Learning as Constraint Reactions*

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

Abstract. A theory of learning is proposed, which extends naturally the classic regularization framework of kernel machines to the case in which the agent interacts with a richer environment, compactly described by the notion of constraint. Variational calculus is exploited to derive general representer theorems that give a description of the structure of the solution to the learning problem. It is shown that such solution can be represented in terms of *constraint reactions*, which remind the corresponding notion in analytic mechanics. In particular, the derived representer theorems clearly show the extension of the classic kernel expansion on support vectors to the expansion on *support constraints*. As an application of the proposed theory three examples are given, which illustrate the dimensional collapse to a finite-dimensional space of parameters. The constraint reactions are calculated for the classic collection of supervised examples, for the case of box constraints, and for the case of hard holonomic linear constraints mixed with supervised examples. Interestingly, this leads to representer theorems for which we can re-use the kernel machine mathematical and algorithmic apparatus.

Giorgio Gnecco
IMT, Piazza S. Ponziano 6, 55100 Lucca, Italy
e-mail: giorgio.gnecco@imtlucca.it

Marco Gori · Stefano Melacci
DIISM – University of Siena, Via Roma 56, 53100 Siena, Italy
e-mail: {marco,mela}@diism.unisi.it

Marcello Sanguineti
DIBRIS – University of Genoa, Via Opera Pia 13, 16145 Genova, Italy
e-mail: marcello.sanguineti@unige.it

* Part of this chapter is an excerpt of the paper "Foundations on Support Constraint Machines," by the same authors, that will appear in Neural Computation, which contains a significantly broader view on the subject along with a more general framework, algorithmic issues, and references to applications.

P. Koprinkova-Hristova et al. (eds.), *Artificial Neural Networks*,
Springer Series in Bio-/Neuroinformatics 4, DOI: 10.1007/978-3-319-09903-3_12

1 Introduction

Examples of constraints in machine learning come out naturally in various situations: constraints may represent, for instance, prior knowledge provided by an expert (e.g., a physician in the case of a medical application: in such a case constraints may be expressed in the form of rules which help in the detection of a disease [14, 16]). The expressive power of constraints becomes particularly significant when dealing with a specific problem, like vision, control, text classification, ranking in hypertextual environment, and prediction of the stock market.

Table 1 provides some examples of constraints that are often encountered in practical problems arising in different domains. The first example (*i*) describes the simplest case in which we handle several classic pairs (x_κ, y_κ) provided for supervised learning in classification, where x_κ is the κ-th supervised example and $y_\kappa \in \{-1, 1\}$ is its label. If $f(\cdot)$ is the function that the artificial agent is expected to compute, then the corresponding real-valued representation of the constraint is just the translation of the classic "robust" sign agreement between the target and the function to be learned. Example *ii* is the normalization of a probability density function, whereas example *iii* (which refers to a binary classification problem) imposes the coherence between the decisions taken on $S_1 x$ and $S_2 x$ for the object x, where S_1 and S_2 are matrices used to select two different views of the same object (see [17]). In the example *iv* we report a constraint from computer vision coming from the classic problem of determining the optical flow. It consists of finding the smoothest solution for the velocity field under the constraint that the brightness of any point in the movement pattern is constant. If $u(x,y,t)$ and $v(x,y,t)$ denote the components of the velocity field and $E(x,y,t)$ the brightness of any pixel (x,y) at time t, then the velocity field satisfies the linear constraint indicated in Table 1 *iv*. Finally, example *v* in the table refers to a document classification problem, and states the rule that all papers dealing with numerical analysis and neural networks are machine-learning papers. Notice that, whereas the first row of example *v* expresses the rule by a first-order logic description, in the second row there is a related constraint expressed by real-valued functions that is constructed using the classic product T-norm [4, 5, 13].

The aim of this chapter is to show how the framework of kernel machines can be extended to support constraint machines by including prior knowledge modeled by several kinds of constraints. In particular, we propose a framework in which the ambient space is described in terms of Reproducing Kernel Hilbert Spaces (RKHSs) of Sobolev type, which has the advantage, over generic RKHSs, of providing optimality conditions expressed as partial differential equations (see Theorems 1 and 2 in Section 2). The general learning paradigm of support constraints machines, its mathematical foundations, representer theorems, and algorithmic issues is presented in [11].

Unlike the classic framework of learning from examples, the beauty and the elegance of the simplicity behind the parsimony principle - for which simple explanations are preferred to complex ones - has not been profitably used yet for the formulation of systematic theories of learning in general constrained environments, although there are some works on learning in specific constrained contexts [2, 6, 12, 15, 21–23]. We propose the study of parsimonious agents interacting

Table 1 Examples of constraints from different environments. For each entry, both a linguistic description of the constraint and its real-valued representation are provided.

i	κ-th supervised pair for classification $y_\kappa \cdot f(x_\kappa) - 1 \geq 0$
ii	normalization of a probability density function $\int_{\mathscr{X}} f(x)dx = 1$, and $\forall x \in \mathscr{X} : \ f(x) \geq 0$
iii	coherence constraint (two classes) $\forall x \in \mathscr{X} : \ f_1(S_1 x) \cdot f_2(S_2 x) > 0$
iv	brightness invariance - optical flow $\frac{\partial E}{\partial x}u + \frac{\partial E}{\partial y}v + \frac{\partial E}{\partial t} = 0$
v	document classification: $\forall x : \ na(x) \wedge nn(x) \Rightarrow ml(x)$ $\forall x \in \mathscr{X} : \ f_{na}(x) f_{nn}(x)(1 - f_{ml}(x)) = 0$

simultaneously with examples and constraints in a multi-task environment with the purpose of developing the simplest (smoothest) vectorial function in a set of feasible solutions. More precisely, we think of an intelligent agent acting on a subset $\mathscr{X}$ of the perceptual space $\mathbb{R}^d$ as one implementing a vectorial function $f := [f_1, \ldots, f_n]' \in \mathscr{F}$, where $\mathscr{F}$ is a space of functions from $\mathscr{X}$ to $\mathbb{R}^n$. Each function f_j is referred to as a *task of the agent*. We assume that additional prior knowledge is available, defined by the fulfillment of constraints modeled as

$$\forall x \in \mathscr{X}_i \subseteq \mathscr{X} : \ \phi_i(x, f(x)) = 0, \ i = 1, \ldots, m, \tag{1}$$

or as

$$\forall x \in \mathscr{X}_i \subseteq \mathscr{X} : \ \check{\phi}_i(x, f(x)) \geq 0, \ i = 1, \ldots, m, \tag{2}$$

where $\phi_i, \check{\phi}_i$ are scalar-valued functions. Following the terminology in variational calculus, when the sets $\mathscr{X}_i$ are open we call (1) *bilateral holonomic constraints* and (2) *unilateral holonomic constraints*. When the sets $\mathscr{X}_i$ are made by finite numbers of points, we replace the term "holonomic" by *point-wise*. The constraints above are called *hard* if they cannot be violated; constraints that can be violated (at the cost of some penalization) play the role of *soft* constraints (this is usually the case for supervised pairs of the learning set). In this way, any cross-dependence amongst the functions f_j is expressed directly by the constraints.

In this chapter, which is an improved and extended version of [8], we investigate theoretically and by means of case studies the problem of learning in a constraint-based environment, taking into account both hard and soft constraints of holonomic and point-wise types. The focus on holonomic constraints is motivated by the fact that they model very general prior knowledge, expressed by universal quantifiers. Examples of learning problems with holonomic constraints are given, e.g, in [5], where the constraints arise by a suitable representation of prior knowledge expressed in terms of first-order-logic clauses. We also consider point-wise constraints, which arise, e.g., in the case of interpolation and approximation problems, given a finite set of examples. However, the proposed framework can be extended to several other

kinds of constraints and their combinations (e.g., isoperimetric ones, box constraints [9, 16], and boundary conditions [10]).

The chapter is organized as follows. In Section 2 we formalize the problems of learning from soft and hard constraints and we present the corresponding representer theorems, which provide information on the form of their solutions. Section 3 is devoted to the concepts of reactions of the constraints and support constraint machines. Section 4 analyzes some practical instances of the proposed framework. Section 5 is a discussion. Finally, the most technical results are detailed in three appendices.

2 Learning from Constraints and Its Representer Theorems

In this chapter, we assume $\mathscr{X}$ 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]. In particular, we consider the case in which, $\forall j \in \mathbb{N}_n := \{1, \ldots, n\}$ and some positive integer k, the function $f_j : \mathscr{X} \to \mathbb{R}$ belongs to the Sobolev space $\mathscr{W}^{k,2}(\mathscr{X})$, i.e., the subset of $\mathscr{L}^2(\mathscr{X})$ whose elements f_j have weak partial derivatives up to the order k with finite $\mathscr{L}^2(\mathscr{X})$-norms. So, the ambient space of the class of proposed problems of learning from constraints is

$$\mathscr{F} := \underbrace{\mathscr{W}^{k,2}(\mathscr{X}) \times \ldots \times \mathscr{W}^{k,2}(\mathscr{X})}_{n \text{ times}}.$$

We take $k > \frac{d}{2}$ since, by the Sobolev Embedding Theorem (see, e.g., Chapter 4 in [1]), for $k > \frac{d}{2}$ each element of $\mathscr{W}^{k,2}(\mathscr{X})$ has a continuous representative, and under such an assumption $\mathscr{F}$ is a RKHS.

We can introduce a seminorm $\| f \|_{P,\gamma}$ on $\mathscr{F}$ via the pair (P, γ), where $P := [P_0, \ldots, P_{l-1}]'$ is a suitable (vectorial) finite-order[1] differential operator of order k with l components and $\gamma \in \mathbb{R}^n$ is a fixed vector with positive components. Let us consider the functional

$$\begin{aligned} \mathscr{E}(f) := \| f \|^2_{P,\gamma} &= \sum_{j=1}^{n} \gamma_j < Pf_j, Pf_j > \\ &= \sum_{j=1}^{n} \gamma_j \left(\sum_{r=0}^{l-1} \int_{\mathscr{X}} (P_r f_j(x) P_r f_j(x)) dx \right). \end{aligned} \tag{3}$$

Note that we overload the notation and use the symbol P for both the (matrix) differential operator acting on f and the (vector) one acting on its components. If we choose for P the form used in Tikhonov's stabilizing functionals [24], for $n = 1$ and $l = k + 1$ we get

$$\| f \|^2_{P,\gamma} = \gamma \int_{\mathscr{X}} \sum_{r=0}^{k} \rho_r(x) \left(D_r f(x) \right)^2 dx,$$

[1] The results can be extended to infinite-order differential operators (see the example in Section 4.3).

where the function $\rho_r(x)$ is nonnegative, $P_r := \sqrt{\rho_r(x)}D_r$, and D_r denotes a differential operator with constant coefficients containing only partial derivatives of order r. In this work, we focus on the case in which the operator P is invariant under spatial shift and has constant coefficients. For a function u and a multiindex α with d nonnegative components α_j, we write $D^\alpha u$ to denote $\frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} u$, where $|\alpha| := \sum_{j=1}^{d} \alpha_j$. So, the generic component P_i of P has 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 $P^\star = [P_0^\star, \ldots, P_{l-1}^\star]'$ whose i-th component $P_i^\star$ is given by $P_i^\star = \sum_{|\alpha| \leq k} (-1)^{|\alpha|} b_{i,\alpha} D^\alpha$. Finally, we define the operators $L := (P^\star)'P$ and, using again an overloaded notation, $\gamma L := [\gamma_1 L, \ldots, \gamma_n L]'$.

2.1 Soft Constraints

We start considering the case of learning from soft constraints, whose representer theorem has a simpler formulation than in the case of hard ones. The problem of learning from soft constraints is based on a direct soft re-formulation of (3). We can associate any holonomic or point-wise unilateral constraint $\breve{\phi}_i(x, f(x)) \geq 0$ with $\phi_i^{\geq}(x, f(x)) = 0$, where $\phi_i^{\geq}(\cdot,\cdot)$ is a non-negative function. Similarly, each holonomic or point-wise bilateral constraint can be expressed via a pair of unilateral constraints. Hence, regardless of bilateral or unilateral constraints, the problem of learning from soft holonomic or point-wise constraints can be formulated as the minimization of the functional

$$\mathscr{L}_s(f) := \frac{1}{2}\mathscr{E}(f) + \sum_{i=1}^{m} \int_{\mathscr{X}} p(x)\ 1_{\mathscr{X}_i}(x) \phi_i^{\geq}(x, f(x)) dx, \tag{4}$$

where $p(x)$ is a (nonnegative) weight function, e.g., a probability density function (this setting can be extended to the case of a generalized probability density function). We use $1_{\mathscr{X}_i}(\cdot)$ to denote the characteristic function of the set $\mathscr{X}_i$ when $\mathscr{X}_i$ is open. In order to keep the notation uniform, we let $1_{\mathscr{X}_i}(\cdot) := \delta(\cdot - x_i)$, where δ denotes the Dirac delta, for a set $\mathscr{X}_i = \{x_i\}$ made up of a single element. Finally, for two vector-valued functions $u^{(1)}$ and $u^{(2)}$ of the same dimensions, $u^{(1)} \otimes u^{(2)}$ represents 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., $v_i := (u^{(1)} \otimes u^{(2)})_i := u_i^{(1)} \otimes u_i^{(2)}$ for each index i.

Theorem 1. *(*REPRESENTER THEOREM FOR SOFT HOLONOMIC AND SOFT POINT-WISE CONSTRAINTS*). Let $p(\cdot)$ be continuous, nonnegative and in $\mathscr{L}^1(\mathscr{X})$, and let f^o be a local minimizer of the functional (4) over $\mathscr{F}$.*

(i) Let also the following hold: $\forall i \in \mathbb{N}_m$, $\mathscr{X}_i \subseteq \mathscr{X}$ is open and $\forall x \in \mathscr{X}_i$, there is an open neighborhood $\mathscr{N}$ of $(x, f^o(x))$ for which $\phi_i^{\geq} \in \mathscr{C}^1(\mathscr{N})$. Then, f^o satisfies on $\mathscr{X}$

$$\gamma L f^o(x) + \sum_{i=1}^{m} p(x) 1_{\mathscr{X}_i}(x) \cdot \nabla_f \phi_i^{\geq}(x, f^o(x)) = 0, \tag{5}$$

where $\gamma L := [\gamma_1 L, \ldots, \gamma_n L]'$ is a spatially-invariant operator[2]*, and $\nabla_f \phi_i^{\geq}$ is the gradient w.r.t. the second vector argument f of the function $\phi_i^{\geq}$.*

(ii) Suppose now that the sets $\mathscr{X}_i$ are disjoint and each set $\mathscr{X}_i$ is made up of a single point x_i, and that $\phi_i^{\geq}$ has the form

$$\phi_i^{\geq}(x, f(x)) = \sum_{j \in \mathbb{N}_n} \phi_{i,j}^{\geq}(x, f_j(x)),$$

where $\phi_{i,j}^{\geq}(x, f_j(x)) := (1 - y_{i,j} \cdot f_j(x))_+$, and the $y_{i,j}$'s belong to the set $\{-1, 1\}$. Then, f^o satisfies on $\mathscr{X}$ (for $j = 1, \ldots, n$)

$$\gamma_j L f_j^o(x) + \sum_{i=1}^{m} p(x) 1_{\mathscr{X}_i}(x) \cdot \overline{\partial_{f_j}} \phi_{i,j}^{\geq}(x, f_j^o(x)) = 0, \tag{6}$$

where $\overline{\partial_{f_j}} \phi_{i,j}^{\geq}(x, f_j^o(x))$ is a suitable element of the subdifferential[3] *$\partial_{f_j} \phi_{i,j}^{\geq}(x, f_j^o(x))$.*

(iii) Let the assumptions of either item (i) or item (ii) hold. If, moreover, $\mathscr{X} = \mathbb{R}^d$, L is invertible on $\mathscr{W}^{k,2}(\mathscr{X})$, and there exists a free-space Green's function g of L that belongs to $\mathscr{W}^{k,2}(\mathscr{X})$, then f^o can be represented as

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) \vec{\otimes} \phi_i^{\geq}(\cdot, f^o(\cdot)), \tag{7}$$

where $g \vec{\otimes} \phi_i^{\geq} := g \otimes \omega_i^{\geq}$ and

$$\omega_i^{\geq} :=\uparrow \phi_i^{\geq}(\cdot, f^o(\cdot)) := -p(\cdot) 1_{\mathscr{X}_i}(\cdot) \nabla_f \phi_i^{\geq}(\cdot, f^o(\cdot))$$

under the assumptions of (i), while the n components of $\omega_i^{\geq}$ are defined as

$$\omega_{i,j}^{\geq} :=\uparrow \phi_i^{\geq}(\cdot, f^o(\cdot)) := -p(\cdot) 1_{\mathscr{X}_i}(\cdot) \overline{\partial_{f_j}} \phi_{i,j}^{\geq}(\cdot, f^o(\cdot))$$

under the assumptions of (ii).

Proof. (i) is proved by fixing arbitrarily $\eta \in \mathscr{C}_0^k(\mathscr{X}, \mathbb{R}^n)$ (the set of functions from $\mathscr{X}$ to $\mathbb{R}^n$ that are continuously differentiable up to order k, and have compact support), then computing

[2] Here we use again an overloaded notation, as made for the operator P.

[3] Let $\Omega \subseteq \mathbb{R}^d$ be a convex set. We recall that the subdifferential of a convex function $u : \Omega \to \mathbb{R}$ at a point $x_0 \in \Omega$ is the set of all the subgradients of u at x_0, that is the set of all vectors $v \in \mathbb{R}^d$ such that $f(x) - f(x_0) \geq v'(x - x_0)$.

$$0 = \lim_{\varepsilon\to 0} \frac{\mathscr{L}_s(f^o+\varepsilon\eta)-\mathscr{L}_s(f^o)}{\varepsilon}$$
$$= \int_{\mathscr{X}} \left(\gamma Lf^o(x) + \sum_{i=1}^{m} p(x) 1_{\mathscr{X}_i}(x) \cdot \nabla_f \phi_i^{\geq}(x, f^o(x)) \right)' \eta(x) dx .$$

The first equality has been derived by the local optimality of f^o, whereas the second one has been derived by exploiting the assumption that $\forall x \in \mathscr{X}_i$ there is an open neighborhood $\mathscr{N}$ of $(x, f^o(x))$ for which $\phi_i^{\geq} \in \mathscr{C}^1(\mathscr{N})$. Finally, the proof is completed applying the fundamental lemma of the calculus of variations (for which we refer, e.g., to Section 2.2 in [7]).

(ii) Let us fix arbitrarily $\eta \in \mathscr{C}_0^k(\mathscr{X}, \mathbb{R}^n)$, with the additional condition that $\eta(x) = 0$ for all $x \in \cup_{i=1}^m \mathscr{X}_i$. Proceeding likewise in the proof of item (i), one obtains

$$\lim_{\varepsilon\to 0} \frac{\mathscr{L}_s(f^o+\varepsilon\eta)-\mathscr{L}_s(f^o)}{\varepsilon} = \int_{\mathscr{X}} (\gamma Lf^o(x))' \eta(x) dx = 0 . \tag{8}$$

Since, for each index $j = 1, \ldots, n$, $\gamma_j Lf_j^o$ is a distribution, formula (8) implies that the support of $\gamma_j Lf_j^o$ is a subset of $\{x_1, \ldots, x_m\}$, which is a set of finite cardinality. By Theorem XXXV in Chapter 3 of [19], $\gamma_j Lf_j^o$ is made up of a finite linear combination of Dirac delta's and their partial derivatives up to some finite order, centered on $x_1, \ldots, x_m$. Now, all the coefficients associated with the partial derivatives of any order of the Dirac delta's are 0, as it can be checked by choosing a function $\eta \in \mathscr{C}_0^\infty(\mathscr{X}, \mathbb{R}^n)$ such that only its j-th component η_j is different from 0, and $\eta_j(x) = 0$ for all $x \in \cup_{i=1}^m \mathscr{X}_i$ (even though some partial derivatives of some order of η_j may be different from 0 for some $x \in \cup_{i=1}^m \mathscr{X}_i$). Concluding, $\gamma_j Lf_j^o$ satisfies on $\mathscr{X}$

$$\gamma_j Lf_j^o(x) = \sum_{i=1}^{m} B_i \delta(x - x_i) , \tag{9}$$

where the B_i's are constants. Notice that (9) is of the same form as (6).

Now, we look for lower and upper bounds on the B_i's. For simplicity of exposition, in the following we suppose $m = 1$, so there is only one constant B_1 (however, the next arguments hold also for the case $m > 1$). We denote by η^{j+} any function in $\mathscr{C}_0^k(\mathscr{X}, \mathbb{R}^n)$ such that only its j-th component η_j^{j+} is different from 0, and $\eta_j^{j+}(x_1) > 0$. Once η^{j+} has been fixed, we denote by η^{j-} the function $-\eta^{j+}$. The following possible cases show up.

Case (a): $(1 - y_{1,j} \cdot f_j^o(x_1))_+ < 0$. In this case, one obtains

$$\lim_{\varepsilon\to 0^+} \frac{\mathscr{L}_s(f^o+\varepsilon\eta^{j+})-\mathscr{L}_s(f^o)}{\varepsilon}$$
$$= \int_{\mathscr{X}} \gamma_j Lf_j^o(x) \eta_j^{j+}(x) dx = B_1 \eta_j^{j+}(x_1) \geq 0 , \tag{10}$$

and

$$\lim_{\varepsilon\to 0^+} \frac{\mathscr{L}_s(f^o+\varepsilon\eta^{j-})-\mathscr{L}_s(f^o)}{\varepsilon} = \int_{\mathscr{X}} \gamma L_j f_j^o(x)\eta_j^{j-}(x)dx = -B_1\eta_j^{j+}(x_1) \geq 0\,, \tag{11}$$

then $B_1 = 0$ (since $\eta_j^{j+}(x_1) > 0$). Notice that the "$\geq$" in formulas (10) and (11) follow by the local optimality of f^o, whereas the first equalities by the left/right differentiability[4] of the function $(\cdot)_+$.

Case (b): $(1 - y_{1,j}\cdot f_j^o(x_1))_+ > 0$. Similarly, in this case, one obtains

$$\begin{aligned}\lim_{\varepsilon\to 0^+} & \frac{\mathscr{L}_s(f^o+\varepsilon\eta^{j+})-\mathscr{L}_s(f^o)}{\varepsilon} \\ &= \int_{\mathscr{X}} \left(\gamma_j L f_j^o(x) - y_{1,j}p(x)1_{\mathscr{X}_i}(x)\right)\eta_j^{j+}(x)dx \\ &= (B_1 - y_{1,j}p(x_1))\eta_j^{j+}(x_1) \geq 0\,,\end{aligned} \tag{12}$$

and

$$\begin{aligned}\lim_{\varepsilon\to 0^+} & \frac{\mathscr{L}_s(f^o+\varepsilon\eta^{j-})-\mathscr{L}_s(f^o)}{\varepsilon} \\ &= \int_{\mathscr{X}} \left(\gamma_j L f_j^o(x) - y_{1,j}p(x)1_{\mathscr{X}_i}(x)\right)\eta_j^{j-}(x)dx \\ &= -(B_1 - y_{1,j}p(x_1))\eta_j^{j+}(x_1) \geq 0\,,\end{aligned} \tag{13}$$

then $B_1 = y_{1,j}p(x_1)$.

Case (c): $(1 - y_{1,j}\cdot f_j^o(x_1))_+ = 0$ and $y_{1,j} = -1$. In this case, one obtains

$$\begin{aligned}\lim_{\varepsilon\to 0^+} & \frac{\mathscr{L}_s(f^o+\varepsilon\eta^{j+})-\mathscr{L}_s(f^o)}{\varepsilon} \\ &= \int_{\mathscr{X}} \left(\gamma_j L f_j^o(x) - y_{1,j}p(x)1_{\mathscr{X}_i}(x)\right)\eta_j^{j+}(x)dx \\ &= (B_1 - y_{1,j}p(x_1))\eta_j^{j+}(x_1) \geq 0\,,\end{aligned} \tag{14}$$

and

$$\lim_{\varepsilon\to 0^+} \frac{\mathscr{L}_s(f^o+\varepsilon\eta^{j-})-\mathscr{L}_s(f^o)}{\varepsilon} = \int_{\mathscr{X}} \gamma L_j f_j^o(x)\eta_j^{j-}(x)dx = -B_1\eta_j^{j+}(x_1) \geq 0\,, \tag{15}$$

then $B_1 \in [y_{1,j}p(x_1), 0] = [-p(x_1), 0]$.

[4] Depending on the sign of $y_{i,j}$.

Case (d): $(1 - y_{1,j} \cdot f_j^o(x_1))_+ = 0$ and $y_{1,j} = 1$. Finally, in this case, one obtains

$$\lim_{\varepsilon \to 0^+} \frac{\mathscr{L}_s(f^o + \varepsilon \eta^{j+}) - \mathscr{L}_s(f^o)}{\varepsilon} = \int_{\mathscr{X}} \gamma_j L f_j^o(x) \eta_j^{j+}(x) dx = B_1 \eta_j^{j+}(x_1) \geq 0, \tag{16}$$

and

$$\begin{aligned} \lim_{\varepsilon \to 0^+} & \frac{\mathscr{L}_s(f^o + \varepsilon \eta^{j-}) - \mathscr{L}_s(f^o)}{\varepsilon} \\ &= \int_{\mathscr{X}} \left(\gamma_j L f_j^o(x) - y_{1,j} p(x) 1_{\mathscr{X}_i}(x) \right) \eta_j^{j-}(x) dx \\ &= -(B_1 - y_{1,j} p(x_1)) \eta_j^{j+}(x_1) \geq 0, \end{aligned} \tag{17}$$

then $B_1 \in [0, y_{1,j} p(x_1)] = [0, p(x_1)]$.

Concluding, one obtains (6) summarizing the results of the analysis of cases (a)-(d), and applying the definition of subdifferentiability to the function $(\cdot)_+$.

(iii) follows by the Euler-Lagrange equations (5) (resp., (6)) of item (i) (resp., (ii)), the definition of the free-space Green's function g of L as the solution of $Lg = \delta$ (where δ denotes the Dirac delta, centered in 0), and the stated assumptions on L and g. □

Item (i) of Theorem 1 applies, e.g., to the case of a function $\phi_i^{\geq}$ that is continuously differentiable everywhere (or at least a function $\phi_i^{\geq}$ that is "seen as" a continuously differentiable function at local optimality, in the sense that $(x, f^o(x))$ is not a point of discontinuity of any partial derivative of $\phi_i^{\geq}$). However, a function $\phi_i^{\geq}$ deriving from a unilateral constraint may not be continuously differentiable everywhere. In such a case, one may approximate such a function by a continuously differentiable approximation, or (for certain $\phi_i^{\geq}$'s) deal directly with the nondifferentiable case, as shown in Theorem 1 (ii) for a particular choice of such functions. We remark that the classic supervised learning is a degenerate case of Theorem 1 (i), in which one sets $p(x)\, 1_{\mathscr{X}_i}(x) = p(x)\delta(x - x_i)$. Such a degenerate case is considered in Theorem 1 (ii) for the case of a particular nondifferentiable function $\phi_i^{\geq}$, but such a result can also be extended to other differentiable or nondifferentiable functions $\phi_i^{\geq}$. Finally, in Theorem 1 (iii) one can recognize both the ingredients of a parsimonious knowledge-based solution, i.e., the free-space Green's function g and the functions $\omega_i^{\geq}$, mixed by convolution. Indeed, note that, by defining $\omega^{\geq} := \sum_{i=1}^m \omega_i^{\geq}$, formula (7) can be re-written as $f^o = \gamma^{-1} g \otimes \omega^{\geq}$ and that, under the assumptions of Theorem 1 (iii), it follows by such an expression that the Fourier transform $\hat{f}^o$ of f^o (by Fourier transform of a vector-valued function we mean the vector of Fourier transforms of each component) is $\hat{f}^o = \gamma^{-1} \hat{g} \cdot \hat{\omega}^{\geq}$. Since under the assumptions of Theorem 1 (iii) the operator L is invertible and has $g\otimes$ as its inverse, from $f^o = \gamma^{-1} g \otimes \omega^{\geq}$ we get also $\omega^{\geq} = \gamma L f^o$, which is just a compact expression of the solution (7) to the Euler-Lagrange equations (5) or (6).

2.2 Hard Constraints

We now consider hard holonomic constraints. The following theorem prescribes the representation of the solution of the associated learning problem. Given a set of m holonomic constraints (defined, in general, on possibly different open subsets $\mathscr{X}_i$), we denote by $m(x)$ the number of constraints that are actually defined in the same point x of the domain. We denote by $\hat{\mathscr{X}}$ any open subset of $\mathscr{X}$, where the same subset of constraints is defined in all its points, in such a way that $m(x)$ is constant on the same $\hat{\mathscr{X}}$. By "cl" we denote the closure in the Euclidean topology. Finally, recall that a constraint $\check{\phi}_i(x, f(x)) \geq 0$ is said to be *active* in $x_0 \in \hat{\mathscr{X}}$ at local optimality iff $\check{\phi}_i(x_0, f^o(x_0)) = 0$, otherwise it is *inactive* in x_0 at local optimality.

Theorem 2. *(*REPRESENTER THEOREM FOR HARD HOLONOMIC CONSTRAINTS, CASE OF FUNCTIONAL LAGRANGE MULTIPLIERS*). Let us consider the minimization of the functional (3) in the case of $m < n$ hard bilateral constraints of holonomic type, which define the subset*

$$\mathscr{F}_\phi := \{f \in \mathscr{F} : \forall i \in \mathbb{N}_m, \forall x \in \mathscr{X}_i \subseteq \mathscr{X} : \phi_i(x, f(x)) = 0\}$$

of the function space $\mathscr{F}$, where $\forall i \in \mathbb{N}_m : \phi_i \in \mathscr{C}^{k+1}(\mathrm{cl}(\mathscr{X}_i) \times \mathbb{R}^m)$. Let f^o be any constrained local minimizer of class $\mathscr{C}^{2k}(\mathscr{X}, \mathbb{R}^n)$ of the functional (3). Let us assume that for any $\hat{\mathscr{X}}$ and for every x_0 in the same $\hat{\mathscr{X}}$ we can find two permutations σ_f and σ_ϕ of the indexes of the n functions f_j and of the m constraints ϕ_i, such that $\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))}$ refer to the constraints actually defined in x_0, and the Jacobian matrix

$$\frac{\partial(\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))})}{\partial(f^o_{\sigma_f(1)}, \ldots, f^o_{\sigma_f(m(x_0))})}, \tag{18}$$

evaluated in x_0, is not singular. Then, the following hold.

(i) There exists a set of functions $\lambda_i : \hat{\mathscr{X}} \to \mathbb{R}, i \in \mathbb{N}_m$, such that, in addition to the above constraints, f^o satisfies on $\hat{\mathscr{X}}$ the Euler-Lagrange equations

$$\gamma L f^o(x) + \sum_{i=1}^{m} \lambda_i(x) 1_{\mathscr{X}_i}(x) \cdot \nabla_f \phi_i(x, f^o(x)) = 0, \tag{19}$$

where $\gamma L := [\gamma_1 L, \ldots, \gamma_n L]'$ is a spatial-invariant operator, and $\nabla_f \phi_i$ is the gradient w.r.t. the second vector argument f of the function ϕ_i.

(ii) Let $\gamma^{-1} g := [\gamma_1^{-1} g, \ldots, \gamma_n^{-1} g]'$. If for all i one has $\mathscr{X}_i = \mathscr{X} = \mathbb{R}^d$, L is invertible on $\mathscr{W}^{k,2}(\mathscr{X})$, and there exists a free-space Green's function g of L that belongs to $\mathscr{W}^{k,2}(\mathscr{X})$, then f^o has the representation

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) \vec{\otimes} \phi_i(\cdot, f^o(\cdot)), \tag{20}$$

where $g \vec{\otimes} \phi_i := g \otimes \omega_i$ and $\omega_i(\cdot) :=\uparrow \phi_i(\cdot, f^o(\cdot)) := -\lambda_i(\cdot) 1_{\mathscr{X}_i}(\cdot) \nabla_f \phi_i(\cdot, f^o(\cdot))$.

(iii) For the case of $m < n$ unilateral constraints of holonomic type, which define the subset $\mathscr{F}_{\breve{\phi}} := \left\{ f \in \mathscr{F} : \forall i \in \mathbb{N}_m, \forall x \in \mathscr{X}_i \subseteq \mathscr{X} : \breve{\phi}_i(x, f(x)) \geq 0 \right\}$ of the function space $\mathscr{F}$, (i) and (ii) still hold (with every occurrence of ϕ_i replaced by $\breve{\phi}_i$) if one requires the nonsingularity of the Jacobian matrix (see (18)) to hold when restricting the constraints defined in x_0 to the ones that are active in x_0 at local optimality. Moreover, each Lagrange multiplier $\lambda_i(x)$ is nonpositive and equal to 0 when the correspondent constraint is inactive in x at local optimality.

Proof. The proof adapts to the case of hard holonomic constraints the one of Theorem 1 above. For completeness, it is detailed in Appendix 2. □

Notice that, due to the definition of ω_i, without loss of generality, one can define $\lambda_i(x) := 0$ for all $x \in \mathscr{X} \setminus \mathscr{X}_i$. Likewise in Theorem 1, by defining $\omega := \sum_{i=1}^{m} \omega_i$, formula (20) can be re-written as $f^o = \gamma^{-1} g \otimes \omega$, and, under the assumptions of Theorem 2 (ii),(iii), one can write $\hat{f}^o = \gamma^{-1} \hat{g} \cdot \hat{\omega}$. We also mention that a similar result can be proven for the case of hard point-wise constraints (in which one has discrete sets $\mathscr{X}_i$ composed of the $|\mathscr{X}_i|$ elements $x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,|\mathscr{X}_i|)}$), and for combinations of soft and hard constraints (e.g., soft point-wise constraints on supervised examples mixed with hard holonomic constraints, in which case the Lagrange multipliers are distributions instead of functions, as detailed in Appendix 2).

3 Support Constraint Machines

The next definition formalizes a concept that plays a basic role in the proposed learning paradigm.

Definition 1. The function $\omega_i^{\geq}$ in Theorem 1 (iii) (resp., the function ω_i in Theorem 2 (ii)) is called the *reaction of the i-th constraint* and $\omega^{\geq} := \sum_{i=1}^{m} \omega_i^{\geq}$ (resp. $\omega := \sum_{i=1}^{m} \omega_i$) is the overall reaction of the given constraints.

We emphasize the fact that the reaction of a constraint is a concept associated with the constrained local minimizer f^o. In particular, two different constrained local minimizers may be associated with different constraint reactions. A similar remark holds for the overall reaction of the constraints. Loosely speaking, under the assumptions of Theorem 1 (iii) or Theorem 2 (ii),(iii), the reaction of the *i*-th constraint provides the way under which such a constraint contributes to the expansion of f^o. So, in this case solving the learning problem is reduced to finding the reactions of the constraints.

Proposition 1. *Under the assumptions of the respective representer theorems (Theorem 1 (iii) and Theorem 2 (ii),(iii)), the reactions of the constraints are uniquely determined by the constrained local minimizer f^o.*

Proof. For the case of soft constraints (Theorem 1 (iii)), the statement follows directly by the definition of the reactions of the constraints $\omega_i^{\geq}(\cdot)$. For the case of hard constraints (Theorem 2 (ii),(iii)), the proof can be given by contradiction. Let us assume that there exist two different sets of Lagrange multipliers associated with

the same constrained local minimizer f^o: $\{\lambda_i,\ i=1\ldots,m\}$ and $\left\{\overline{\lambda}_i,\ i=1\ldots,m\right\}$, with at least one $\lambda_i \neq \overline{\lambda}_i$. According to Theorem 2 (i), f^o satisfies the Euler-Lagrange equations (19). Without loss of generality, for each $x \in \hat{\mathscr{X}}$, one can re-order the constraints and the associated Lagrange multipliers in such a way that the first $m(x)$ constraints are the ones actually defined in $x \in \hat{\mathscr{X}}$, and assume that $\lambda_i(x) = \overline{\lambda}_i(x) = 0$ for all indexes $i > m(x)$, as the corresponding constraint reactions are equal to 0 in x due to the definition of ω_i. Subtracting the two expressions of f^o in terms of the two sets of Lagrange multipliers, one obtains $\sum_{i=1}^{m}(\lambda_i - \overline{\lambda}_i)\nabla_f \phi_i = (\lambda_{(D)} - \overline{\lambda}_{(D)})' \frac{\partial(\phi_1,\ldots,\phi_{m(x)})}{\partial(f_1,\ldots,f_{m(x)})} = 0$, where $\lambda_{(D)} := [\lambda_1,\ldots,\lambda_m(x)]'$ and $\overline{\lambda}_{(D)} := [\overline{\lambda}_1,\ldots,\overline{\lambda}_{m(x)}]'$. Now, distinct multipliers are only compatible with the singularity of the Jacobian matrix, which contradicts the assumption on the invertibility of (18). □

A remarkable difference between the case of soft and hard constraints is the following. For soft constraints, the solution provided by Theorem 1 (iii) is based on the assumption of knowing the probability density of the data $p(\cdot)$. For hard constraints, instead (see Theorem 2 (ii),(iii)), one needs to compute the Lagrange multipliers $\lambda_i(\cdot)$ associated with the constraints, and also to check the (hard) satisfaction of the constraints.

Summing up, and removing the superscript "$\geq$" in $\phi_i^{\geq}$ and $\omega_i^{\geq}$ when the meaning is clear from the context, the solution of the learning problem is fully representable by the reactions ω_i of the constraints ϕ_i, as it is depicted in Fig. 1, where ∇_f denotes the gradient with respect to f, $\lambda_i(x)$ is the Lagrange multiplier, and $p(x)$ is the probability density. Interestingly, in the two cases, each constraint reaction has exactly the same dependency on the gradient of the constraint with respect to its second vector-valued argument but, while in the case of hard constraints the Lagrange multipliers need to be determined so as to impose the hard fulfillment of the constraints, in the case of soft constraints one exploits the probability density of

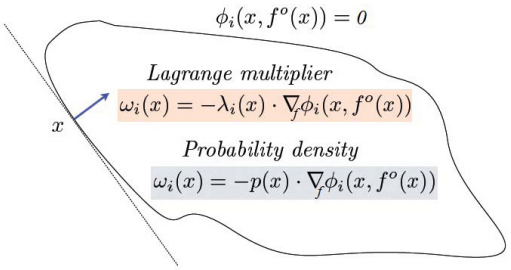

Fig. 1 Constraint reactions in the generic point $x \in \mathscr{X}$ corresponding to the cases of hard and soft constraints, where one can see, resp., the roles of the Lagrange multiplier associated with each constraint and of the probability density in x. For illustrative purposes, the case $n = 2$ is considered here. The reaction of the constraint ϕ_i in x is a vector orthogonal to the level lines of $\phi_i(x, f^o(x))$, interpreted as a function of its second vector-valued argument only.

the data - which comes from the problem formulation - in the representation of the constraint reactions. Both $\lambda_i(x)$ and $p(x)$ play a crucial role in determining the constraint reactions. For a given point x, the weights $\lambda_i(x)$ need to be computed in such a way not to violate the constraints at x, whereas in case of soft-fulfillment, $p(x)$ - which is the typical weight that is high (low) in regions of high (low) data density - is used to compute the constraint reactions. Now, we introduce the following two concepts.

Definition 2. A *support constraint* is a constraint associated with a reaction that is different from 0 at least in one point of the domain $\mathscr{X}$. A *support constraint machine* is any learning machine capable of finding a (local or global) solution to either of the problems of learning from constraints formulated in Theorems 1 or 2, when such a solution is expressed by either the representation (7) or the one (20).

So, under the assumptions of Theorem 1 (iii) or Theorem 2 (ii),(iii), the solution f^o can be obtained by the knowledge of the reactions associated merely with the support constraints. This motivates the use of the terminology "support constraints" as an extension of the classical concept of "support vectors" used in kernel methods [20]. Interestingly, support vectors are particular cases of support constraints. Indeed, the connection with kernel methods arises 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., [10]). Finally, we mention that for convex problems of learning from constraints, convex optimization algorithms can be used to find the reactions of the constraints (hence, to determine the support constraints).

4 Case Studies

4.1 Supervised Learning from Examples

The classic formulation is based on soft constraints and consists in finding $f^\star \in \operatorname{argmin}_{f\in\mathscr{F}} \mathscr{L}_s(f)$, where $\mathscr{L}_s(f)$ is given in formula (4) with $p(x)1_{\mathscr{X}_i}(x) = p(x)\delta(x - x_i)$ and each set $\mathscr{X}_i$ is a singleton. By $y_{i,j}$ we denote a real number for regression and an element of the set $\{-1,1\}$ for classification. As before, we denote by f^o a local minimizer (of which a global one $f^\star$ is a particular case). There are different possible choices for $\phi_i^{\geq}(x, f(x))$, which typically depend mostly on whether one faces regression or classification problems. The *quadratic loss* $V_Q(u) := \frac{1}{2}u^2$ is associated with the bilateral constraints $\phi_{i,j}(f_j(x)) = (y_{i,j} - f_j(x))$, which originate[5] - in the case of soft constraints - the term $\phi_i^{\geq}(f(x)) = \sum_{j\in\mathbb{N}_n} V_Q \circ \phi_{i,j}(f_j(x)) = \frac{1}{2}\sum_{j\in\mathbb{N}_n}(y_{i,j} - f_j(x))^2$. For every $j \in \mathbb{N}_n$ and $x \in \mathscr{X}$, by Theorem 1 the j-th component of the reaction of the i-th constraint is

[5] In the following, we write $\phi_i^{\geq}(f(x))$ instead of $\phi_i^{\geq}(x, f(x))$ since there is no explicit dependence on x.

$$\begin{aligned}\omega_{i,j}^{\geq}(x) &= (\uparrow \phi_i^{\geq}(f^o(x)))_j = -p(x)1_{\mathscr{X}_i}(x)\frac{\partial}{\partial f_j}\phi_i^{\geq}(f^o(x)) \\ &= -p(x)\delta(x-x_i)\frac{\partial}{\partial f_j}\left(\frac{1}{2}\sum_{h\in\mathbb{N}_n}(y_{i,h}-f_h^o(x))^2\right) \\ &= p(x)(y_{i,j}-f_j^o(x))\delta(x-x_i)\,.\end{aligned}$$

The *hinge loss* $V_H(u) = (u)_+$ is associated with the unilateral constraint $\breve{\phi}_{i,j}(f(x)) = 1 - y_{i,j}\cdot f_j(x)$, which gives rise to $\phi_i^{\geq}(f(x)) = \sum_{j\in\mathbb{N}_n} V_H \circ \breve{\phi}_{i,j}(f_j(x)) = \sum_{j\in\mathbb{N}_n}(1 - y_{i,j}\cdot f_j(x))_+$. In this case, the reaction of the i-th constraint is given by

$$\begin{aligned}\omega_{i,j}^{\geq}(x) &= (\uparrow \phi_i^{\geq}(f^o(x)))_j = -p(x)1_{\mathscr{X}_i}(x)\overline{\partial_{f_j}}\phi_i^{\geq}(f^o(x)) \\ &= -p(x)\delta(x-x_i)\overline{\partial_{f_j}}\left((1-y_{i,j}\cdot f_j^o(x))_+\right)\,,\end{aligned}$$

where $-\overline{\partial_{f_j}}\left((1-y_{i,j}\cdot f_j^o(x))_+\right)$ is equal to 0 if $(1-y_{i,j}\cdot f_j^o(x)) < 0$ and to $y_{i,j}$ if $(1-y_{i,j}\cdot f_j^o(x)) > 0$, whereas if $(1-y_{i,j}\cdot f_j^o(x)) = 0$, $-\overline{\partial_{f_j}}\left((1-y_{i,j}\cdot f_j^o(x))_+\right)$ denotes an element (to be found) either of the set $[0,1]$, when $y_{i,j} = 1$, or of $[-1,0]$, when $y_{i,j} = -1$.

In both cases, due to the presence of the Dirac delta, we end up with

$$f_j^o(x) = \frac{1}{\gamma_j}\sum_{i=1}^m g\otimes\omega_{i,j}^{\geq}(x) = \sum_{i=1}^m \alpha_{i,j}g(x-x_i), \tag{21}$$

where the $\alpha_{i,j}$'s are suitable scalar coefficients (different in the case of hinge or quadratic loss). The classical solution schemes of Ridge Regression and Support Vector Machines can be applied to find the $\alpha_{i,j}$'s.

We conclude with a remark on the notion of constraint reaction in the classic case of supervised learning from examples. In the case of the quadratic loss, it is clear that there is a non-null reaction whenever the associated hard constraint is not satisfied. This happens iff $y_{i,j} \neq f_j^o(x_i)$ for at least one index j. This corresponds to the well-known fact that usually all the examples are support vectors (apart from the case of an interpolating solution). On the opposite side, a set of support vectors that is a proper subset of all the examples usually arises in the hinge loss case.

4.2 Linear Constraints with Supervised Examples Available

Let $\mathscr{X} = \mathbb{R}^d$ and $\forall i \in \mathbb{N}_m$, $\forall x \in \mathscr{X}$ let $\phi_i(f(x)) := a_i' f(x) - b_i(x) = 0$, where $a_i \in \mathbb{R}^n$ and $b_i(x)$ is a real-valued function. We consider hard holonomic bilateral constraints that can be written as $Af(x) = b(x)$, where $A \in \mathbb{R}^{m,n}$, and $b \in \mathscr{C}_0^{2k}(\mathscr{X},\mathbb{R}^m)$ is a smooth vector-valued function with compact support. We assume $n > m$ and $\mathrm{rank}(A) = m$. We discuss the solution for the class of so-called *rotationally-symmetric differential operators P*, as defined by [10]. These are operators of the form $P := [\sqrt{\rho_0}D_0, \sqrt{\rho_1}D_1, \ldots, \sqrt{\rho_\kappa}D_\kappa, \ldots, \sqrt{\rho_k}D_k]'$, where the

operators D_κ satisfy $D_{2r} = \Delta^r = \nabla^{2r}$ and $D_{2r+1} = \nabla\nabla^{2r}$ (Δ denotes the Laplacian operator and ∇ the gradient, with the additional condition $D_0 f = f$, see also [18, 25]), $\rho_0, \rho_1, \ldots, \rho_\kappa, \ldots, \rho_k \geq 0$, and $\rho_0, \rho_k > 0$. Such operators correspond via $L = (P^\star)'P$ to $L = \sum_{\kappa=0}^{k} (-1)^\kappa \rho_\kappa \nabla^{2\kappa}$, which is an invertible operator on $\mathscr{W}^{k,2}(\mathbb{R}^d)$ (see, e.g., Lemma 5.1 in [10]). In addition, we assume that $\gamma = \bar{\gamma} > 0$, where $\bar{\gamma}$ has constant and equal components, and again, an overloaded notation is used. We also assume that m_d additional supervised pairs (x_κ, y_κ) ($\kappa = 1, \ldots, m_d$) induce soft constraints expressed in terms of the quadratic loss. A slight variation of Theorem 2 (see Theorem 3, reported for completeness in Appendix 2, and applied here with $\mu = 1$), implies that a constrained local minimizer f^o of the associated functional satisfies the Euler-Lagrange equations

$$\bar{\gamma} L f^o + A'\lambda + \frac{1}{m_d} \sum_{\kappa=1}^{m_d} (f^o(\cdot) - y_\kappa)\delta(\cdot - x_\kappa) = 0. \tag{22}$$

After some straightforward computations (see Appendix 3 for details), one obtains for the overall constraint reaction (of both hard and soft constraints) the expression

$$\omega(x) = c(x) + \bar{\gamma} \sum_{\kappa=1}^{m_d} Q \alpha_\kappa^{(ql)} \delta(x - x_\kappa),$$

where the $\alpha_\kappa^{(ql)}$'s ($\kappa = 1, \ldots, m_d$) are suitable coefficients to be determined (and "ql" stands for "quadratic loss"), whereas $c(x) := \bar{\gamma} A'(AA')^{-1} L b(x)$ and $Q := I_n - A'[AA']^{-1}A$, where I_n is the identity matrix of size n. Finally, as a unilateral variation of this example, we mention the remarkable case of a unilateral constraint $f(x) \geq 0$ (componentwise), which makes sense when the components of f represent, e.g., mass or probability densities.

4.3 Box Constraints

To fix ideas, let us consider, as a simple sketch, the case of the constraint defined by the rule $\forall x \in \mathscr{B} \subset \mathscr{X} : f(x) - 1 = 0$ [16], with $\mathscr{B} = [a, b] \subset \mathbb{R}$. As depicted in Fig. 2(b), after softening the constraint in the way illustrated by [16], the reaction of the constraint becomes a rectangular impulse, instead of a Dirac distribution as in the case of supervised learning from point-wise examples. However, the latter can be still thought as a degenerate case of the rectangular impulse. For the case in which l in (3) is not finite and the infinite-order differential regularization operator that corresponds to the Gaussian kernel with width σ is used, Theorem 1 (iii) for a single box provides for the solution the representation (up to a positive constant)

$$(g \otimes 1_{\mathscr{B}})(x) \propto erf((x-a)/\sigma) - erf((x-b)/\sigma),$$

where $1_{\mathscr{B}}(\cdot)$ is the characteristic function of $\mathscr{B}$. This clearly indicates that the solution can be thought of as the response of a system with a certain free-space Green's function g, which we call *plain kernel*, to a Dirac delta (supervised pair) or to a

rectangular impulse (box constraint). The latter case is just an example to show that the representation of the solution is not based on the plain kernel anymore, but on a function that arises from its marriage with the reaction of the constraint. Basically, the emergence of plain kernels is just a consequence of the degeneration of the reaction of the constraint to a Dirac distribution.

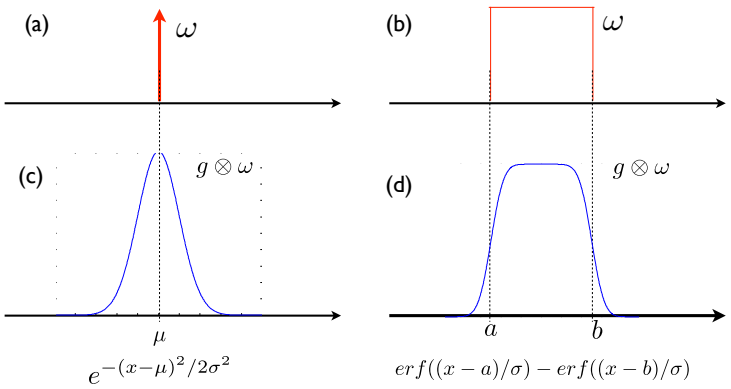

Fig. 2 (a) Constraint reactions corresponding to a classic supervised pair (b) and to the (softened) constraint $\forall x \in [a,b]: \ f(x) = 1$ (box constraint). In (c) and (d) we can see the emergence, resp., of the plain and box kernel. Here, the infinite-order differential regularization operator in (3) is such that the free-space Green's function of L yields the classic Gaussian kernel.

5 Discussion

We have introduced a general framework of learning that involves agents acting in a constraint-based environment, for hard and soft constraints of holonomic type and for soft point-wise constraints. The application of the theory to the chosen case studies illustrates the generality of the approach, which can be fully grasped as we acquire the notion of constraint reaction. The theory presented in this chapter extends the framework of kernel machines to more general hard and soft constraints, and opens the doors to an in-depth re-thinking of the notion of plain kernel that, under some assumptions, was proved to be the Green's function of the differential operator [10] used in the formulation of the learning problem. Interestingly, the notion of constraint reaction and the corresponding representer theorems show that the solution to the learning problem is given in terms of new kinds of kernels that, unlike the plain kernels, also involve the structure of the corresponding constraints: indeed, they originate from the marriage of the plain kernels with the reactions of the constraints. Finally, when the probability density of the data is unknown, the theory suggests to explore the numerical solution of the Euler-Lagrange equations by using unsupervised data, e.g., to learn the probability density itself.

Acknowledgements. G. Gnecco and M. Sanguineti were partially supported by the project "Methodologies for the Approximate Solution of Team Optimization and Strategic Interaction Problems" granted by INDAM-GNAMPA (National Institute of High Mathematics - National Group for Mathematical Analysis, Probability, and Their Application) and the Progetto di Ricerca di Ateneo 2012 "Models and Computational Methods for High-Dimensional Data", granted by the University of Genoa.

Appendix 1

The next lemma, which is a consequence of the Implicit Function Theorem, is exploited in the proof of Theorem 2 in Section 2.2. For a scalar-valued function u of various vector arguments, we denote by $\nabla_i u$ the column vector of partial derivatives of u with respect to all the components of the i-th vector argument. Instead, for a vector-valued function u of various vector arguments, $\nabla_i u$ denotes the matrix whose h-th row is the transpose of the column vector $\nabla_i u_h$.

Lemma 1. *Let $\Omega \subseteq \mathbb{R}^d$, $\mathscr{Y} \subseteq \mathbb{R}^{n_1}$, $\mathscr{Z} \subseteq \mathbb{R}^{n_2}$ be open subsets, and $\phi : \Omega \times \mathscr{Y} \times \mathscr{Z} \to \mathbb{R}^{n_2}$ a given function. Let also $y : \Omega \to \mathscr{Y}$ and $z : \Omega \to \mathscr{Z}$ be other given functions, which satisfy the (vector-valued) holonomic and bilateral constraint*

$$\phi(x, y(x), z(x)) = 0\,, \forall x \in \Omega\,.$$

Suppose also that $\phi \in \mathscr{C}^{k+1}(\Omega \times \mathscr{Y} \times \mathscr{Z}, \mathbb{R}^{n_2})$ for some positive integer $k \geq 1$ and that, for each $x \in \Omega$, the Jacobian matrix

$$\nabla_3 \phi(x, y(x), z(x)) := \begin{pmatrix} \frac{\partial \phi_1(x,y(x),z(x))}{\partial z_1} & \cdots & \frac{\partial \phi_1(x,y(x),z(x))}{\partial z_{n_2}} \\ \cdots & \cdots & \cdots \\ \frac{\partial \phi_{n_2}(x,y(x),z(x))}{\partial z_1} & \cdots & \frac{\partial \phi_{n_2}(x,y(x),z(x))}{\partial z_{n_2}} \end{pmatrix} \tag{23}$$

is nonsingular (possibly after interchanging locally some components of $y(x)$ by an equal number of components of $z(x)$, and redefining the function ϕ and the vectors $y(x)$ and $z(x)$ according to such a replacement). Now, let η_y be an arbitrary function in $\mathscr{C}_0^k(\Omega, \mathbb{R}^{n_1})$ with compact support Ω_C contained in an open ball of sufficiently small radius, and consider a perturbation $\Delta y(x) := \varepsilon \eta_y(x)$ of the function $y(x)$, where $\varepsilon \in \mathbb{R}$ is sufficiently small. Then, there exists a unique function $\eta_z \in \mathscr{C}_0^k(\Omega, \mathbb{R}^{n_2})$ with compact support Ω_C such that the perturbed holonomic and bilateral constraint

$$\phi(x, y(x) + \Delta y(x), z(x) + \Delta z(x)) = 0\,, \forall x \in \Omega$$

is satisfied for $\Delta z(x)$ of the form

$$\Delta z(x) = \varepsilon \eta_z(x) + O(\varepsilon^2)\,, \tag{24}$$

where the "hidden constant" inside the "big O" notation above does not depend[6] *on x, and* $\eta_z(x)$ *has the expression*

$$\eta_z(x) = -(\nabla_3 \phi(x, y(x), z(x)))^{-1} (\nabla_2 \phi(x, y(x), z(x))) \eta_y(x) . \tag{25}$$

Moreover, for each $h \in \{1, \ldots, k\}$ *and* $i \in \{1, \ldots, n_2\}$*, one has, for the i-th component* Δz_i *of* Δz*,*

$$\frac{\partial^h}{\partial x_{j_1} \ldots \partial x_{j_h}} \Delta z_i(x) = \varepsilon \frac{\partial^h}{\partial x_{j_1} \ldots \partial x_{j_h}} \eta_{z_i}(x) + O(\varepsilon^2) , \tag{26}$$

where, again, the "hidden constants" inside the "big O" notations above do not depend on x.

Proof. Fix $x = x_0 \in \Omega$. Since $\phi \in \mathscr{C}^{k+1}(\Omega \times \mathscr{Y} \times \mathscr{Z}, \mathbb{R}^{n_2})$ for $k \geq 1$ and the Jacobian matrix (23) is nonsingular, one can apply the Implicit Function Theorem, according to which, on a suitable open ball $\mathscr{B}$ of $(0,0)$ of sufficiently small radius $\varepsilon > 0$, there exists a unique function $u \in \mathscr{C}^{k+1}(\mathscr{B}, \mathbb{R}^{n_2})$ such that $u(0,0) = 0$ and

$$\phi(x + \Delta x, y(x) + \Delta y, z(x) + u(\Delta x, \Delta y)) = 0 , \quad \forall (\Delta x, \Delta y) \in \mathscr{B} . \tag{27}$$

Moreover, since[7] $k + 1 \geq 2$, each component $u_i(\Delta x, \Delta y)$ of the function $u(\Delta x, \Delta y)$ has the multivariate Taylor expansion

$$u_i(\Delta x, \Delta y) = \sum_{|\alpha|=1} D^\alpha u_i(0,0) (\Delta x, \Delta y)^\alpha + O(\|(\Delta x, \Delta y)\|^2) , \tag{28}$$

where $(\Delta x, \Delta y)^\alpha := \prod_{j=1}^{d} (\Delta x_j)^{\alpha_j} \prod_{j=1}^{n_1} (\Delta y_j)^{\alpha_{d+j}}$, and the term $O(\|(\Delta x, \Delta y)\|^2)$ denotes a function of class $\mathscr{C}^{k+1}(\mathscr{B})$, infinitesimal at $(0,0)$ with order at least 2, where the "hidden" constant inside the "big O" notation above depends only on the local behavior of ϕ on a neighborhood of $(x, y(x), z(x))$, and is independent from x itself, provided that, after the initial choice x_0 for x, x varies inside a

[6] In this formula and in the next one (26) there is, instead, a dependence of the hidden constants on the specific choice of η_y, which may be removed by further assuming $\|\eta_y\|_{\mathscr{C}_0^k(\Omega, \mathbb{R}^{n_1})} \leq M_y$ for some given positive constant M_y.

[7] In the lemma, we have made the assumption $\phi \in \mathscr{C}^{k+1}(\Omega \times \mathscr{Y} \times \mathscr{Z}, \mathbb{R}^{n_2})$ instead of the looser one $\phi \in \mathscr{C}^k(\Omega \times \mathscr{Y} \times \mathscr{Z}, \mathbb{R}^{n_2})$ in order to be able to express the remainder in Taylor's polynomial (28) by the integral Lagrange's form, instead, e.g., of the Peano's form (however, for simplicity of notation, in formula (28) we have not reported the explicit expression of the remainder in the integral Lagrange's form). Considering for simplicity the case of a scalar-valued function $u(x)$ of class $\mathscr{C}^2$ depending on a scalar argument x, we recall that one has the expression

$$f(x + \Delta x) = f(x) + f'(x) \Delta x - \int_0^{\Delta x} (t - \Delta x) f''(x + t) dt ,$$

where the last term is the remainder expressed in the integral Lagrange's form. This formula can be generalized to the multivariate case, and such an extension is used to be able to obtain terms of order $O(\varepsilon^2)$ in (26).

compact subset Ω_C of the projection of the set[8] $\mathscr{B} + (x_0, y(x_0))$ on Ω. Now, let $\eta_y \in \mathscr{C}_0^k(\Omega_C, \mathbb{R}^{n_1}) \subseteq \mathscr{C}_0^k(\Omega, \mathbb{R}^{n_1})$ and set $\Delta x = 0$ and $\Delta y = \Delta y(x) := \varepsilon \eta_y(x)$. Then, we define each component $\Delta z_i(x)$ of the function $\Delta z(x)$ as

$$\begin{aligned} \Delta z_i(x) &:= u_i(0, \varepsilon \eta_y(x)) \\ &= \sum_{|\alpha|=1} D^\alpha u_i(0,0)(0, \varepsilon \eta_y(x))^\alpha + O(\|(0, \varepsilon \eta_y(x))\|^2) \\ &= \varepsilon \sum_{|\alpha|=1} D^\alpha u_i(0,0)(0, \eta_y(x))^\alpha + O(\varepsilon^2), \end{aligned} \tag{29}$$

where the replacement of the term $O(\|(0, \varepsilon \eta_y(x))\|^2)$ by the one $O(\varepsilon^2)$ follows by the fact that $\eta_y(x)$ is fixed and uniformly bounded. Then, (24) follows by setting

$$\eta_{z,i}(x) := \sum_{|\alpha|=1} D^\alpha u_i(0,0)(0, \eta_y(x))^\alpha,$$

which shows that the function $\eta_{z,i}$ is in $\mathscr{C}_0^k(\Omega_C, \mathbb{R}) \subseteq \mathscr{C}_0^k(\Omega, \mathbb{R})$, likewise η_y is in $\mathscr{C}_0^k(\Omega_C, \mathbb{R}^{n_1}) \subseteq \mathscr{C}_0^k(\Omega, \mathbb{R}^{n_1})$. Finally, the application of the Implicit Function Theorem shows also that the vector $\eta_z(x)$ with components $\eta_{z,i}(x)$ has the expression

$$\eta_z(x) = -(\nabla_3 \phi(x, y(x), z(x)))^{-1} (\nabla_2 \phi(x, y(x), z(x))) \eta_y(x).$$

Finally, (26) is derived directly by (24), by computing its partial derivatives of order h (i.e., exploiting the expression of the remainder of Taylor's polynomial (28) in Lagrange's integral form, the rule of differentiation under the integral's sign, the chain rule, and the fact that each component of the function η_y is bounded on Ω_C, together with its partial derivatives - up to the order k - with respect to the components of x). □

The meaning of Lemma 1 is the following: in order to be still able to satisfy the holonomic and bilateral constraint, a perturbation $\Delta y(x) := \varepsilon \eta_y(x)$ of the function $y(x)$ implies a perturbation $\Delta z(x) := \varepsilon \eta_z(x)$ (apart from an infinitesimal of order greater than ε) of the function $z(x)$, where η_z depends only on η_y and suitable partial derivatives of ϕ evaluated at the current solution $(x, y(x), z(x))$, but does not depend on ε. The formula (26) shows that also the partial derivatives of $\Delta z(x)$ up to the order k have similar expressions.

Appendix 2

This appendix reports the complete proof of Theorem 2 in Section 2.2.

Proof. (i) Let f^o be a constrained local minimizer over $\mathscr{F}$ of the functional $\mathscr{E}(f) = \| f \|_{P,\gamma}^2$ defined in formula (3). Fix $x_0 \in \hat{\mathscr{X}}$ and a compact subset $\mathscr{X}_C \subset \hat{\mathscr{X}}$ contained in an open ball of sufficiently small radius, and containing x_0, and, after

[8] Here, we denote by $\mathscr{B} + (x_0, y(x_0))$ the translation of the set $\mathscr{B}$ by $(x_0, y(x_0))$.

performing the permutations σ_ϕ and σ_f, re-order the constraints (and the components of f, resp.) in such a way that the ones with indexes $\sigma_\phi(1),\ldots,\sigma_\phi(m(x_0))$ ($\sigma_f(1),\ldots,\sigma_f(m(x_0))$, resp.) are the first $m(x_0)$ ones. Due to an application of Lemma 1 in Appendix 1, if one fixes arbitrarily the functions $\eta_i \in \mathscr{C}_0^k(\mathscr{X}_C)$ for $i = m(x_0)+1, m(x_0)+2,\ldots,n$, then, for every sufficiently small $|\varepsilon| > 0$, the bilateral holonomic constraints are met for a function f whose components f_j have the following expressions:

$$\begin{aligned}
f_1 &= f_1^o + \varepsilon\eta_1 + O(\varepsilon^2),\\
f_2 &= f_2^o + \varepsilon\eta_2 + O(\varepsilon^2),\\
&\ldots\\
f_{m(x_0)} &= f_{m(x_0)}^o + \varepsilon\eta_{m(x_0)} + O(\varepsilon^2),\\
f_{m(x_0)+1} &= f_{m(x_0)+1}^o + \varepsilon\eta_{m(x_0)+1},\\
f_{m(x_0)+2} &= f_{m(x_0)+2}^o + \varepsilon\eta_{m(x_0)+2},\\
&\ldots\\
f_n &= f_n^o + \varepsilon\eta_n,
\end{aligned} \tag{30}$$

where the functions $\eta_i \in \mathscr{C}_0^k(\mathscr{X}_C)$, for $i = 1,\ldots,m(x_0)$, are still determined by Lemma 1. In particular, by setting $y(x) = [f_{m(x_0)+1}^o(x), f_{m(x_0)+2}^o(x),\ldots,f_n^o(x)]'$, $z(x) = [f_1^o(x),\ldots,f_{m(x_0)}^o(x)]'$, $\phi = [\phi_1,\ldots,\phi_{m(x_0)}]'$, $\eta_y = [\eta_{m(x_0)+1},\eta_{m(x_0)+2},\ldots,\eta_n]'$, and $\eta_z = [\eta_1,\ldots,\eta_{m(x_0)}]'$, one has

$$\eta_z(x) = -(\nabla_3\phi(x,y(x),z(x)))^{-1}(\nabla_2\phi(x,y(x),z(x)))\eta_y(x). \tag{31}$$

Moreover, due to (26), the partial derivatives, up to the order k, of the first $m(x_0)$ components of f, have expressions similar to (30), and contain terms of order $O(\varepsilon^2)$. This implies that $\mathscr{E}(f)$ can be written as

$$\begin{aligned}
\mathscr{E}(f) &= \sum_{j=1}^n \gamma_j < P(f^o+\varepsilon\eta)_j, P(f^o+\varepsilon\eta)_j > + O(\varepsilon^2)\\
&= \sum_{j=1}^n \gamma_j < Pf_j^o, Pf_j^o > + 2\varepsilon \sum_{j=1}^n \gamma_j < Pf_j^o, P\eta_j >\\
&\quad + \varepsilon^2 \sum_{j=1}^n \gamma_j < P\eta_j, P\eta_j > + O(\varepsilon^2)\\
&= \sum_{j=1}^n \gamma_j < Pf_j^o, Pf_j^o > + 2\varepsilon \sum_{j=1}^n \gamma_j < Pf_j^o, P\eta_j > + O(\varepsilon^2).
\end{aligned}$$

Moreover, by an application of Green's formula (see, e.g., Proposition 5.6.2 in [3]), we have

$$< Pf_j^o, P\eta_j > = < (P^\star)' Pf_j^o, \eta_j > = < Lf_j^o, \eta_j >,$$

where $P^\star$ is the formal adjoint of the operator P. Now, we define locally the row vector function $\lambda(x)$ as follows:

$$\lambda(x) := -[\gamma_1(Lf^o)_1(x),\ldots,\gamma_{m(x_0)}(Lf^o)_{m(x_0)}(x)](\nabla_3\phi(x,y(x),z(x)))^{-1}. \quad (32)$$

Then, with such a definition, and exploiting formula (31), one obtains

$$\begin{aligned}\sum_{j=1}^{m(x_0)}\gamma_j < Pf_j^o, P\eta_j > &= \sum_{j=1}^{m(x_0)}\gamma_j < Lf_j^o, \eta_j > \\ &= \int_{\mathscr{X}}\lambda(x)(\nabla_2\phi(x,y(x),z(x)))\eta_y(x)dx.\end{aligned}$$

Summing up, one has

$$\begin{aligned}\mathscr{E}(f)-\mathscr{E}(f^o) = 2\varepsilon\int_{\mathscr{X}}\Big([\gamma_{m(x_0)+1}(Lf^o)_{m(x_0)+1}(x),\ldots,\gamma_n(Lf^o)_n(x)] \\ +\lambda(x)(\nabla_2\phi(x,y(x),z(x)))\Big)\eta_y(x)dx+O(\varepsilon^2).\end{aligned}$$

Now, since $\mathscr{E}(f)-\mathscr{E}(f^o)\geq 0$ for $|\varepsilon|>0$ sufficiently small due to the local optimality of f^o, and $\eta_y\in\mathscr{C}_0^k(\mathscr{X}_C,\mathbb{R}^{n-m(x_0)})$ is arbitrary, by applying the fundamental lemma of the calculus of variations (see, e.g., Section 2.2 in [7]) we conclude that

$$[\gamma_{m(x_0)+1}(Lf^o)_{m(x_0)+1}(x),\ldots,\gamma_n(Lf^o)_n(x)]+\lambda(x)(\nabla_2\phi(x,y(x),z(x)))=0$$

on $\mathscr{X}_C$. This, together with the definition (32) of $\lambda(x)$, shows that (19) holds on $\mathscr{X}_C$. Finally, by varying the point x_0, one obtains (19) on the whole $\hat{\mathscr{X}}$.

(ii) follows by (19), the definition of the free-space Green's function g of L as the solution of $Lg=\delta$ (where δ denotes the Dirac delta, centered in 0), and the stated assumptions on L and g.

(iii) For the case of unilateral constraints, of course the constraints that are inactive in x_0 at local optimality are not taken into account locally, so the condition about the nonsingularity of the Jacobian matrix has to be referred only to the constraints that are active in x_0 at local optimality. Moreover, all the arguments used to derive (i) and (ii) still hold (of course, restricting the analysis to the active constraints in x_0 at local optimality, and replacing the ϕ_i's by the $\check{\phi}_i$'s), since, for every sufficiently small $|\varepsilon|>0$, a function f constructed as in the proof of (i) still satisfies with equality the active constraints in x_0 at local optimality.

Finally, we show that each Lagrange multiplier function $\lambda_i(x)$ is nonpositive. Without loss of generality, we can restrict the analysis to the points of continuity of $\lambda_i(x)$. Suppose by contradiction that there exists one such point $\hat{x}_0\in\hat{\mathscr{X}}$ such that $\lambda_i(\hat{x}_0)>0$. Then, by continuity $\lambda_i(x)>0$ on a sufficiently small open ball centered on $\hat{x}_0$. For simplicity of notation, we also suppose that all the constraints defined on $\hat{x}_0$ are active in $\hat{x}_0$ at local optimality. Then, due to the condition about the

nonsingularity of the Jacobian matrix, there is a vector $u = [u_1, \ldots, u_{m(\hat{x}_0)}]'$ such that $\nabla_3 \breve{\phi}(\hat{x}_0, y(\hat{x}_0), z(\hat{x}_0))u = e_i$, where e_i is a column vector of all 0's, with the exception of the i-th component, which is 1. Then, by an application of the Implicit Function Theorem (likewise in the proof of Lemma 1), for every sufficiently small $\varepsilon > 0$ (but in this case, not for every sufficiently small $\varepsilon < 0$) one can construct a feasible smooth perturbation $f(x)$ of $f^o(x)$ such that its components f_j satisfy

$$\begin{aligned}
f_1(x) &= f_1^o(x) + \varepsilon\eta_1(x) + O(\varepsilon^2), \\
f_2(x) &= f_2^o(x) + \varepsilon\eta_2(x) + O(\varepsilon^2), \\
&\cdots \\
f_{m(\hat{x}_0)}(x) &= f_{m(\hat{x}_0)}^o(x) + \varepsilon\eta_{m(\hat{x}_0)}(x) + O(\varepsilon^2), \\
f_{m(\hat{x}_0)+1}(x) &= f_{m(\hat{x}_0)+1}^o(x), \\
f_{m(\hat{x}_0)+2}(x) &= f_{m(\hat{x}_0)+2}^o(x), \\
&\cdots \\
f_n(x) &= f_n^o(x),
\end{aligned} \tag{33}$$

for suitable functions $\eta_1, \ldots, \eta_{m(\hat{x}_0)} \in \mathscr{C}_0^k(\mathscr{X}_C)$ such that $\eta_1(\hat{x}_0) = u_1$, $\eta_2(\hat{x}_0) = u_2$, $\ldots$, $\eta_{m(\hat{x}_0)}(\hat{x}_0) = u_{m(\hat{x}_0)}$, and such that $\mathscr{E}(f) - \mathscr{E}(f^o)$, apart from an infinitesimal of order $O(\varepsilon^2)$, is directly proportional to

$$\varepsilon[\gamma_1 (Lf^o)_1(\hat{x}_0), \ldots, \gamma_{m(x_0)} (Lf^o)_{m(\hat{x}_0)}(\hat{x}_0)]u = -\varepsilon\lambda(\hat{x}_0)e_i = -\varepsilon\lambda_i(\hat{x}_0) < 0,$$

which contradicts the local optimality of f^o. Then, one has $\lambda_i(\hat{x}_0) \leq 0$. □

The following theorem is a slight variation of Theorem 2, and is exploited in the example in Section 4.2.

Theorem 3. *(*REPRESENTER THEOREM FOR HARD HOLONOMIC CONSTRAINTS MIXED WITH SOFT QUADRATIC POINT-WISE CONSTRAINTS, CASE OF DISTRIBUTIONAL LAGRANGE MULTIPLIERS*). Let us consider the minimization of the functional*

$$\begin{aligned}
\mathscr{L}_s'(f) &:= \frac{1}{2} \parallel f \parallel_{P,\gamma}^2 + \frac{\mu}{m_d} \sum_{\kappa=1}^{m_d} \sum_{j=1}^{n} V_Q(y_{\kappa,j} - f_j(x_\kappa)) \\
&= \frac{1}{2} \sum_{j=1}^{n} \gamma_j < Pf_j, Pf_j > + \frac{\mu}{2m_d} \sum_{\kappa=1}^{m_d} \sum_{j=1}^{n} (y_{\kappa,j} - f_j(x_\kappa))^2
\end{aligned} \tag{34}$$

(a particular case of the functional (4)), where $\mu \geq 0$ and m_d is the number of supervised examples, in the case of $m < n$ hard bilateral constraints of holonomic type, which define the subset

$$\mathscr{F}_\phi := \{f \in \mathscr{F} : \forall i \in \mathbb{N}_m, \forall x \in \mathscr{X}_i \subseteq \mathscr{X} : \phi_i(x, f(x)) = 0\}$$

of the function space $\mathscr{F}$, where $\forall i \in \mathbb{N}_m : \phi_i \in \mathscr{C}^\infty(\mathrm{cl}(\mathscr{X}_i) \times \mathbb{R}^m)$. Let $f^o \in \mathscr{F}_C$ be any constrained local minimizer of (34), and let the holonomic constraints be defined in such a way that either $Lf^o \in \mathscr{C}^0(\mathscr{X}, \mathbb{R}^n)$ or they are of the form $Af(x) = b(x)$, where $A \in \mathbb{R}^{m,n}$ with $m < n$ and $\mathrm{rank}(A) = m$, and $b \in \mathscr{C}_0^{2k}(\mathscr{X}, \mathbb{R}^m)$. Let us assume that for any $\hat{\mathscr{X}}$ and for every x_0 in the same $\hat{\mathscr{X}}$ we can find two permutations σ_f and σ_ϕ of the indexes of the n functions f_j and of the m constraints ϕ_i, such that $\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))}$ refer to the constraints actually defined in x_0, and the Jacobian matrix

$$\frac{\partial(\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))})}{\partial(f^o_{\sigma_f(1)}, \ldots, f^o_{\sigma_f(m(x_0))})}, \tag{35}$$

evaluated in x_0, is not singular. Suppose also that (35) is of class $\mathscr{C}^\infty(\hat{\mathscr{X}}, \mathbb{R}^n)$. Then, the following hold.

(i) There exists a set of distributions λ_i defined on $\hat{\mathscr{X}}$, $i \in \mathbb{N}_m$, such that, in addition to the above constraints, f^o satisfies on $\hat{\mathscr{X}}$ the Euler-Lagrange equations

$$\gamma L f^o + \sum_{i=1}^{m} \lambda_i 1_{\mathscr{X}_i}(\cdot) \cdot \nabla_f \phi_i(\cdot, f^o(\cdot)) + \frac{\mu}{m_d} \sum_{\kappa=1}^{m_d} (f^o(\cdot) - y_\kappa)\delta(\cdot - x_\kappa) = 0, \tag{36}$$

where $\gamma L := [\gamma_1 L, \ldots, \gamma_n L]'$ is a spatially-invariant operator, and $\nabla_f \phi_i$ is the gradient w.r.t. the second vector argument f of the function ϕ_i.

(ii) Let $\gamma^{-1} g := [\gamma_1^{-1} g, \ldots, \gamma_n^{-1} g]'$. If for all i one has $\mathscr{X}_i = \mathscr{X} = \mathbb{R}^d$, L is invertible on $\mathscr{W}^{k,2}(\mathscr{X})$, and there exists a free-space Green's function g of L that belongs to $\mathscr{W}^{k,2}(\mathscr{X})$, then f^o has the representation

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) \vec{\otimes} \phi_i(\cdot, f^o(\cdot)) - \frac{\mu}{m_d} \sum_{\kappa=1}^{m_d} (f^o(\cdot) - y_\kappa)\gamma^{-1} g(\cdot - x_\kappa), \tag{37}$$

where $g \vec{\otimes} \phi_i := g \otimes \omega_i$ and $\omega_i(\cdot) :=\uparrow \phi_i(\cdot, f^o(\cdot)) := -\lambda_i(\cdot) 1_{\mathscr{X}_i}(\cdot) \nabla_f \phi_i(\cdot, f^o(\cdot))$.

(iii) For the case of $m < n$ unilateral constraints of holonomic type, which define the subset $\mathscr{F}_{\check{\phi}} := \left\{ f \in \mathscr{F} : \ \forall i \in \mathbb{N}_m, \forall x \in \mathscr{X}_i \subseteq \mathscr{X}, \ \check{\phi}_i(x, f(x)) \geq 0 \right\}$ of the function space $\mathscr{F}$, (i) and (ii) still hold (with every occurrence of ϕ_i replaced by $\check{\phi}_i$) if one requires the nonsingularity of the Jacobian matrix (see (35)) to hold when restricting the constraints defined in x_0 to the ones that are active in x_0 at local optimality. Moreover, each Lagrange multiplier λ_i is nonpositive and locally equal to 0 *when the correspondent constraint is locally inactive at local optimality.*

Proof. For $\mu = 0$ (or equivalently, when no supervised examples are available) and an additional smoothness assumption on f^o, the theorem reduces to Theorem 2. For the general case $\mu \geq 0$, one can show that the differences with respect to the proof of Theorem 2 are the following:

- there is an additional term $\frac{\mu}{m_d} \sum_{\kappa=1}^{m_d} (f^o(x) - y_\kappa)\delta(x - x_\kappa)$ in the Euler-Lagrange equations, due to the presence of the supervised examples;

- in general, the Lagrange multipliers $\lambda_i(\cdot)$ are not functions, likewise in Theorem 2, but distributions, obtained by a variation of formula (32), which is well-defined in a distributional sense since the Jacobian matrix (35) is locally invertible and infinitely smooth, and since either $Lf^o \in \mathscr{C}^0(\mathscr{X}, \mathbb{R}^n)$ or $Af(x) = b(x)$ hold (with the stated assumptions on A and b). More precisely, formula (32) is replaced by

$$\begin{aligned}\lambda := &- [\gamma_1 (Lf^o)_1, \ldots, \gamma_{m(x_0)} (Lf^o)_{m(x_0)}](\nabla_3 \phi(\cdot, y(\cdot), z(\cdot)))^{-1} \\ &+ \left(\frac{\mu}{m_d} \sum_{\kappa=1}^{m_d} [(y_{\kappa,1} - f_1^o), \ldots, (y_{\kappa,m(x_0)} - f^o_{m(x_0)})] \right. \\ &\left. \delta(\cdot - x_\kappa) \right) (\nabla_3 \phi(\cdot, y(\cdot), z(\cdot)))^{-1}, \end{aligned} \tag{38}$$

  where now λ is a row vector distribution;
- differently from Theorem 2, additional smoothness of f^o is not required, since only (35) is required to be infinitely smooth. □

Appendix 3

This appendix reports the complete derivations for the determination of the constraint reactions in the example of Section 4.2.

Let us determine the vector of distributional Lagrange multipliers λ. We start noting that

$$\begin{aligned} ALf(x) &= A \sum_{\kappa=0}^{k} (-1)^\kappa \rho_\kappa \nabla^{2\kappa} f(x) \\ &= \sum_{\kappa=0}^{k} (-1)^\kappa \rho_\kappa A \nabla^{2\kappa} f(x) \\ &= \sum_{\kappa=0}^{k} (-1)^\kappa \rho_\kappa \nabla^{2\kappa} Af(x) \\ &= \sum_{\kappa=0}^{k} (-1)^\kappa \rho_\kappa \nabla^{2\kappa} b(x) \\ &= Lb(x), \end{aligned}$$

where $Lb \in \mathscr{C}_0^0(\mathscr{X}, \mathbb{R}^m)$ has compact support. Hence, from (22) we get

$$\bar{\gamma} Lb(x) + A \left[A' \lambda(x) + \sum_{\kappa=1}^{m_d} \frac{(f^o(x) - y_\kappa)}{m_d} \delta(x - x_\kappa) \right] = 0.$$

So, the Lagrange multiplier distribution λ is given by

$$\lambda = -[AA']^{-1} \left(\bar{\gamma} Lb + \frac{1}{m_d} \sum_{\kappa=1}^{m_d} A(f^o(\cdot) - y_\kappa) \delta(\cdot - x_\kappa) \right).$$

Now, if we plug this expression for λ into the Euler-Lagrange equations (22), we get

$$\bar{\gamma} L f^o(x) = c(x) + \frac{1}{m_d} \sum_{\kappa=1}^{m_d} Q(y_\kappa - f^o(x)) \delta(x - x_\kappa),$$

where $c(x) := \bar{\gamma} A'(AA')^{-1} Lb(x)$ and $Q := I_n - A'[AA']^{-1} A$. Let $\alpha_\kappa^{(ql)} := \frac{1}{m_d} \bar{\gamma}^{-1}(y_\kappa - f^o(x_\kappa))$. By inverting the operator L, we get

$$f^o(x) = \bar{\gamma}^{-1} \int_{\mathscr{X}} g(\zeta) c(x - \zeta) d\zeta + \sum_{\kappa=1}^{m_d} Q \alpha_\kappa^{(ql)} g(x - x_\kappa). \tag{39}$$

So, the overall constraint reaction (of both hard and soft constraints) is

$$\omega(x) = c(x) + \bar{\gamma} \sum_{\kappa=1}^{m_d} Q \alpha_\kappa^{(ql)} \delta(x - x_\kappa).$$

The coefficients $\alpha_\kappa^{(ql)}$ can be determined by the following scheme. Denote by $y := [y_1, \ldots, y_{m_d}] \in \mathbb{R}^{n,m_d}$ the matrix of targets, where the κ-th column is associated with the corresponding example x_κ, and $\alpha^{(ql)} := [\alpha_1^{(ql)}, \ldots, \alpha_{m_d}^{(ql)}] \in \mathbb{R}^{n,m_d}$. By the definition of $\alpha^{(ql)}$ we get

$$\bar{\gamma} m_d \alpha^{(ql)} + Q \alpha^{(ql)} G = y - \bar{\gamma}^{-1} \int_{\mathscr{X}} g(\zeta) H(\zeta) \, d\zeta \,,$$

where G is the Gram matrix of the input data and the kernel g, and $H : \mathscr{X} \to \mathbb{R}^{n,m_d}$ is the matrix-valued function whose κ-th column is given by the function $c(x_\kappa - \cdot)$. The existence of a solution $\alpha^{(ql)}$ to the linear system above follows by a slight modification of Theorem 1 in [10] (since for $\rho_0 > 0$, $\|\cdot\|_{P,\bar{\gamma}}$ is a Hilbert-space norm on $\mathscr{W}^{k,2}(\mathbb{R}^d)$ by Proposition 3 in [10], and the square loss is convex and continuous) and the nonsingularity of the Jacobian matrix (35) associated with the set of hard constraints $Af(x) = b(x)$.

We conclude discussing the admissibility of the obtained solution (39). By an application of Theorem 3 in [10] about the smoothness properties of free-space Green's functions, it follows that, for this problem, $g \in \mathscr{W}^{k,2}(\mathbb{R}^d)$. This implies that $f^o \in \mathscr{F}$, $\mathscr{L}_s'(f^o)$ is finite, and f^o is a constrained global minimizer, too (thanks to the convexity of the problem).

References

1. Adams, R.A., Fournier, J.F.: Sobolev Spaces, 2nd edn. Academic Press (2003)
2. Argyriou, A., Micchelli, C.A., Pontil, M.: When is there a representer theorem? Vector versus matrix regularizers. Journal of Machine Learning Research 10, 2507–2529 (2009)
3. Attouch, H., Buttazzo, G., Michaille, G.: Variational Analysis in Sobolev and BV Spaces. Applications to PDEs and Optimization. SIAM, Philadelphia (2006)

4. Diligenti, M., Gori, M., Maggini, M., Rigutini, L.: Multitask kernel-based learning with logic constraints. In: Proc. 19th European Conf. on Artificial Intelligence, pp. 433–438 (2010)
5. Diligenti, M., Gori, M., Maggini, M., Rigutini, L.: Bridging logic and kernel machines. Machine Learning 86, 57–88 (2012)
6. Dinuzzo, F., Schoelkopf, B.: The representer theorem for Hilbert spaces: A necessary and sufficient condition. In: Proc. Neural Information Processing Systems (NIPS) Conference, pp. 189–196 (2012)
7. Giaquinta, M., Hildebrand, S.: Calculus of Variations I, vol. 1. Springer (1996)
8. Gnecco, G., Gori, M., Melacci, S., Sanguineti, M.: Learning with hard constraints. In: Mladenov, V., Koprinkova-Hristova, P., Palm, G., Villa, A.E.P., Appollini, B., Kasabov, N. (eds.) ICANN 2013. LNCS, vol. 8131, pp. 146–153. Springer, Heidelberg (2013)
9. Gnecco, G., Gori, M., Melacci, S., Sanguineti, M.: A theoretical framework for supervised learning from regions. Neurocomputing 129, 25–32 (2014)
10. Gnecco, G., Gori, M., Sanguineti, M.: Learning with boundary conditions. Neural Computation 25, 1029–1106 (2013)
11. Gnecco, G., Gori, M., Melacci, S., Sanguineti, M.: Foundations of support constraints machines. Neural Computation (to appear)
12. Gori, M., Melacci, S.: Constraint verification with kernel machines. IEEE Transactions on Neural Networks and Learning Systems 24, 825–831 (2013)
13. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer (2000)
14. Kunapuli, G., Bennett, K.P., Shabbeer, A., Maclin, R., Shavlik, J.: Online knowledge-based support vector machines. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part II. LNCS (LNAI), vol. 6322, pp. 145–161. Springer, Heidelberg (2010)
15. Mangasarian, O.L., Wild, E.W.: Nonlinear knowledge-based classification. IEEE Transactions on Neural Networks 19, 1826–1832 (2008)
16. Melacci, S., Gori, M.: Learning with box kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11), 2680–2692 (2013)
17. Melacci, S., Maggini, M., Gori, M.: Semi–supervised learning with constraints for multi–view object recognition. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009, Part II. LNCS, vol. 5769, pp. 653–662. Springer, Heidelberg (2009)
18. Poggio, T., Girosi, F.: A theory of networks for approximation and learning. Technical report. MIT (1989)
19. Schwartz, L.: Théorie des distributions. Hermann, Paris (1978)
20. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
21. Sun, Z., Zhang, Z., Wang, H., Jiang, M.: Cutting plane method for continuously constrained kernel-based regression. IEEE Transactions on Neural Networks 21, 238–247 (2010)
22. Suykens, J.A.K., Alzate, C., Pelckmans, K.: Primal and dual model representations in kernel-based learning. Statistics Surveys 4, 148–183 (2010)
23. Theodoridis, S., Slavakis, K., Yamada, I.: Adaptive learning in a world of projections. IEEE Signal Processing Magazine 28, 97–123 (2011)
24. Tikhonov, A.N., Arsenin, V.Y.: Solution of ill-posed problems. W.H. Winston, Washington, DC (1977)
25. Yuille, A.L., Grzywacz, N.M.: A mathematical analysis of the motion coherence theory. International Journal of Computer Vision 3, 155–175 (1989)