. 6, pp. 1668–1681, 2021.

[16] M. Scandella, A. Moreschini, and T. Parisini, "Kernel-Based Continuous-Time System Identification: A Parametric Approximation," in *Proc. 62nd IEEE Conf. Decis. Control (CDC)*, pp. 1492–1497, 2023.

[17] M. Scandella, M. Mazzoleni, S. Formentin, and F. Previdi, "A Note on the Numerical Solutions of Kernel-Based Learning Problems," *IEEE Trans. Autom. Control*, vol. 66, no. 2, pp. 940–947, 2021.

[18] G. Pillonetto, M. H. Quang, and A. Chiuso, "A New Kernel-Based Approach for NonlinearSystem Identification," *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2825–2840, 2011.

[19] M. Scandella, M. Bin, and T. Parisini, "Kernel-Based Identification of Incrementally Input-to-State Stable Nonlinear Systems," in *Proc. 22nd IFAC World Congr.*, vol. 56, pp. 5127–5132, 2023.

[20] H. S. Witsenhausen, "A Counterexample in Stochastic Optimum Control," *SIAM J. Control*, vol. 6, no. 1, pp. 131–147, 1968.

[21] Y.-C. Ho, "Review of the Witsenhausen problem," in *Proc. 47th IEEE Conf. Decis. Control (CDC)*, pp. 1611–1613, 2008.

[22] R. P. Feynman, *Feynman Lectures on Computation*. CRC Press, 2018.

[23] J. Stoer, R. Bulirsch, R. Bartels, W. Gautschi, and C. Witzgall, *Introduction to numerical analysis*, vol. 2. Springer, 1980.

[24] A. Moreschini, M. Bin, A. Astolfi, and T. Parisini, "A generalized passivity theory over abstract time domains," *IEEE Trans. Autom. Control*, 2024.

[25] A. Moreschini, M. Bin, A. Astolfi, and T. Parisini, "On ϱ-passivity," in *Proc. 22nd IFAC World Congr.*, vol. 56, pp. 8556–8561, 2023.

[26] Y. Kawano, A. Moreschini, and M. Cucuzzella, "Krasovskii passivity for sampled-data stabilization and output consensus," *IEEE Transactions on Automatic Control*, pp. 1–16, 2024.

[27] A. Moreschini, S. Monaco, and D. Normand-Cyrot, "Dirac structures for a class of port-Hamiltonian systems in discrete time," *IEEE Trans. Autom. Control*, vol. 69, no. 3, pp. 1999–2006, 2024.

[28] R. I. McLachlan, G. R. W. Quispel, and N. Robidoux, "Geometric integration using discrete gradients," *Phil. Trans. R. Soc. A.*, vol. 357, no. 1754, pp. 1021–1045, 1999.

[29] V. Grimm, R. I. McLachlan, D. I. McLaren, G. Quispel, and C. Schönlieb, "Discrete gradient methods for solving variational image regularisation models," *J. Phys. A: Math. Theor.*, vol. 50, no. 29, p. 295201, 2017.

[30] A. Macchelli, "Trajectory tracking for discrete-time port-Hamiltonian systems," *IEEE Control Syst. Lett.*, vol. 6, pp. 3146–3151, 2022.

[2] The validation of the local minimum can be easily made by perturbing the initial conditions.