Learning with mixed hard/soft pointwise constraints

Giorgio Gnecco, Marco Gori, Stefano Melacci, and Marcello Sanguineti

Abstract—A learning paradigm is proposed and investigated, in which the classical framework of learning from examples is enhanced by the introduction of hard pointwise constraints, i.e., constraints, imposed on a finite set of examples, that cannot be violated. Such constraints arise, e.g., when imposing coherent decisions of classifiers acting on different views of the same pattern. The classical examples of supervised learning, which can be violated at the cost of some penalization (quantified by the choice of a suitable loss function) play the role of soft pointwise constraints. Constrained variational calculus is exploited to derive a representer theorem that provides a description of the functional structure of the optimal solution to the proposed learning paradigm. It is shown that such an optimal solution can be represented in terms of a set of support constraints, that generalize the concept of support vectors. The general theory is applied to deriving the representation of the optimal solution to the problem of learning from hard linear pointwise constraints combined with soft pointwise constraints induced by supervised examples. In some cases, closed-form optimal solutions are obtained.

Index Terms—Parsimony principle, hard and soft constraints, constrained variational calculus, representer theorems, support constraints.

I. INTRODUCTION

In the last few years, the classic framework of learning from examples has been remarkably enriched to better model complex interactions of intelligent agents with the environment. A specific research direction consists in re-framing the learning process in a context described by a collection of constraints, which clearly incorporates learning from examples.

An insight into such an idea was given in [1], but the first systematic study along this direction is [2], where, in addition to some artificial problems, the authors applied the theory to the automatic tagging of bibtex entries. Related studies can be found in [3]–[6]. An application to text categorization that involves supervised learning and prior knowledge was addressed in [7], with experimental results on the CORA database (http://people.cs.umass.edu/~mccallum/data.html). An attempt to perform First Order Logic (FOL) verification can be found in [8]. Kernel-based representations of the optimal solutions have been used also in cases where prior knowledge is not given in terms of logic expressions. Remarkable results were obtained by imposing the classification consistency of different views of the same object [9] and probabilistic constraints [10]. A preliminary study on the benefit deriving from the restriction to convex constraints is [11]. A detailed investigation of the special case of constraints deriving from propositional descriptions was provided in [12]. Several experiments, such as those presented in [2, Section 6], show that the addition of constraints representing prior knowledge on the learning problem under study improves the generalization capability of the learned model, compared to the case in which the same constraints are not taken into account in the problem formulation. In [13], constraints expressed by boundary conditions were also studied, by exploiting tools from calculus of variations and statistical learning theory. Some of the experimental results obtained in the above-mentioned works are summarized in Table I, which emphasizes the importance of making use of known constraints in learning problems.

An in-depth analysis of learning from constraints, stimulated by the above-mentioned papers, suggests the need of a theory to better devise effective algorithms for such a learning framework, and motivates the present work. The framework considered in this paper is a prototype for learning problems in which hard and soft constraints are combined. The distinction between hard and soft constraints is in the way the constraints are embedded into the problem formulation: in the hard case, they restrict the set of feasible solutions, whereas in the soft case their violation is penalized through terms containing a loss function, which are included in the objective of the optimization problem. In particular, here we focus on pointwise constraints (i.e., constraints defined on a finite set of examples, where each element of the set is associated with one...
such constraint), as they model very general knowledge and are often encountered in machine learning problems. However, extensions can be obtained to other settings, as discussed in Section IX.

We investigate learning in a constrained-based environment that takes into account at the same time both hard pointwise constraints and soft pointwise constraints. The former may encode very precise prior knowledge coming from rules, applied, e.g., to a large collection of unsupervised examples (usually, sampled independently according to a possibly unknown probability distribution), whereas the latter may be associated to more unreliable knowledge, corresponding, e.g., to examples more heavily corrupted by noise. The emphasis on the unsupervised examples associated with hard pointwise constraints is motivated by the fact that often they are more widely available than supervised ones, due to a possibly high cost of supervision. In general, dealing with hard constraints is more difficult than dealing with soft ones. So, sometimes optimization problems containing hard constraints are solved by considering sequences of problems in which they are replaced by soft constraints, associated with larger and larger penalties. Under suitable conditions, the optimal solutions to such “relaxed” problems converge to the optimal solutions of the original one [14, Section 10.11]. We remark that this is not the point of view of the present work, in which, instead, we are interested in finding structural properties of the optimal solutions (and in some cases, even closed-form optimal solutions) directly for the original learning problem with hard constraints.

Compared to [2], [9], and [11], here we provide a much more detailed theoretical investigation of the problem of learning in a constrained-based environment. In particular, we investigate issues such as the existence and uniqueness of an optimal solution and we derive necessary optimality conditions in the form of representers theorems. Moreover, the kinds of constraints considered in this paper are different, too: hard and soft pointwise constraints are present simultaneously in the problem formulation, which is given in terms of a Sobolev space. With respect to classical representers theorems, those obtained here provide necessary optimality conditions expressed in the forms of (distributional) partial differential equations. This is due to the choices of (i) a Sobolev space as ambient space of the optimization problem, and (ii) a regularization term expressed in integro-differential form. This makes it possible to exploit our analysis tools from functional analysis, such as Green’s functions, and constrained variational calculus, which are not used in the classical representation theory for machine learning. A similar approach was adopted in [15], but in the absence of hard constraints. The work [13] provides the mathematical foundations of the regularization principles adopted here. We emphasize that our theory provides the classical representers theorems as particular cases.

The paper is organized as follows. In Section II we formulate the problem of learning from examples with mixed hard/soft pointwise constraints in the presence of finite-order differential regularization operators. Section III investigates the existence and uniqueness of an optimal solution. A representers theorem is derived in Section IV. Section V introduces the concept of constraint reaction and discusses some computational issues. In Section VI, extensions to infinite-order differential regularization operators are briefly discussed through a representative example. In Section VII, the general theory is applied to the case of learning in the presence of both linear hard pointwise constraints and soft pointwise constraints on supervised examples. Section VIII deals with two case studies and provides a numerical example. Section IX is a discussion. Finally, the Appendix collects technical definitions and proofs.

II. FORMULATION OF THE LEARNING PROBLEM

We think of an intelligent agent acting on a subset \( \mathcal{X} \) of the perceptual space \( \mathbb{R}^d \) as one implementing a vector function \( f := [f_1, \ldots, f_n]^t \in F \), where \( F \) is a space of functions from \( \mathcal{X} \) to \( \mathbb{R}^n \). Each function \( f_j \) is referred to as a task of the agent.

We assume that prior knowledge is available, modeled by the exact fulfillment of constraints that are expressed in one of the two following ways:

\[
\forall x^{(h)} \in \mathcal{X}_H \subseteq \mathcal{X} : \phi_i(x^{(h)}), f(x^{(h)}) = 0, i = 1, \ldots, m(x^{(h)}), \quad (1)
\]
\[
\forall x^{(h)} \in \mathcal{X}_H \subseteq \mathcal{X} : \tilde{\phi}_i(x^{(h)}), f(x^{(h)}) \geq 0, i = 1, \ldots, \tilde{m}(x^{(h)}), \quad (2)
\]

where the set \( \mathcal{X}_H := \{x^{(1)}, x^{(2)}, \ldots, x^{(|\mathcal{X}_H|)}\} \) is made up of a finite number of points, \( \phi_i \) and \( \tilde{\phi}_i \) are scalar-valued functions, and \( m(x^{(h)}) \) (resp. \( \tilde{m}(x^{(h)}) \)) is the number of constraints of the form (1) (resp., (2)) defined in \( x^{(h)} \). We denote by \( C \) a collection of constraints of the form (1) or (2).

We call (1) hard bilateral pointwise constraints and (2) hard unilateral pointwise constraints. The term “hard” is motivated by the fact that such constraints cannot be violated. Instead, constraints that can be violated - at the cost of some penalization - are called soft constraints. For instance, soft constraints associated with a finite supervised set, corresponding to the classical framework of learning from examples, can be expressed via

\[
V(f(\tilde{y}^{(c)}), \tilde{y}^{(c)})
\]

where \( V : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty] \) is a loss function (i.e., a non-negative function such that, \( \forall z \in \mathbb{R}^n, V(z, z) = 0 \)), \( \tilde{y}^{(c)} \) varies in a finite subset \( X_S := \{\tilde{y}^{(1)}, \tilde{y}^{(2)}, \ldots, \tilde{y}^{(|X_S|)}\} \subseteq \mathcal{X} \), and \( \tilde{y}^{(c)} \in \mathbb{R}^n \) is the associated label. It is worth remarking that, in principle, any constraint can be regarded as hard or soft, depending on the way it is encoded in the optimization problem modeling learning with constraints.

We assume \( \mathcal{X} \) to be either the whole \( \mathbb{R}^d \), or an open, bounded and connected subset of \( \mathbb{R}^d \), with strongly local Lipschitz continuous boundary\(^1\) [16]. In particular, we consider the case in which, for every \( j \in \mathbb{N}_n := \{1, \ldots, n\} \) and some positive integer \( k \), the function \( f_j : \mathcal{X} \to \mathbb{R} \) belongs to the Sobolev space \( W^{k,2}(\mathcal{X}) \), i.e., the subset of \( L^2(\mathcal{X}) \) whose elements \( f_j \) have weak partial derivatives up to the order \( k \) with finite \( L^2(\mathcal{X}) \)-norms. So,

\[
F := W^{k,2}(\mathcal{X}) \times \ldots \times W^{k,2}(\mathcal{X})
\]

\[^{1}\]This assumption is motivated by technical reasons, related to the applicability of some results from the theory of Sobolev spaces.
We assume $k > \frac{k}{2}$ since in such a case, by the Sobolev Embedding Theorem [16, Chapter 4], every element of $W^{k,2}(X)$ has a continuous representative, on which the constraints (1) and (2) can be evaluated unambiguously and $\mathcal{F}$ is a Reproducing Kernel Hilbert Space (RKHS) (see also [17, Chapter 6] for a proof of this fact).

We introduce on $\mathcal{F}$ a seminorm $\|f\|_{p,\gamma}$ by the pair $(P, \gamma)$, where $P := \{P_0, \ldots, P_{\gamma-1}\}$ is a (vector) finite-order differential operator of order $k$ with $l$ components and $\gamma \in \mathbb{R}^n$ is a fixed vector of positive components. Let

$$\|f\|_{P, \gamma}^2 := \sum_{r=0}^{\gamma} \int_X (P_r f_j(x) P_r f_j(x)) dx,$$

and $\mu \geq 0$ be a fixed constant. Given a finite set $X_S \subset X$ and a set of supervised examples $\{f(x^{(\kappa)}), \tilde{y}^{(\kappa)}\}, \kappa = 1, \ldots, |X_S|$, where $|X_S|$ denotes the cardinality of $X_S$, we denote by

$$\mathcal{E}_S(f) := \frac{1}{2} \|f\|_{\mathcal{P}, \mu} + \frac{\mu}{|X_S|} \sum_{\kappa=1}^{\gamma} \sum_{\alpha, \beta} V(f(x^{(\kappa)}), \tilde{y}^{(\kappa)})$$

the objective functional to be minimized (which includes the soft pointwise constraints), whereas we let $\mathcal{F}_C \subset \mathcal{F}$ be the subset of functions that belong to the functional space $\mathcal{F}$ (see (3)) and are compatible with a given collection $C$ of hard pointwise constraints of the form (1) or (2). We state the following optimization problem.

**Problem LMPC (Learning with Mixed hard/soft Pointwise Constraints).** The problem of determining a constrained (local or global) minimizer $f^\ast$ over $\mathcal{F}_C$ is referred to as learning from the hard pointwise constraint collection $C$ and the soft pointwise constraints induced by the functional $\mathcal{E}_S$.

If we choose for $P$ the form used in Tikhonov stabilizing functionals [18], for $n = 1$ and $l = k + 1$ we get

$$\|f\|_{P}^2 = \int_X \sum_{r=0}^{k} \rho_r(x) (D_r f(x))^2 dx,$$

where the function $\rho_r(\cdot)$ is non-negative. $P_r := \sqrt{\rho_r(\cdot)} D_r$, and $D_r$ denotes a differential operator with constant coefficients and containing only partial derivatives of order $r$. An interesting case corresponds to the choices $\rho_r(\cdot) := \rho_r \geq 0$, $D_{2r} := \Delta^r = \nabla^{2r}$, and $D_{2r+1} := \nabla \Delta^r = \nabla \nabla^{2r}$, where $\Delta$ denotes the Laplacian operator and $\nabla$ the gradient, with the additional condition $D_0 f = f$ (see [15], [19]). For instance, $D_1 = \nabla$, $D_2 = \Delta = \nabla^2$, and $D_3 = \Delta^2 = \nabla^{2r}$.

According to (4), when $n > 1$ the operator $P$ acts separately on all the components of $f$, i.e., $P f := \{P f_1, P f_2, \ldots, P f_n\}$. Note that in this case we have overloaded the notation and used the symbol $P$ both for the (matrix) differential operator acting on $f$ and for the (vector) differential operator acting on the components. We focus on the case in which the operator $P$ is invariant under spatial shift and has constant coefficients. We use the following notation. For a function $u$ and a multi-index $\alpha$ with $d$ non-negative components $\alpha_j$, we write $D^\alpha u$ to denote $\frac{\partial^{\alpha}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} u$, where $|\alpha| := \sum_{j=1}^{d} \alpha_j$.

So, the generic component $P_i$ of the operator $P$ takes on the expression $P_i = \sum_{|\alpha| \leq k} b_i,\alpha D^\alpha$, where the $b_i,\alpha$’s are suitable real coefficients. Then, the formal adjoint of $P$ is defined as the operator $P^* := \{P_0^*, \ldots, P_{\gamma-1}^*\}$ whose $i$-th component $P_i^*$ has the form $P_i^* := \sum_{|\alpha| \leq k} (\alpha_i) b_i,\alpha D^\alpha$. Finally, we define the operators $L := \{P_i^*\} P$ and, using again an overloaded notation, $\gamma L := \{\gamma_1 L, \ldots, \gamma_n L\}$.

### III. Existence and Uniqueness of an Optimal Solution to Problem LMPC

The following theorem provides sufficient conditions for the existence and uniqueness of a global minimizer for Problem LMPC.

**Theorem III.1.** If $\| \cdot \|_{P, \gamma}$ is a RKHS norm on $\mathcal{F}$, the loss function $V$ is convex and continuous, and the set $\mathcal{F}_C$ is nonempty, closed, and convex, then Problem LMPC has a unique global optimal solution.

For $n = 1$, examples of problems for which $\| f \|_{P, \gamma}$ is a RKHS norm on $\mathcal{F}$ are provided in [13] for the class of rotationally-symmetric differential operators defined therein. Such examples extend readily to the case $n > 1$, as the operator $P$ acts on each component of $f$ separately. Examples of problems for which $\mathcal{F}_C$ is nonempty, closed, and convex arise when the hard pointwise constraints are bilateral and the associated constraint functions $\phi_i(\cdot, \cdot)$ are linear with respect to the second vector-valued argument. For hard unilateral pointwise constraints, the assumptions of closedness and convexity of $\mathcal{F}_C$ hold when the associated constraint functions $\tilde{\phi}_i(\cdot, \cdot)$ are concave with respect to the second vector-valued argument and continuous.

### IV. The Representer Theorem for Problem LMPC

The next theorem provides a representation for a (local or global) optimal solution $f^\ast$ to Problem LMPC. Without a significant loss of generality, we assume that the sets $\mathcal{X}_S$ and $\mathcal{X}_H$ where the soft and hard pointwise constraints are defined, resp., have empty intersections. Before stating the theorem, we introduce the following notations and definitions. For two vector-valued functions $u^{(1)}$ and $u^{(2)}$ of the same dimension, $u^{(1)} + u^{(2)}$ denotes the vector-valued function $v$ whose first component is the convolution of the first components of $u^{(1)}$ and $u^{(2)}$, the second component is the convolution of the second components of $u^{(1)}$ and $u^{(2)}$, and so on, i.e., $u_j := (u_j^{(1)} + u_j^{(2)})$ for each index $j$. A hard constraint $\tilde{\phi}_j(x^{(\kappa)}), f(x^{(\kappa)}) \geq 0$ is active at local optimality iff $\tilde{\phi}_j(x^{(\kappa)}), f(x^{(\kappa)}) = 0$, otherwise it is inactive at local optimality. We denote by $m := \max_{x^{(\kappa)} \in \mathcal{X}_H} m(x^{(\kappa)})$ the maximum number of constraints defined on a point $x^{(\kappa)} \in \mathcal{X}_H$. Finally, we recall that a free-space Green’s function associated to a differential operator $O$ is a solution $g$ to the distributional differential equation $O g = \delta$, where $\delta$ is the Dirac distribution centered at the origin.

---

2Boundary-constrained Green’s functions for soft-constrained learning problems were investigated, together with free-space Green’s functions, in [13].
Theorem IV.1. (Representer theorem for mixed hard/soft pointwise constraints). Let us consider an agent that minimizes the functional (4) on $F$ consistently with a given set of bilateral pointwise constraints given by
\[
\forall x^{(h)} \in X_H, \quad \phi_i(x^{(h)}, f(x^{(h)})) = 0, \quad i = 1, \ldots, m(x^{(h)}),
\]
where $\phi_i \in C^1(\mathbb{X} \times \mathbb{R}^n)$, $F$ be any constrained local minimizer of the functional (4). Assume that for any $x^{(h)}$ in the finite set $X_H$ one has $m(x^{(h)}) \leq n$ and one can find a permutation of the $n$ functions $f^o$ such that the Jacobian matrix
\[
\frac{\partial (\phi_1, \ldots, \phi_m(x^{(h)}))}{\partial (f^o_1, \ldots, f^o_m(x^{(h)}))},
\]
evaluated in $x^{(h)}$, is non-singular. Let also $X_H \cap X_S = \emptyset$. Then, the following hold.

(i) There exist $|X_H|$ vectors $\lambda^i \in \mathbb{R}^{m(x^{(h)})}$, each one associated with a point $x^{(h)} \in X_H$, such that, in addition to the above constraints, the function $f^o$ satisfies on $X$ the Euler-Lagrange equations
\[
\gamma L f^o(x) + \sum_{h=1}^{|X_H|} \lambda^i \delta(x - x^{(h)}) \nabla f \phi_i(x^{(h)}, f^o(x^{(h)})) + \frac{\mu}{|X_S|} \delta(x - x^{(k)}) \nabla f \phi_k(x^{(k)}, f^o(x^{(k)})) = 0,
\]
where $\gamma L := [\gamma_1, \ldots, \gamma_n L]$ is a spatially-invariant operator, $\lambda^i$ is the $i$-th component of the vector $\lambda^i$, and $\nabla f \phi_k$ is the gradient w.r.t. the second vector argument $f$ of the function $\phi_k$.

(ii) Let $\gamma^{-1} g := \gamma_1^{-1} g_1, \ldots, \gamma_n^{-1} g_n$ and let us extend the vectors $\lambda^i$ to vectors in $\mathbb{R}^n$ (still denoted by $\lambda^i$), by adding as many 0’s as required. If $X = \mathbb{R}^d$, $L$ is invertible on $W^{h,2}(X)$, and there exists a free-space Green’s function $g$ of $L$ that belongs to $W^{h,2}(X)$, then $f^o$ has the representation
\[
f^o(\cdot) = \gamma^{-1} g(\cdot) \left( \sum_{i=1}^{m} \omega^H_i(\cdot) + \sum_{k=1}^{m} \omega^S_k(\cdot) \right),
\]
where
\[
\omega^H_i(\cdot) := - \sum_{h=1}^{x^{(h)}} \lambda^i \delta(- x^{(h)}) \nabla f \phi_i(x^{(h)}, f^o(x^{(h)})) ,
\]
\[
\omega^S_k(\cdot) := - \frac{\mu}{|X_S|} \delta(- x^{(k)}) \nabla f \phi_k(x^{(k)}, f^o(x^{(k)}), \bar{y}^{(k)}) ,
\]
and
\[
\omega(\cdot) := \left( \sum_{i=1}^{m} \omega^H_i(\cdot) + \sum_{k=1}^{m} \omega^S_k(\cdot) \right).
\]

(iii) For the case of unilateral pointwise constraints given by
\[
\forall x^{(h)} \in X_H, \quad \phi_i(x^{(h)}, f(x^{(h)})) \geq 0, \quad i = 1, \ldots, m(x^{(h)}),
\]
where $\phi_i \in C^1(\mathbb{X} \times \mathbb{R}^n)$, (i) and (ii) still hold if one requires the non-singularity of (5) to hold on the constraint that are active at local optimality (of course, replacing $\phi_i$ by $\hat{\phi}_i$ and $m(x^{(h)})$ by $\hat{m}(x^{(h)})$). Moreover, each Lagrange multiplier $\lambda^i$ is non-positive and equal to 0 when the correspondent constraint is inactive at local optimality.

Remark IV.2. The assumption $g \in W^{h,2}(X)$ in Theorem IV.1 can be enforced, e.g., under the assumptions of Theorem 3 in [13] about the smoothness properties of free-space Green’s functions.

Remark IV.3. Theorem IV.1 can be extended to deal with the case of certain non-smooth convex loss functions, like the hinge loss (see [20], [21] for a proof in the presence of soft pointwise constraints only). In this case, subgradients of the loss function would appear in the correspondent modification of the Euler-Lagrange equations (6). Moreover, Theorem IV.1 can be also extended to the case in which one has bilateral and unilateral hard pointwise constraints simultaneously (in addition to the soft pointwise constraints).

V. CONSTRAINT REACTIONS AND COMPUTATIONAL ISSUES

A. Constraint reactions

The next definition formalizes a concept that plays a basic role in our study.

Definition V.1. The distributions $\omega^H_i$ and $\omega^S_k$ in Theorem IV.1 are called the reaction of the $i$-th hard pointwise constraint and the reaction of the $k$-th soft pointwise constraint, resp., whereas $\omega$ is the overall reaction of the given constraints.

We emphasize that the (hard or soft) reaction of a constraint is a concept associated with the constrained local minimizer $f^o$ of the functional (4). In particular, two different constrained local minimizers $f^o$ may be associated with different reactions. A similar remark holds true for the overall reaction of the constraints. Loosely speaking, under the assumptions of Theorem IV.1 (i), (ii), the reaction of each (hard or soft) constraint provides the way under which such constraint contributes to the expansion of $f^o$, as specified by (7). So, in this case solving Problem LMPC is reduced to finding the reactions of the constraints.

The next proposition states that, under suitable assumptions, the reactions of the constraints are uniquely determined by the constrained local minimizer $f^o$. Although the latter is in general not known a-priori, this is a structural property of local constrained minimizers, which can be useful, e.g., when searching for them.

Proposition V.2. Under the assumptions of Theorem IV.1 (ii),(iii), the reactions of the constraints are uniquely determined by the constrained local minimizer $f^o$.

Under the assumptions of Theorem IV.1 (ii),(iii), a constrained local minimizer $f^o$ for Problem LMPC can be obtained merely by the knowledge of the reactions associated with the constraints whose reactions are different from 0, that we call support constraints. For the case of soft pointwise
constraints, this coincides with the notion of support vectors used in kernel methods [22], since these are just the data points \( x^{(k)} \) whose associated soft constraint has a reaction that is different from 0. The connection with kernel methods arises also because, under quite general conditions, the free-space Green's function \( g \) associated with the operator \( L \) is a kernel of a RKHS (see, e.g., [13] and the references therein). In such cases, for suitable choices of the constraints, various kernel methods are obtained as particular instances of the proposed learning framework. For instance, in the case of soft constraints expressed by the quadratic loss, and in the absence of hard constraints, one re-obtains support vector machines (see [20], [21] for a derivation of this connection).

The emergence of constraints whose reactions are identically 0 at local optimality is particularly evident for the case of hard unilateral pointwise constraints. Indeed, under the assumptions of Theorem IV.1 (iii), a hard constraint \( \phi \) that is inactive at local optimality for all \( f^{(h)} \in X_H \) is associated with a Lagrange multiplier \( \lambda \) that is identically 0, so its reaction is identically 0, too.

It is interesting to discuss the case of an instance of Problem LMPC in which one of the hard pointwise constraints is redundant\(^5\), in the sense that the fulfillment of all the other hard pointwise constraints guarantees its fulfillment, too. Without loss of generality, such a redundant constraint can be discarded from the problem formulation and, provided that the assumptions of Theorem IV.1 (ii),(iii) hold, one still has the representation (7) for the constrained local minimizer \( f^o \), where the Lagrange multiplier associated with the redundant constraint is 0 (so, also the reaction for that constraint is 0).

\(^5\)Of course, redundant hard constraints can appear only if the assumption on the invertibility of (5) is violated.

B. Computational issues

Although Problem LMPC has been reduced to finding the reactions of the constraints, a serious issue in the application of this recipe is that it requires the knowledge of the Lagrange multipliers. Indeed, in addition to the fact that \( f^o \) has to satisfy the Euler-Lagrange equations (6), the entire collection of hard pointwise constraints guarantees its fulfillment, too. Without loss of generality, such a redundant constraint can be discarded from the problem formulation and, provided that the assumptions of Theorem IV.1 (ii),(iii) hold, one still has the representation (7) for the constrained local minimizer \( f^o \), where the Lagrange multiplier associated with the redundant constraint is 0 (so, also the reaction for that constraint is 0).

VI. EXTENSION TO INFINITE-ORDER DIFFERENTIAL OPERATORS

Theorem IV.1 shows circumstances under which an optimal solution to Problem LMPC can be represented as a linear combination of kernel functions obtained by translations of a kernel \( g \) that is a free-space Green's function of the operator \( L \). A particularly interesting case is the one in which \( g \) is a Gaussian, since the latter is often used as a kernel in kernel methods. However, for any finite-order differential operator \( L \), a Gaussian function cannot be its free-space Green's function, as in this case the first hand-side of the distributional differential equation \( Lg = \delta \) would be smooth, in contrast with the right-hand side. Nevertheless, if the operator \( P \) is replaced by an infinite-order differential operator with constant coefficients, then it is possible to get a Gaussian as a free-space Green's function. Indeed, in [27] it was shown that the

\(^5\)See https://sites.google.com/site/semanticbasedregularization/home/software for a software simulator that allows one to find the reactions for various kinds of soft pointwise constraints, in real-world problems.
Gaussian kernel with mean 0 and variance $\sigma^2$ is a free-space Green’s function for the differential operator of infinite order

$$L = L(\infty) := \sum_{i=0}^{\infty} (-1)^i a_i \nabla^{2i},$$

(10)

where $a_i := \frac{\sigma^{2i}}{2^{2i}}$ (see also [15, Section 5.1.2]). This can be proved via Fourier transforms, by observing that the Fourier transforms of $(-1)^i \nabla^{2i}$ and $\delta$ are, resp., the function $\|2\pi \xi\|^{2i}$ and the constant function 1. Then, with such coefficients $a_i$, by exploiting the Taylor series $\exp(t) = \sum_{i=0}^{\infty} \frac{t^i}{i!}$, we get

$$\hat{g}(\xi) = \hat{g}(\infty)(\xi) := \left(\sum_{i=0}^{\infty} a_i \|2\pi \xi\|^{2i}\right)^{-1} = \exp\left(-\frac{\sigma^2\|2\pi \xi\|^2}{2}\right),$$

whose inverse Fourier transform is the Gaussian

$$g(x) = g(\infty)(x) := \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right).$$

(11)

The simplest way to apply the results of the previous sections to this case consists in replacing, for some positive integer $k$, the infinite-order operator (10) with its truncation

$$L = L(k) := \sum_{i=0}^{k} (-1)^i a_i \nabla^{2i},$$

(12)

which is a finite-order differential operator. Then, the free-space Green’s function $g(x) = g(k)(x)$ associated with $L(k)$ is the inverse Fourier transform of

$$\hat{g}(\xi) = \hat{g}(k)(\xi) := \left(\sum_{i=0}^{k} a_i \|2\pi \xi\|^{2i}\right)^{-1},$$

(13)

which for $k$ sufficiently large is a good approximation of the Gaussian kernel. Of course, a similar method can be used for other infinitely-smooth kernels that are free-space Green’s functions of infinite-order differential operators. Otherwise, direct extensions of the results to infinite-order differential operators may be obtained by setting Problem LMPC on Sobolev spaces of infinite order [28].

VII. COMBINING HARD LINEAR AND SOFT QUADRATIC POINTWISE CONSTRAINTS

In this section, we consider the application of Theorem IV.1 to the case of hard linear pointwise constraints combined with soft quadratic pointwise constraints. We first address bilateral hard linear pointwise constraints, then unilateral ones.

6Of course, differential operators $P$ associated with $L$ by $L := (P^*)^\dagger P$ can be constructed in several ways.

7Here, we use the unitary definition of the Fourier transform in terms of the frequency vector $\xi \in \mathbb{R}^d$, i.e., $\hat{g}(\xi) := \int g(x) \exp(-2\pi i \xi \cdot x) dx$ and $g(x) := \int \hat{g}(\xi) \exp(i(2\pi \xi \cdot x)) dx$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^d$. A. Hard bilateral constraints

A particularly interesting case of application of Theorem IV.1 (ii) is the one in which one has hard linear bilateral pointwise constraints of the form

$$\forall x^{(h)} \in \mathcal{X}_H, \phi_i(x^{(h)}, f(x^{(h)})) = \begin{cases} A(x^{(h)}) f(x^{(h)}) - b(x^{(h)}) & i = 1, \ldots, m(x^{(h)}) \end{cases} = 0,$$

(14)

where $A(x^{(h)})$ and $b(x^{(h)})$ are matrices and column vectors of dimensions $m(x^{(h)}) \times n$ and $m(x^{(h)})$, resp., and soft quadratic constraints expressed by the quadratic loss

$$V(f(\tilde{x}^{(c)}), \tilde{y}^{(c)}) := \frac{1}{2} \sum_{j=1}^{n} (f_j(\tilde{x}^{(c)}) - \tilde{y}_j^{(c)})^2.$$  

(15)

The first kind of constraints express a-priori knowledge, whereas the second ones model the classical way of dealing softly with supervised examples.

For this case, denoting by $A_i^r(x^{(h)})$ the transpose of the $i$-th row of $A(x^{(h)})$, we get the following expressions for the hard and soft constraint reactions:

$$\omega^H_i(\cdot) := -\sum_{h=1}^{\min(|\mathcal{X}_H|)} \lambda^h_i \delta(\cdot - x^{(h)}) A_i^r(x^{(h)}),$$

(16)

$$\omega^S_{\kappa}(\cdot) := -\frac{\mu}{|\mathcal{X}_S|} \delta(\cdot - \tilde{x}^{(\kappa)}) \left( f(\tilde{x}^{(\kappa)}) - \tilde{y}^{(\kappa)} \right),$$

(17)

Then, a function $f^o$ that satisfies all the hard pointwise constraints and the Euler-Lagrange equations (6) is obtained by solving the linear system of equations

$$\left\{ \begin{array}{l} A(x^{(h)}) f^o(x^{(h)}) = b(x^{(h)}), \quad h = 1, \ldots, |\mathcal{X}_H|, \\ f^o(x^{(h)}) = \gamma^{-1} g(x) * \omega(x) |_{x=x^{(h)}}, \quad h = 1, \ldots, |\mathcal{X}_H|, \\ f^o(x^{(\kappa)}) = \gamma^{-1} g(x) * \omega(x) |_{x=\tilde{x}^{(\kappa)}}, \quad \kappa = 1, \ldots, |\mathcal{X}_S|, \end{array} \right.$$  

where $\omega$ is the overall reaction of the constraints obtained by formulas (9), (16), and (17). Notice that the linear system has

$$m(|\mathcal{X}_H| + |\mathcal{X}_S|) + \sum_{h=1}^{\min(|\mathcal{X}_H|)} m(x^{(h)})$$  

unknowns and the same number of equations (which are linearly independent when the assumptions of Theorem III.1 hold). Finally, once the linear system above has been solved, $f^o$ has the representation

$$f^o(x) = \gamma^{-1} g(x) * \omega(x), \forall x \in \mathcal{X} = \mathbb{R}^d.$$  

(18)

We conclude discussing the admissibility of such a function $f^o$. Since $g \in \mathcal{W}^{k,2}(\mathbb{R}^d)$, by the representation (18) we get $f^o \in \mathcal{F}$. Then, $E_s(f^o)$ is finite, and $f^o$ is a constrained global minimizer, too (thanks to the convexity of the problem).
B. Hard unilateral constraints

As a variation of the example in the previous subsection, one can consider the case in which the hard bilateral pointwise constraints of the form (14) are replaced by hard unilateral pointwise constraints of the form

\[ \forall x^{(h)} \in X_H, \quad \phi_i(x^{(h)}, f(x^{(h)})) = \left( A(x^{(h)}) f(x^{(h)}) - b(x^{(h)}) \right)_i \geq 0, \quad i = 1, \ldots, m(x^{(h)}) \]

(19)

In this case, the optimality conditions obtained from Theorem IV.1 (iii) can be translated into the system

\[
\begin{align*}
A(x^{(h)}) f_i(x^{(h)}) &\geq b(x^{(h)}) , \\
\lambda_i^{(h)} &\leq 0 , \\
f_i^{\mu}(x^{(h)}) &= \gamma^{-1} g(x) * \omega(x)_{|x=x^{(h)}} , \\
f_i^{\nu}(x^{(h)}) &= \gamma^{-1} g(x) * \omega(x)_{|x=x^{(h)}} ,
\end{align*}
\]

for \( h = 1, \ldots, |X_H| \) and \( \kappa = 1, \ldots, |X_S| \), where \( \phi \) denotes the Hadamard (i.e., entry-wise) product, and the inequalities are to be interpreted component-wise. Again, the system above has a solution whenever the assumptions of Theorem III.1 hold.

VIII. CASE STUDIES

Here we discuss the application of the results of Section VII to two case studies.

A. Hard bilateral constraints on supervised examples

As a particular instance of the case of hard linear bilateral pointwise constraints combined with soft quadratic pointwise constraints, here we address the following variation of the classical supervised learning framework.

We are interested in enforcing the hard fulfillment of some constraints on supervised examples, instead of merely their soft fulfillment. More precisely, we consider hard bilateral pointwise constraints given by

\[ \forall x^{(h)} \in X_H, \quad \phi_i(x^{(h)}, f(x^{(h)})) = f(x^{(h)}) - y_i^{(h)} = 0, \quad i = 1, \ldots, m(x^{(h)}) = n . \]

(21)

So, for each supervised example \( x^{(h)} \) associated with a hard bilateral constraint of the form (21), one has \( A(x^{(h)}) = I_n \), the identity matrix of dimension \( n \times n \). Notice that in this case, for each \( x^{(h)} \in X_H \) the number of hard bilateral pointwise constraints \( m(x^{(h)}) \) is equal to \( n \), which is the number of components of \( f \). We can promptly see that the non-singularity of the Jacobian matrix (5) - which is required for the application of Theorem IV.1 - holds true, since for this case (5) is the identity matrix. In addition to the hard bilateral pointwise constraints (21), other soft pointwise constraints expressed by the quadratic loss (15) may or may not be present in the formulation of the learning problem.

The motivation for this case study is that sometimes, in machine learning, some supervised examples are “more important” than others (e.g., because the latter may be more heavily corrupted by noise). So, one can associate the first ones to hard bilateral constraints (thus, preventing their violation), and the others to soft constraints (allowing their violation, at the cost of some penalization).

Now, we specialize the linear system (18) to such a case study. To simplify some of the next formulas, we assume that each component of \( \gamma \) is equal to the same positive scalar \( \gamma \) (this assumption may be relaxed). In order to use a compact notation, we define the following matrices:

\[
F_H := \left( f_H^{(1)}, f_H^{(2)}, \ldots, f_H^{(n)} \right) := \left( f_1^{(1)}, f_2^{(1)}, \ldots, f_n^{(1)} \right) \cdot 1, \quad F_S := \left( f_S^{(1)}, f_S^{(2)}, \ldots, f_S^{(n)} \right) := \left( f_1^{(2)}, f_2^{(2)}, \ldots, f_n^{(2)} \right), \quad Y_H := (y_H^{(1)}, y_H^{(2)}, \ldots, y_H^{(n)}) := \left( y_1^{(1)}, y_2^{(1)}, \ldots, y_n^{(1)} \right) \cdot 1, \quad Y_S := (y_S^{(1)}, y_S^{(2)}, \ldots, y_S^{(n)}) := \left( y_1^{(2)}, y_2^{(2)}, \ldots, y_n^{(2)} \right), \quad \Lambda := (\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(n)}) := \left( \lambda_1^{(1)}, \lambda_2^{(1)}, \ldots, \lambda_n^{(1)} \right) \cdot 1, \quad \Lambda := (\Lambda^{(1)}, 1, \ldots, \Lambda^{(n)}) := \left( \lambda_1^{(1)}, \lambda_2^{(1)}, \ldots, \lambda_n^{(1)} \right).
\]

The matrices \( G_{HH} \in \mathbb{R}^{|X_H| \times |X_H|}, G_{HS} \in \mathbb{R}^{|X_H| \times |X_S|}, G_{SH} \in \mathbb{R}^{|X_S| \times |X_H|} \), and \( G_{SS} \in \mathbb{R}^{|X_S| \times |X_S|} \) are defined in terms of their elements as \( G_{HH,H,P,Q} := g(x_p - x_Q) \), \( G_{HS,H,P,Q} := g(x_p - x_Q) \), \( G_{SH,H,P,Q} := g(x_p - x_Q) \), and \( G_{SS,H,P,Q} := g(x_p - x_Q) \). By making the dependencies on \( \Lambda, F(H), \) and \( F(S) \) explicit, the function \( f^o \) can be expressed as \( f^o(\Lambda, F(H), F(S)) \), via obvious substitutions into (16), (17), and (18). So, one has to find suitable values for the unknowns \( \Lambda, F(H), F(S) \). These are found by solving the three groups of equations of the linear system (18), which can be expressed, in terms of \( \Lambda, F(H), F(S) \), as

\[
\begin{align*}
F_H &= Y_H, \\
F_H &= -\gamma^{-1} \left( G_{HH} \Lambda + \frac{\mu}{|X_S|} G_{HS} (F_S - Y_S) \right), \\
F_S &= -\gamma^{-1} \left( G_{SH} \Lambda + \frac{\mu}{|X_S|} G_{SS} (F_S - Y_S) \right).
\end{align*}
\]
Now, the linear system above can be decomposed into the $n$ decoupled linear systems
\[
\begin{aligned}
F_H^{(j)} &= Y_H^{(j)}, \\
F_H^{(j)} &= -\bar{\gamma}^{-1} \left[ G_{HH}\Lambda^{(j)} + \frac{\mu}{|x_i|} G_{HS}(F_S^{(j)} - \bar{Y}_S^{(j)}) \right], \\
\bar{F}_S^{(j)} &= -\bar{\gamma}^{-1} \left[ G_{SH}\Lambda^{(j)} + \frac{\mu}{|x_i|} G_{SS}(\bar{F}_S^{(j)} - \bar{Y}_S^{(j)}) \right],
\end{aligned}
\]
\[j = 1, \ldots, n\] (this decoupling is been allowed by the form (21) of the hard bilateral constraints and may hold also for other choices of the linear hard bilateral constraints). Then, each of the linear systems above can be written as
\[
\begin{aligned}
F_H^{(j)} &= Y_H^{(j)}, \\
\left( \begin{array}{c}
\bar{F}_S^{(j)} \\
\bar{F}_S^{(j)}
\end{array} \right) &= \left( \begin{array}{cc}
\bar{\gamma}^{-1}G_{HH} & \bar{\gamma}^{-1} \frac{\mu}{|x_i|} G_{HS} \\
\bar{\gamma}^{-1}G_{SH} & \bar{\gamma}^{-1} \frac{\mu}{|x_i|} G_{SS} + I_{|x_i|}
\end{array} \right) \left( \begin{array}{c}
\Lambda^{(j)} \\
\bar{F}_S^{(j)}
\end{array} \right).
\end{aligned}
\]
\[(22)\]
Finally, the system matrix
\[
\left( \begin{array}{cc}
\bar{\gamma}^{-1}G_{HH} & \bar{\gamma}^{-1} \frac{\mu}{|x_i|} G_{HS} \\
\bar{\gamma}^{-1}G_{SH} & \bar{\gamma}^{-1} \frac{\mu}{|x_i|} G_{SS} + I_{|x_i|}
\end{array} \right)
\]
\[(23)\]
can be also expressed as
\[
\left( \begin{array}{cc}
\bar{\gamma}^{-1}G_{HH} & \bar{\gamma}^{-1} \frac{\mu}{|x_i|} G_{HS} \\
\bar{\gamma}^{-1}G_{SH} & \bar{\gamma}^{-1} \frac{\mu}{|x_i|} G_{SS} + I_{|x_i|}
\end{array} \right) \left( \begin{array}{c}
0 \\
0
\end{array} \right) = \left( \begin{array}{c}
\bar{I}_{|x_i|} \\
0
\end{array} \right)
\]
\[(24)\]
which is clearly invertible whenever the Gram matrix
\[
\bar{\gamma}^{-1} \left( \begin{array}{cc}
G_{HH} & G_{HS} \\
G_{SH} & G_{SS}
\end{array} \right)
\]
is invertible (which happens, e.g., when the Green’s function $g$ is positive definite: two such cases are the Gaussian and its approximations considered in Section VI). Concluding, under such an assumption one can solve the linear system (22) by obtaining first the vector $F_H^{(j)}$ from the equation $F_H^{(j)} = Y_H^{(j)}$, then inverting the system matrix (23) to find $\Lambda^{(j)}$ and $\bar{F}_S^{(j)}$.

**Remark VIII.1.** The computations above extend to the degenerate cases in which (i) there are no hard bilateral constraints, but only soft constraints, or (ii) there are no soft constraints but only hard bilateral constraints. For instance, in the case (i) (which corresponds to kernel ridge regression, basically equivalent to least-square support vector machines [29, Section 9.3.1]), the linear system (22) becomes
\[
\left( \begin{array}{c}
\bar{F}_S^{(j)}
\end{array} \right) = \bar{\gamma}^{-1} \left[ G_{SS} + I_{|x_i|} \right] \left( \begin{array}{c}
\bar{F}_S^{(j)}
\end{array} \right),
\]
\[(25)\]
whereas in the case (ii) it becomes
\[
\left\{ \begin{array}{l}
F_H^{(j)} = Y_H^{(j)}, \\
\bar{F}_S^{(j)} = -\bar{\gamma}^{-1} G_{HH}\Lambda^{(j)} = -F_H^{(j)}.
\end{array} \right.
\]
\[(26)\]
Compared to kernel ridge regression, the case of hard linear bilateral pointwise constraints combined with soft quadratic pointwise constraints, and the case (ii) above, “give more importance” to some constraints, enforcing them in a hard way. Interestingly, as shown by formulas (22) and (26), closed-form optimal solutions are obtained even in these cases (only matrix inversions are required to find them).

**A numerical example.** To illustrate the difference between learning from soft constraints only, learning with mixed hard/soft constraints, and the limit case in which there are only hard constraints, we consider a simple instance with $d = 2$ and $n = 1$. In the example shown in Fig. 2, the set $\{(-1, 0), (-2, -1), (-3, -4), (0, 3)\}$ of positive supervised examples associated with the label $y^{(h)} = 1$ is given, together with the set $\{(3, -3), (2, -2), (3, 2), (0, -4)\}$ of negative supervised examples associated with the label $y^{(h)} = -1$. The Gaussian kernel (11) - whose application is made possible by the discussion in Section VI - is used, with the value $\sigma = 1$, whereas the choices $\mu = 30$ and $\gamma = 1$ are made in the formulation of Problem LMPC. Fig. 2 (a) shows the optimal solution $f^*$ to Problem LMPC obtained when all the supervised examples (represented by crosses) are associated with soft constraints expressed by the quadratic loss (15) (this is the case of kernel ridge regression). Fig. 2 (b) shows the optimal solution obtained when the negative supervised examples (represented by crosses) are still associated with soft constraints expressed by the quadratic loss (15), but the positive supervised examples (represented by circles) are associated with hard bilateral constraints of the form (21). Finally, Fig. 2 (c) shows the optimal solution obtained when all the supervised examples (represented by circles) are associated with hard bilateral constraints of the form (21).

**B. Hard unilateral constraints of non-negativeness**

As a particular instance of the case of hard linear unilateral pointwise constraints combined with soft quadratic pointwise constraints, we consider here the remarkable case in which the hard linear unilateral pointwise constraints (19) take on the specific form
\[
\forall x^{(h)} \in X_H, \quad \hat{f}_i(x^{(h)}), f(x^{(h)}) = f_i(x^{(h)}) \geq 0, \quad i = 1, \ldots, m(x^{(h)}) = n,
\]
\[(27)\]
which has applications, e.g., whenever the components of $f$ represent mass or probability densities. Likewise in Section VIII-B, we assume that, in addition to such hard constraints, one has soft constraints on supervised examples, expressed by the quadratic loss (15). Using the same notation, the system (20) can be written in the form
\[
\begin{aligned}
F_H &\geq 0, \\
\Lambda &\leq 0, \\
F_H \circ \Lambda &\leq 0, \\
\bar{F}_S &\leq 0.
\end{aligned}
\]
\[(28)\]
Again, a decoupling is possible, so (28) provides the $n$ decoupled systems.
for $j = 1, \ldots, n$. Then, expressing $\tilde{F}^{(j)}_S$ as an affine function of $\Lambda^{(j)}$, we get

$$\tilde{F}^{(j)}_S(\Lambda^{(j)}) = -\left( I_{|X_S|} + \gamma^{-1} \frac{\mu}{|X_S|} G_{SS} \right)^{-1} \gamma^{-1} \left[ G_{SH} \Lambda^{(j)} - \frac{\mu}{|X_S|} G_{SS} \tilde{Y}^{(j)}_S \right], \quad (30)$$

and similarly, also $F^{(j)}_H$ can be expressed as an affine function of $\Lambda^{(j)}$, by the replacement $\tilde{F}^{(j)}_S = F^{(j)}_S(\Lambda^{(j)})$ in

$$F^{(j)}_H = -\gamma^{-1} \left[ G_{SH} \Lambda^{(j)} + \frac{\mu}{|X_S|} G_{SS} (\tilde{F}^{(j)}_S - \tilde{Y}^{(j)}_S) \right]. \quad (31)$$

Finally, with the obtained expressions of $F^{(j)}_H(\Lambda^{(j)})$ and $\tilde{F}^{(j)}_S(\Lambda^{(j)})$, each system (29) can be written as

$$\begin{cases} F^{(j)}_H(\Lambda^{(j)}) \geq 0, \\
\Lambda^{(j)} \leq 0, \\
F^{(j)}_H(\Lambda^{(j)}) \circ \Lambda^{(j)} = 0. \end{cases} \quad (32)$$

Interestingly, taking into account all the $n$ decoupled systems, the first group of inequalities of (32) expresses the requirement of searching for a feasible solution to this instance of Problem LMPC. The second group imposes the feasibility of the optimal solution to its dual problem [14, Section 8.6], whereas the third group states the Karush-Kuhn-Tucker complementarity conditions. Recall that the dual problem associated with this instance of Problem LMPC consists in maximizing the dual objective function $d(\Lambda)$ over the set $\Lambda \geq 0$, where $d(\lambda)$ is obtained by replacing $f = f(S, \Lambda, F^{(H)}(\Lambda), \tilde{F}^{(S)}(\Lambda))$ into the Lagrangian functional $E_S(f) + \sum_{i=1}^{n} |\lambda_i| \sum_{j=1}^{n} \lambda^{(j)} \in (x^{(j)}_i)$. Since $E_S(f)$ is quadratic and $F^{(H)}(\Lambda), \tilde{F}^{(S)}(\Lambda)$ have affine dependencies on $\Lambda$, such a dual problem is a Quadratic Programming (QP) problem with linearly inequality constraints. We refer the reader to [23] for various techniques developed to solve such a kind of problems. We also observe that, thanks to the decoupling the objective function $d(\Lambda)$ of the dual problem takes on the separable form $d(\Lambda) = \sum_{j=1}^{n} d^{(j)}(\Lambda^{(j)})$, where the functions $d^{(j)}(\Lambda^{(j)})$ are the dual objective functions of the primal problems in the decomposition.

**Two numerical examples.** As a first numerical example, we consider $d = 2$ and $n = 1$, the Gaussian kernel with $\sigma = 2$, and the choices $\mu = 30$, $\gamma = 1$. In Fig. 3 (a), there are two negative supervised examples $(-3,0)$ and $(5,0)$, with labels, resp., equal to $-3.772$ and $-1$, which are associated with soft constraints expressed by the quadratic loss (15). The supervised examples together with their labels are represented by crosses; they are dealt with in a soft way. In Fig. 3 (b), the unsupervised example $(-0,0)$ has been added, for which a hard unilateral constraint of the form (27) has been imposed. In the figure, we represent such a point by a star (located at $y = 0,2$ instead of $y = 0$, in order to make it visible). In this case, after solving the system (32), the Lagrange multiplier associated with the hard unilateral constraint is different from 0 (it is equal to $-3$), and there is a nonzero constraint reaction associated with such a hard unilateral constraint. This

![Figure 2: Optimal solutions to Problem LMPC for the numerical example presented in Section VIII-A. See the text for the explanations.](image-url)
is illustrated by the emergence of a positive peak in \( f^o \) around the unsupervised example, which enforces the hard unilateral constraint to hold with the equality. For simplicity, we have considered the case in which there is only one unsupervised point associated with a hard unilateral constraint of the form (27) but, of course, the example can be extended to the case of many such points (also in the presence of supervised examples dealt with in a soft way).

As a second numerical example, we consider the case in which the system (32) is solved by the choice \( \Lambda^{(j)} = 0 \). This happens whenever the system (in the unknowns \( F_H^{(j)} \) and \( F_S^{(j)} \))

\[
\begin{align*}
F_H^{(j)} &= -\gamma^{-1} \left[ \frac{\mu}{|A_S|} G_{HS}(\tilde{F}_S^{(j)} - \tilde{Y}_S^{(j)}) \right] \geq 0, \\
F_S^{(j)} &= \left[ I_{|A_S|} + \gamma^{-1} \frac{\mu}{|A_S|} G_{SS} \right]^{-1} \gamma^{-1} \frac{\mu}{|A_S|} G_{SS} \tilde{Y}_S^{(j)},
\end{align*}
\]

has a solution. In such a situation, all the inequality constraints are inactive at local optimality. Figure 4 shows an example (with \( d = 2 \) and \( n = 1 \)) in which this happens, referring to the Gaussian kernel with \( \sigma = 2, \mu = 30, \) and \( \gamma = 1 \) in the formulation of Problem LMPC. In this case, the set \( \{(-1, 0), (-2, -1), (-3, -4), (0, 3)\} \) of positive supervised examples associated with the label \( y^{(h)} = 1 \) is given (and dealt with by using the soft quadratic loss (15)), together with the set of unsupervised examples \( \{(-5, 5), (-6, -2), (1, 7), (-5, -5)\} \), for which one requires \( f(x^{(h)}) \geq 0 \). Again, in the figure the supervised examples associated with the soft pointwise constraints are represented by crosses, whereas the unsupervised examples associated with the hard unilateral pointwise constraints are represented by stars (placed at \( y = 0.5 \) instead of \( y = 0 \), to make them better visible). In this case, all the Lagrange multipliers are equal to 0, as all the hard unilateral constraints are inactive at local (in this case, also global) optimality. The optimal solution \( f^o \) is the same that one would have obtained by training only on the positive supervised examples. Indeed, one can check a-posteriori that such a solution also satisfies the hard unilateral pointwise constraints on the unsupervised examples.

![Figure 3: Optimal solutions to Problem LMPC for the first numerical example presented in Section VIII-B. See the text for the explanations.](image)

![Figure 4: Optimal solution to Problem LMPC for the second numerical example presented in Section VIII-B. See the text for the explanations.](image)

Remark VIII.2. Although a hard unilateral pointwise constraint that is inactive at local optimality is always associated with a 0 Lagrange multiplier (hence a null constraint reaction), there may be cases in which the Lagrange multiplier is 0 (hence the constraint reaction is 0) even if the constraint is active at local optimality (i.e., it holds with the equality). In this case, we say that the hard unilateral constraint reaction is degenerate. Although we expect this situation to be rare in practice, a similar phenomenon is encountered sometimes in finite-dimensional constrained optimization problems (see, e.g., [30, Example 3.3.2]).

IX. DISCUSSION

We have addressed one of the remarkable open problems concerning the representation and the development of learning schemes, for agents acting in a constrained-based environment in the presence of mixed hard/soft pointwise constraints. Our study has focused on the issue of designing intelligent agents with effective learning capabilities in complex contexts where sensorial data (typically represented by soft constraints, as it happens, e.g., in the case of finite sets of supervised examples) are combined with knowledge-based descriptions of the tasks (represented by hard constraints). The interest in hard pointwise constraints is motivated by the fact that they model
very general prior knowledge. Such knowledge is obtained via discretizations of universal quantifiers by using a possibly large collection of unsupervised examples, which, in general, are cheaper to be obtained than supervised examples. Indeed, hard pointwise constraints can be regarded as discretizations of hard holonomic constraints (obtained when the set $X_H$ of finite cardinality in (1) or (2) is replaced by open sets), which are studied in [20], [21]. Extensions to constraints of different natures (e.g., holonomic, isoperimetric) and their combinations, algorithmic issues, and references to applications are also deeply investigated in [20].

In the present work, the investigation of learning with pointwise constraints has been performed from an optimization viewpoint, studying structural properties of the optimal solutions to the associated optimization problems, and finding them in closed form under some conditions. A possible extension concerns the application of tools from statistical learning theory, such as Rademacher’s complexity (see, e.g., [31]), to investigate how the presence of the constraints influences the generalization capability of the learned model. As an example, we mention that in the case of hard constraints, the set of admissible functions is restricted by their presence, so one expects a smaller upper bound on the corresponding Rademacher’s complexity, hence better bounds from statistical learning theory. However, this kind of investigation is outside the scope of this work. Here, we only mention that in [13] we applied tools from statistical learning theory to investigate the case of learning from supervised examples in the presence of additional boundary conditions.

ACKNOWLEDGEMENTS

This research was partially supported by the grant PRIN2009 “Learning Techniques in Relational Domains and Their Applications” (2009LNP494) from the Italian MURST and by the “Progetto di Ricerca di Ateneo 2013” “Dealing with High-Dimensional Data with Applications to Life Sciences”, granted by the University of Genoa. G. Gnecco and M. Sanguineti are members of the “Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni” (GNAMPA) of the “Istituto Nazionale di Alta Matematica” (INdAM).

APPENDIX: TECHNICAL DEFINITIONS AND PROOFS

In order to investigate the existence and uniqueness of an optimal solution to Problem LMPC (see Theorem III.1), we need the following definitions and the next Lemma IX.1.

Recall that for a Hilbert space $H$ with inner product $(\cdot, \cdot)_H$, a sequence $\{f^{(i)}\}$ in $H$ converges weakly to a function $f^* \in H$ iff for every $f \in H$ one has $(f, f^*)_H \to (f^*, f)_H$. A subset $S$ of $H$ is weakly closed iff the weak limit of each weakly convergent sequence $\{f^{(i)}\} \subseteq S$ belongs to $S$. A set $S \subseteq H$ is weakly compact $^9$ iff each sequence $\{f^{(i)}\} \subseteq S$ has a subsequence that converges weakly to some $f^* \in S$. A functional $\Phi$ on a nonempty and weakly closed subset $S$ of a Hilbert space $H$ is weakly lower semicontinuous iff for every $f \in S$ and every sequence $\{f^{(i)}\} \subseteq S$ weakly convergent to $f$, one has $\Phi(f) \leq \lim \inf_{i \to \infty} \Phi(f^{(i)})$.

The next lemma summarizes elementary properties of weakly lower semicontinuous functionals (see, e.g., [33, Appendix D]). It is exploited in the proof of Theorem III.1.

**Lemma IX.1.** The following hold.

(i) Every closed and convex subset of a Hilbert space is weakly closed (Mazur’s Theorem [33, p. 639]).

(ii) Every closed and bounded set of a Hilbert space is weakly relatively compact (i.e., its closure in the weak topology is compact). A weakly closed and bounded subset of a Hilbert space is weakly compact.

(iii) By (i) and (ii), every convex closed bounded set of a Hilbert space is weakly compact.

(iv) Every convex and continuous functional on a Hilbert space is weakly lower semicontinuous.

Proof of Theorem III.1. Since $\| f \|_{P, \gamma}$ is a RKHS norm on $F$ and $V$ is convex and continuous, the functional $\frac{1}{2} \| f \|_{P, \gamma}^2$ is strictly convex and continuous on $F$, whereas $\frac{1}{2} \sum_{l=1}^{L} \sum_{k=0}^{K_l} V(f(l, \kappa), \gamma^{(k)})$ is convex and continuous on $F$ (by the continuity of the evaluation functional in a RKHS). Then, it follows by Lemma IX.1 (iv) that the functional $E_\gamma(f)$ is weakly lower semicontinuous on $F$. Moreover, since for any sufficiently large $C \in \mathbb{R}$ the set $S_C := \{ f \in F_c \mid E_\gamma(f) \leq C \}$ is nonempty, closed, bounded, and convex, by Lemma IX.1 (iii) $S_C$ is nonempty and weakly compact. Finally, since the sum of a strictly convex functional and a convex one is strictly convex, by Lemma IX.1 (v) the global minimizer is unique.

**Proof of Theorem IV.1.**

(i) Recall that the set $X_H$ has the definition $X_H := \{ x^{(1)}, x^{(2)}, \ldots, x^{(l)}, \ldots, x^{(|X_H|)} \}$ and $m(x(l))$ is the number of hard pointwise constraints defined in the generic $x(l) \in X_H$. For $l = 1, \ldots, |X_H|$, let the auxiliary functions $^{10}$ $\eta^{(l)}, \ldots, \eta^{(l)}_{m(x(l))} \in C_b^L(X, \mathbb{R}^n)$ (i.e., the set of vector-valued functions from $X$ to $\mathbb{R}^n$ that are continuously differentiable up to the order $k$ and have compact supports) be chosen according to the following rules:

- the supports of $\eta^{(l)}, \ldots, \eta^{(l)}_{m(x(l))}$ are contained in an open ball of sufficiently small radius, centered in $x^{(l)}$;

These different definitions are needed because for infinite-dimensional spaces, compactness of a set does not coincide in general with its simultaneous boundedness and closedness (see, e.g., [32, Chapter 4]).

9In general, closed and bounded sets are not weakly compact in Hilbert spaces (e.g., the set consisting of an orthonormal basis in an infinitely-dimensional Hilbert space is closed and bounded but not weakly compact, as it does not contain 0).

10Here, to improve the readability of the notation, the second index appears as a subscript as done before to denote the components of a vector, but in this case $\eta^{(l)}, \ldots, \eta^{(l)}_{m(x(l))}$ are vectors.
the auxiliary functions correspondent to different indices $l$ (hence to different points $x_r^{(l)}$) have disjoint supports;
• the $m(x_r^{(l)}) \times m(x_r^{(l)})$ matrix
  $$\frac{\partial \phi_1, \ldots, \phi_m(x_r^{(l)})}{\partial (f_1, \ldots, f_m)} \eta^{(l)}_1, \ldots, \eta^{(l)}_m(x_r^{(l)})$$
evaluated in $x_r^{(l)}$, is non-singular\footnote{Choosing the functions $\eta^{(l)}_1, \ldots, \eta^{(l)}_m(x_r^{(l)})$ in such a way is always possible, thanks to the assumed non-singularity of the Jacobian matrix (5).}, where $\eta^{(l)}_1, \ldots, \eta^{(l)}_m(x_r^{(l)})$ denotes the $n \times m(x_r^{(l)})$ matrix obtained by a concatenation of the column vectors $\eta^{(l)}_1, \ldots, \eta^{(l)}_m(x_r^{(l)})$.

Now, let us choose an arbitrary function $\eta \in C^k_b(X, \mathbb{R}^n)$ and, for $\varepsilon, \varepsilon_r^{(l)} \in \mathbb{R}$ ($l = 1, \ldots, |X_H|, r = 1, \ldots, m(x_r^{(l)}))$, consider the problem of minimizing the function

$$F(\varepsilon, \varepsilon_r^{(l)}):= E_\varepsilon \left( f^0 + \varepsilon \eta + \sum_{l=1}^{|X_H|} \sum_{r=1}^{m(x_r^{(l)})} \varepsilon_r^{(l)} \eta^{(l)}_r \right)$$

(where $E_\varepsilon$ has the expression (4)) subject to the set of equality constraints given by

$$\forall x_h^{(l)} \in X_H, \forall \varepsilon \in \mathbb{N}_{m(x_r^{(l)})},
G_h^{(l)}(\varepsilon, \varepsilon_r^{(l)}) := \phi_i \left( x_h^{(l)}, f^0 + \varepsilon \eta + \sum_{l=1}^{|X_H|} \sum_{r=1}^{m(x_r^{(l)})} \varepsilon_r^{(l)} \eta^{(l)}_r \right) (x_h^{(l)}) = 0.$$ 

Note that, due to the construction of the functions $\eta^{(l)}_r$, the constraints above also can be written as

$$\forall x_h^{(l)} \in X_H, \forall \varepsilon \in \mathbb{N}_{m(x_r^{(l)})},
G_h^{(l)}(\varepsilon, \varepsilon_r^{(l)}) := \phi_i \left( x_h^{(l)}, f^0 + \varepsilon \eta + \sum_{r=1}^{m(x_r^{(l)})} \varepsilon_r^{(l)} \eta^{(l)}_r \right) (x_h^{(l)}) = 0.$$ 

Let us collect the functions $G_h^{(l)}(\varepsilon, \varepsilon_r^{(l)})$ into the column vector $G(\varepsilon, \varepsilon_r^{(l)})$ of dimension $M:= \sum_{l=1}^{|X_H|} m(x_r^{(l)})$ (using, e.g., the lexicographical order). Of course, being $f^0$ a constrained local minimizer of $E_\varepsilon$ implies that $(0, \ldots, 0)$ is a constrained local minimizer of $F(\varepsilon, \varepsilon_r^{(l)})$ under the constraints (33). Since the non-singularity of (5) provides the qualification\footnote{In particular, in this case, the so-called linear independence constraint qualification (LICQ) \cite{30}.} of the set of hard constraints $G(\varepsilon, \varepsilon_r^{(l)}) = 0$ in $(0, \ldots, 0)$, one can apply the theory of Lagrange multipliers in infinite-dimensional spaces (see, e.g., \cite[Chapter 3]{30}) to conclude that there exist Lagrange multipliers $\{\lambda_h^{(l)}, h = 1, \ldots, |X_H|, i = 1, \ldots, m(x_h^{(l)})\}$, collected into a column vector $\lambda \in \mathbb{R}^M$, such that

$$\nabla \varepsilon \left( \nabla (\varepsilon_r^{(l)}) \right) F(0, \ldots, 0) + \nabla \varepsilon \left( \nabla (\varepsilon_r^{(l)}) \right) G(0, \ldots, 0) \lambda = 0,$$

$$\nabla \varepsilon \left( \nabla (\varepsilon_r^{(l)}) \right) G(0, \ldots, 0) \lambda = 0$$

subject to a set of inequality constraints given by

$$\nabla \varepsilon \left( \nabla (\varepsilon_r^{(l)}) \right) F(0, \ldots, 0) + \nabla \varepsilon \left( \nabla (\varepsilon_r^{(l)}) \right) G(0, \ldots, 0) \lambda = 0$$

where $\nabla \varepsilon (\varepsilon_r^{(l)}) F$ and $\nabla \varepsilon (\varepsilon_r^{(l)}) G$ denote the gradient of $F$ with respect to the variables $\{\varepsilon_r^{(l)}\}$ and $\varepsilon$, resp., whereas $\nabla \varepsilon (\varepsilon_r^{(l)}) G$ and $\nabla \varepsilon (\varepsilon_r^{(l)}) G$ are the Jacobian of $G$ with respect to the variables $\{\varepsilon_r^{(l)}\}$ and $\varepsilon$, resp. Notice that both terms $\nabla \varepsilon (\varepsilon_r^{(l)}) F(0, \ldots, 0)$ and $(\nabla \varepsilon (\varepsilon_r^{(l)}) G)'(0, \ldots, 0)$ do not depend on the choice of the arbitrary function $\eta$ and that the term $\nabla \varepsilon (\varepsilon_r^{(l)}) G(0, \ldots, 0)$ is equal to the block-diagonal matrix

$$J := \begin{pmatrix} J^{(1)} & 0 & \cdots & 0 \\ 0 & J^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J^{(|X_H|)} \end{pmatrix},$$

where, for $h = 1, \ldots, |X_H|,$

$$J^{(h)} := \frac{\partial \phi_i \ldots \phi_m(x_h^{(l)})}{\partial (f_1^{(l)} \ldots f_m^{(l)})} \eta^{(h)}_1, \ldots, \eta^{(h)}_m(x_h^{(l)})],$$
evaluated in $x_h^{(l)}$, so $J$ is invertible by construction. Concluding, formula (34) allows one to compute the vector $\lambda$, regardless of the specific choice of the function $\eta$. Finally, by making explicit the partial derivatives, formula (35) is equivalent to

$$\int_{X} \left[ \gamma L f^0(x) + \sum_{\lambda=1}^{m(x_h^{(l)})} \lambda^{(h)} (x - x_h^{(l)}) \right] \eta(x) dx = 0,$$

where the term in $L f^0(x)$ of (36) follows by

$$\frac{1}{2} \langle P f^0 + \varepsilon \eta, P f^0 + \varepsilon \eta \rangle =$$

$$= \frac{1}{2} \langle P f^0, P f^0 \rangle + \varepsilon \langle P f^0, P \eta \rangle + O(\varepsilon^2) =$$

$$= \frac{1}{2} \langle P f^0, P f^0 \rangle + \varepsilon \langle L f^0, \eta \rangle + O(\varepsilon^2),$$

where $O(\cdot)$ denotes the “big O” notation. The Dirac delta terms in (36) arise from the facts that $\nabla \varepsilon (\varepsilon_r^{(l)}) G(0, \ldots, 0)$ depends only the values assumed by the vector $\eta(x)$ for $x = x_h^{(l)}, h = 1, \ldots, |X_H|$ and the loss $V$ is evaluated only in correspondence of the elements $\varepsilon^{(l)}$ of the set $X_H$. Finally, (6) follows from (36), since $\eta \in C^0_b(X, \mathbb{R}^n)$ is arbitrary.
\[ \forall x^{(h)} \in \mathcal{X}_H, \forall i \in \mathbb{N}_n(x^{(h)}) , \quad G^i(h)(\varepsilon, \varepsilon^{(l)}) := \delta \left( x^{(h)}, \left( f^a + \varepsilon h + \sum_{i=1}^{\|X_H\|} \sum_{r=1}^{m} \varepsilon_r \eta_r \right)(x^{(h)}) \right) \geq 0. \]

Since the active constraints at local optimality are qualified by assumption, one can proceed similarly as in the proof of part (i), by applying in addition the Karush-Kuhn-Tucker necessary conditions for local optimality, which provide the correct sign of the Lagrange multipliers associated with the inequality constraints.

**Proof of Proposition V.2.** For what concerns the reactions \( \omega^S \) of the soft constraints, the uniqueness follows directly by (8). Regarding the reactions \( \omega^H \) of the hard constraints, we prove the uniqueness by contradiction. Let us assume that there exist two different sets of Lagrange multipliers associated with the same constrained local minimizer \( f^a \), namely, \( \{ \lambda^i(h), i = 1 \ldots , m, h = 1 \ldots , |X_H| \} \) and \( \{ \lambda_i(h) \}, i = 1 \ldots , m, h = 1 \ldots , |X_H| \), with at least one \( \lambda^i(h) \neq \lambda_i(h) \). According to Theorem IV.1 (i), \( f^a \) satisfies the Euler-Lagrange equations (6). Subtracting the two expressions of \( f^a \) obtained in terms of the two sets of Lagrange multipliers, we get

\[
\sum_{i=1}^{\|X_H\|} \sum_{i=1}^{m} \left( \lambda^i(h) - \lambda_i(h) \right) \delta(x - x^{(h)}) \nabla f_i(x^{(h)}, f^a(x^{(h)})) = 0,
\]

Now, distinct multipliers are only compatible with the singularity of the Jacobian matrix, which contradicts the assumption on the invertibility of (5).

**REFERENCES**


