Foundations of support constraint machines

Foundations of Support Constraint Machines

Giorgio Gnecco[1], Marco Gori[2], Stefano Melacci[2], Marcello Sanguineti[3]

[1] Institute for Advanced Studies (IMT), Lucca
giorgio.gnecco@imtlucca.it

[2] University of Siena
{marco,mela}@diism.unisi.it

[3] University of Genova
marcello.sanguineti@unige.it

Abstract

The mathematical foundations of a new theory for the design of intelligent agents are presented. The proposed learning paradigm is centered around the concept of *constraint*, to represent the interactions with the environment, and the parsimony principle. The classical regularization framework of kernel machines is naturally extended to the case in which the agents interact with a richer environment, where abstract granules of knowledge, compactly described by different linguistic formalisms, can be translated into the unified notion of constraint for defining the hypothesis set. Constrained variational calculus is exploited to derive general representation theorems that provide a description of the *optimal body of the agent* (i.e., the functional structure of the optimal solution to the learning problem), which is the basis for devising new learning algorithms. It is shown that, regardless of the kind of constraints, the optimal body of the agent is a *Support Constraint Machine* (*SCM*) based on representer theorems that extend classical results for kernel machines and provide new representations. In a sense, the expressiveness of constraints yields a *semantic-based regularization theory*, which strongly restricts the hypothesis set of classical regularization. Some guidelines to unify continuous and discrete computational mechanisms are given, so as to accommodate in the same framework various kinds of stimuli, like supervised examples and logic predicates. The proposed view of learning from constraints incorporates classical learning from examples and extends naturally to the case in which the examples are subsets of the input space, which is related to learning propositional logic clauses.

Keywords: kernel machines, learning from constraints, learning with prior knowledge, multi-task learning, support constraints, constrained variational calculus, representer theorems.

Now published in Neural Computation (https://direct.mit.edu/neco)
Full citation: Giorgio Gnecco, Marco Gori, Stefano Melacci, Marcello Sanguineti; Foundations of Support Constraint Machines. Neural Computation (2015) 27 (2): 388–480. DOI: https://doi.org/10.1162/NECO_a_00686]

Contents

List of Figures

List of Tables

Summary of the main notations

In the next table we collect some symbols and notations widely used in the paper.

$\mathcal{X}$	subset of the perceptual space $\mathbb{R}^d$
$\mathcal{F}$	task space
f_j	j-th task of the agent
ϕ	function expressing a holonomic bilateral constraint
$\check{\phi}$	function expressing a holonomic unilateral constraint
Φ	functional expressing an isoperimetric bilateral constraint
$\check{\Phi}$	functional expressing an isoperimetric unilateral constraint
$\mathcal{F}_i \subseteq \mathcal{F}$	set of admissible tasks wrt the i-th constraint
$\mathcal{C}$	constraint collection
$\mathcal{F}_{\mathcal{C}} \subseteq \mathcal{F}$	set of admissible tasks with respect to the constraint collection $\mathcal{C}$
$\mathcal{W}^{k,2}(\mathcal{X})$	Sobolev space, with $k > d/2$
$D^{\alpha}u$	$\frac{\partial^{\lvert\alpha\rvert}}{\partial\alpha_1 \ldots \partial\alpha_n}u$, $\lvert\alpha\rvert := \sum_{j=1}^{n} \alpha_j$
P_i	finite-order linear differential operator $\sum_{\lvert\alpha\rvert \leq k_i} b_{i,\alpha} D^{\alpha}$
P	vectorial finite-order linear differential operator $[P_0, \ldots, P_{l-1}]'$
$P^{\star}$	formal adjoint of P
L	$(P^{\star})'P$
$\parallel f_j \parallel_P^2$	$\langle Pf_j, Pf_j \rangle = \sum_{r=0}^{l-1} \int_{\mathcal{X}} (P_r f_j(x))^2 dx$
$\parallel f \parallel_{P,\gamma}^2$	$\sum_{j=1}^{n} \gamma_j \parallel f_j \parallel_P^2$, $\gamma \in \mathbb{R}^n$ vector of positive components
$\mathcal{E}(\cdot)$	parsimony index $\parallel \cdot \parallel_{P,\gamma}^2$
$\mu_{\mathcal{C}}$	degree of mismatch of $\mathcal{C}$
$\mathcal{E}_{\mathcal{C}}^{\text{soft}}(\cdot)$	$\frac{1}{2} \parallel \cdot \parallel_{P,\gamma}^2 + \mu_{\mathcal{C}}(\cdot)$
f^*	global optimum
f^o	local optimum
g	free-space Green function of an operator
ω_i	reaction of the i-th constraint
ω	overall reaction of the constraints
V_Q	quadratic loss
V_H	hinge loss

Table 1: Main notations and their meanings

1. Introduction

1.1 Motivations

This work focuses on the open issue of designing intelligent agents with effective learning capabilities in complex environments, where sensorial data are combined with knowledge-based descriptions of the tasks. Unlike the classical framework of learning from examples, in those cases the beauty and the elegance of simplicity behind the *parsimony principle*[1] has not been profitably used yet for the formulation of systematic theories of learning. Most solutions are essentially based on hybrid systems, in which there is a mere combination of different modules that are separately charged of handling the prior knowledge on the problem at hand and providing the inductive behavior naturally required in some tasks. The investigation of more unified approaches is not only of interest per se, but also, and perhaps primarily, because this crafting of knowledge with learning can give rise to interesting induction/deduction processes that are likely to be very effective in complex real-world problems. Our theory is centered around the parsimony principle: we consider intelligent agents interacting with constraints in a multi-task environment, with the purpose of developing the simplest (smoothest) vectorial function in a set of feasible solutions.

A first insight into the idea of learning from constraints on which this paper is based was given in (Gori (2009)), but the first systematic study along this direction is given in (Diligenti et al. (2012)), 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 (Diligenti et al. (2010); Saccà et al. (2011b,a); Diligenti et al. (2011)). An application to text categorization that involves supervised learning and prior knowledge has been addressed in (Frandina et al. (2012)), while a first attempt to perform First-Order Logic (FOL) verification can be found in (Gori and Melacci (2013)). Kernel-based representations of the optimal solutions to constrained learning problems have been used also in other cases, where the prior knowledge is not given in terms of logic expressions. Remarkable results have been obtained by imposing the classification consistency of different views of the same object (Melacci et al. (2009)) and probabilistic constraints in (Melacci and Gori (2011)). A preliminary study on the benefit deriving from the restriction to convex constraints is given in (Gori and Melacci (2010)). An in-depth analysis of the special case of constraints deriving from propositional descriptions is given in (Melacci and Gori (2013)). Theoretical results for some kinds of constraints are contained in (Gnecco et al. (2013a)). The regularization principles adopted in this paper were exploited in (Gnecco et al. (2013b)), together with tools from Statistical Learning Theory, to investigate learning from constraints expressed as boundary conditions.

The main objective of this study consists in developing mathematical foundations for the paradigm of learning from constraints, with the aim of providing the mathematical and algorithmic apparatus (Section 5) for facing typical machine tasks encountered in applications, such as those presented in Section 5.3 (see, in particular, Table 4).

1.2 Types of constraints and their roles

In order to provide a unified context for manipulating perceptual data and granules of knowledge, we propose to use the unifying concept of *constraint.* It is sufficiently general to represent different kinds of sensorial data along with their relations, as well as to express abstract knowledge on the tasks and to embrace logic descriptions. Examples of constraints come out naturally in different contexts: one might want to enforce the probabilistic normalization of a set of functions modeling a classification task, the probabilistic normalization of a density function, or to impose coherent decisions of the classifiers acting on different views of the same pattern (Melacci et al. (2009)). The expressive power of constraints becomes more significant when dealing with specific problems coming from, amongst others, vision, control, text classification, ranking in hyper-textual environment, and prediction of the stock market. While the linguistic description to express a constraint can be of many different types, including those based on logic formalisms, in order to describe knowledge granules we can always end up with real-valued multi-variable functions involving the inputs and the learning tasks.

We propose to build an interaction amongst different tasks by introducing various kinds of constraints; they are summarized in the following definition, which follows the terminology used in variational calculus.

Definition 1 [TYPES OF CONSTRAINTS] *Let $\mathcal{X}$ denote a subset of the perceptual space $\mathbb{R}^d$, $\mathcal{F}$ a space of functions $f : \mathcal{X} \to \mathbb{R}^n$, $\mathcal{X}_i$ open subsets of $\mathcal{X}$, $\phi_i : \mathcal{X}_i \times \mathbb{R}^n \to \mathbb{R}$ and $\check{\phi}_i : \mathcal{X}_i \times \mathbb{R}^n \to \mathbb{R}$ continuous functions, $\Phi_i : \mathcal{F} \to \mathbb{R}$ and $\check{\Phi}_i : \mathcal{F} \to \mathbb{R}$*

1. This principle has been massively exploited in decision making to promote the preference for the least complex explanation of observations, as well as for deriving laws of Nature, especially in physics (Basdevant (2006)).

continuous functionals, and $m_H, m_I, \check{m}_H$, and $\check{m}_I$ positive integers.

$i.$ Holonomic (ho) bilateral (bi) : $\forall x \in \mathcal{X}_i \subseteq \mathcal{X} : \ \phi_i(x, f(x)) = 0, \ i = 1, \ldots, m_H$.

$ii.$ Holonomic (ho) unilateral (un) : $\forall x \in \mathcal{X}_i \subseteq \mathcal{X} : \ \check{\phi}_i(x, f(x)) \geq 0, \ i = 1, \ldots, \check{m}_H$.

$iii.$ Isoperimetric (is) bilateral (bi) : $\Phi_i(f) = 0, \ i = 1, \ldots, m_I$.

$iv.$ Isoperimetric (is) unilateral (un) : $\check{\Phi}_i(f) \geq 0, \ i = 1, \ldots, \check{m}_I$.

$v., vi.$ Pointwise (pw) bilateral (bi) and pointwise (pw) unilateral (un) : as i. and ii., respectively, with each $\mathcal{X}_i$ made up of finitely many points (in this case, the continuity of ϕ_i - respectively, of $\check{\phi}_i$ - is required with respect to the second vector argument) .

For notational simplicity, when dealing with constraints of the same type, the notations "m_H", "m_I", "$\check{m}_H$" and "$\check{m}_I$" will be replaced simply by "m". It is worth remarking that holonomic constraints express local properties, since they hold $\forall x \in \mathcal{X}_i$. Instead, isoperimetric constraints express global properties of f (apart from degenerate cases, such as the one in which they are expressed by integrand functions[2] with compact support)[3].

We consider both the case in which perfect constraint satisfaction on a whole subset of the perceptual space is required, and the case where constraint violations are allowed, at the cost of some penalization, quantified by a loss. The former situation corresponds to a *hard (hr) interpretation* of the constraints, whereas the latter to their *soft (sf) interpretation*. For the sake of simplicity and with a little abuse of terminology, we refer to these cases as *hard constraints* and *soft constraints*, respectively. Soft constraints arise often in real-word problems, e.g., in case of collections of supervised data in which labeling errors are quite common. Moreover, sometimes it may be desirable a total fulfillment (i.e., hard constraints), which, however, may be too difficult to achieve. So, the original problem with hard constraints is approximated by a sequence of problems with soft constraints, whose optimal solutions converge to the "hard" one (Luenberger (1969)). In summary, the theory that we propose is based on using both supervised pairs and hard/soft constraints of more general nature.

Based on the notion of constraint, we can accommodate into the same framework stimuli of very different kinds, like those shown in Table 2. The labels $(ho, nh, is, bi, un, pw, hr, sf)$ are used to classify the different constraints. The classification considers three categorical variables that represent the specific local/global nature (pw, ho, nh, and is stand for *pointwise*, *holonomic*, *non-holonomic*, and *isoperimetric*, respectively), the *bilateral/unilateral* formulation (bi, un), and the *hard/soft* enforcement (hr, sf). It is worth mentioning that sometimes a given problem can receive different classifications and, consequently, can be attacked using the corresponding different methodologies. For example, the classical learning from examples corresponds to constraints that can be regarded either as pointwise (see Table 2-i) or special isoperimetric. The second interpretation is specifically covered in Section 4.1. The table gives examples of constraints that are described in column 2 either by informal (i through vi) or formal languages (vii and $viii$). The informal description in the table is kept concise but could be much more detailed. Column 3 contains a translation into constraints re-written by real-valued functions, that are the formal environmental description of our agents.

The example i describes the simplest case in which we handle the classical pair (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 is the function that the agent is expected to compute, the corresponding real-valued representation of the constraint, which is reported in column 3, is just the translation of the classical "robust" sign agreement between the target and the function to be learned.

2. This is the case of the functions ψ_i in (14) (see Section 2), when they have compact support.

3. The meaning of "holonomic" is "integrable". Constraints of the forms

$$\forall x \in \mathcal{X}_i \subseteq \mathcal{X} : \ \phi_i(x, f(x), Qf(x)) = 0, \ i = 1, \ldots, m_{NH}, \tag{1}$$

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

where Q is a (vector) linear differential operator, which cannot be transformed by integration into the holonomic forms of Definition 1 $i.$ and $ii.$, respectively, are called *non-holonomic constraints* (nh). They will not be considered in this paper. Any holonomic bilateral constraint can be transformed by differentiation into a constraint of the form (1), and one can re-gain the original constraint by integration. However, the vice-versa does not hold true (Giaquinta and Hildebrand (1996), p. 98). The name "isoperimetric" derives from the classical isoperimetric problem of the calculus of variations, which consists in determining a plane figure of the largest possible area, whose boundary is constrained to have a specified length.

Examples ii and iii are classical probabilistic normalizations, while example iv imposes the coherence between the decisions taken on x_1 and x_2, for the object x, where x_1 and x_2 are two different views of the same object x (see Melacci et al. (2009)).

Example v describes the constraints needed to impose consistency in portfolio asset allocation when investing money (USD and Euro) in bonds and stocks. Here f_c^d, f_b^d, f_s^d denote the allocations in cash, bond, and stock in USD on the basis of the financial feature vector x, while f_c^e, f_b^e, f_s^e are the corresponding allocations in Euro. The constraints simply express the consistency imposed by the overall amount of available money, denoted by T (in USD), being c the Euro/USD conversion factor.

In the example vi, we report a constraint coming from computer vision, concerning the classical problem of determining the optical flow. It consists in finding the smoothest solution for the velocity field under the constraint that the brightness of any point in the movement pattern is constant. The smoothness of the velocity field can be measured according to formula (11) introduced in Section 2, by choosing a linear differential operator such as the gradient or the Laplacian. If $u(x, y, t)$ and $v(x, y, t)$ denote the components of the velocity field and $E(x, y, t)$ denotes the brightness of any pixel (x, y) at time t, then the velocity field satisfies the linear constraint reported in Table 2 vi (Horn and Schunck (1981)).

In the example vii, the i-th rule expresses propositions on some linguistic features $P1$, $P2$, $P5$, $N5$, $V0$, $V1$ detected by proper linguistic tests for the diagnosis of Wernicke's aphasia (Tsakonasa et al. (2004)). The propositional description of column 2 can be equivalently split into a supervised pair $(\mathcal{X}_i, y_{we}^i)$, where $y_{we}^i = 1$ and $\mathcal{X}_i = (3, +\infty) \times (-\infty, 4] \times (2, \infty) \times (-\infty, 22] \times (-\infty, 62] \times (38, +\infty)$. Basically, as it will be shown in Section 2, any proposition i in this example can be associated with a correspondent set $\mathcal{X}_i \in 2^{\mathcal{X}}$, where $\mathcal{X} := \mathbb{R}^6 = \{(P1, P2, P5, N5, V0, V1)\}$ is the feature space. The last row of the table refers to a document classification problem and states that all papers dealing with numerical analysis and neural networks are also machine learning papers. Notice that whereas column 2 expresses the rule by a logic description, in column 3 there is a related constraint expressed by real-valued functions, according to the classical product T-norm (Klement et al. (2000); Diligenti et al. (2010, 2012)).

1.3 Lex parsimoniae principles

A specific interpretation of the parsimony principle, inspired by Occam's razor, provides the unified methodology for the construction of the theory of learning from constraints presented in this paper. The parsimony principle has been used in decision making by promoting the idea that simple explanations are preferable, and it has been proposed in that context as a mysterious heuristics to guide towards the understanding of the laws of Nature. Although in the scientific method, the preference for the simplest explanation is not necessarily considered an irrefutable principle and, therefore, it is arguable and falsifiable on the basis of experimental results, typically the laws derived via the parsimony principle provide very accurate descriptions of reality. The principle of least action, the studies on electromagnetic fields, the development of quantum mechanics are, amongst others, some of the many examples of successful application of the *lex parsimoniae* to physics (Basdevant (2006)).

There are arguments to claim that Occam's razor should not be regarded as a theory in the classical sense of being a model that explains physical observations. According to James Gleick (Gleick (1992), pp. 60-61), "Where Newton's methods left scientists with a feeling of comprehension, minimum principles left a sense of mystery", which is nicely reinforced in David Park's challenging question: "How does the ball know which path to choose?". The experimental validation is the only way to claim that the principle holds. Aesthetic considerations are not enough to claim that a ball or a planet are condemned to follow a predetermined path. This issue becomes central when trying to capture complex cognitive processes or to conceive models for decision making. We can think of cognitive processes that obey a sort of "cognitive law" based on the minimization of a generalized version of the Dirichlet integral in Physics, that yields simplicity in its optimal solution. Interestingly, within this context, the above Park's question on mechanics sheds light on the truly inductive nature of Occam's razor. The introduction of constraints, along with quantifiers, opens the doors to an interpretation of the parsimony principle that better resembles physical laws like Netwon's laws. Indeed, constraints translate rules that are not only expected to hold for single examples. As it will be shown, a remarkable improvement of the learning paradigm developed in this paper with respect to classical frameworks of learning from examples, is that these "cognitive laws" can naturally be checked on tons of unsupervised examples.

The idea of simplicity that is generally expressed by Occam's razor can be given different formulations, which are partially related. Within the framework of the Minimum Description Length (MDL) principle, which gives preference to models that lead to the best compression of the data, the invariance theorem states that, for a long sequence, the difference between any two optimal descriptions of it is negligible compared to the size of the se-

Examples of constraints				
#	*linguistic description*	*real-valued representation*	*classification*	*typical interpretation*
$i.$	i-th supervised pair for classification	$y_\kappa \cdot f(x_\kappa) - 1 \geq 0$	(pw,un)	(sf)
$ii.$	probabilistic normalization for classification	$\forall x \in \mathcal{X}$: $f_1(x) + f_2(x) + f_3(x) = 1$;	(ho,bi)	(hr)
		$\forall x \in \mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \mathcal{X}$: $f_j(x) \geq 0$ ($j = 1, 2, 3$)	(ho,un)	(hr)
$iii.$	probabilistic normalization of a density function	$\int_{\mathcal{X}} f(x)dx = 1$;	(is,bi)	(hr)
		$\forall x \in \mathcal{X}$: $f(x) \geq 0$	(ho,un)	(hr)
$iv.$	coherence constraint (2 classes)	$\forall x = (x_1, x_2) \in \mathcal{X}$: $f_1(x_1) \cdot f_2(x_2) \geq 0$	(ho,un)	(sf)
$v.$	asset allocation cash, bond, and stock in USD cash, bond, and stock in Euro overall invest. in USD and Euro	$\forall x \in \mathcal{X}$: $f_c^d(x) + f_b^d(x) + f_s^d(x) = t_d(x)$; $f_c^e(x) + f_b^e(x) + f_s^e(x) = t_e(x)$; $t_d(x) + c \cdot t_e(x) = T$	(ho,bi)	(hr)
$vi.$	optical flow	$\frac{\partial E}{\partial x}u + \frac{\partial E}{\partial y}v + \frac{\partial E}{\partial t} = 0$	(ho,bi)	(sf)
$vii.$	Wernicke's aphasia (i-th) rule if $P1 > 3$ and $P2 \leq 4$ and $P5 > 2$ and $N5 \leq 22$ and $V0 \leq 62$ and $V1 > 38$ then W	$\forall x \in \mathcal{X}_i$: $y_{we}^i \cdot f_{we}(x) - 1 \geq 0$	(ho,un)	(sf)
$viii.$	document classification: $\forall x : na(x) \wedge nn(x) \Rightarrow ml(x)$	$f_{na}(x) \cdot f_{nn}(x) \cdot (1 - f_{ml}(x)) \leq \epsilon$ ($\epsilon > 0$ and $\epsilon \simeq 0$)	(ho,un)	(sf)

Table 2: Examples of constraints from different environments. The first four are problem-independent, whereas the others are descriptions of granules of knowledge in specific domains. The classification of the constraints in the second to last column concerns local/global nature (pw, ho, nh, and is stand for *pointwise*, *holonomic*, *non-holonomic*, and *isoperimetric*, respectively), the *bilateral/unilateral* formulation (bi, un), and the *hard/soft* enforcement (hr, sf). The classification hard vs soft constraint in the last column is referred to the way in which the constraints are usually dealt with in practice.

quence (Kolmogorov (1965); Solomonoff (1964); Chaitin (1966)), which gave rise to the notion of universal code and Kolmogorov complexity. However, the concept of simplest explanation given by Vapnik in (Vapnik (1998), Chapter 6) with the introduction of the concept of structural risk minimization, rooted in the Vapnik-Chervonenkis (VC) dimension, is likely to be the better starting point to grasp the idea of simplicity followed in our approach (which, like for support vectors machines, is not simply associated with the number of free parameters).

Regardless of the contiguities with other approaches and the specific discussion on parsimony issues, the proposed approach is primarily related to kernel machines. Our investigation was stimulated by the idea that, so far, restricting to environments based on examples only has given quite a limited picture of the potential use of the parsimony principle, as the examples alone are not a very expressive description of most interesting cognitive tasks.

1.4 Overview of the main results

The present paper summarizes and cements a series of works from the authors, by providing a unifying perspective and contextually new theoretical results. The paper presents a new learning paradigm, which we call *Support Constraint Machines (SCMs)*, and can be regarded as a generalization of standard Support Vector Machines (SVMs) and other kernel methods, where the classical supervision based on training samples is interpreted in terms of pointwise constraints. Based on this novel view, new types of possibly non-pointwise constraints are taken into account, and a classification of constraints that can be handled by the new learning paradigm is provided. Similarly to SVMs and other kernel methods, learning takes place by finding a function that best satisfies the constraints (in a hard or soft way) and is as smooth as possible, according to an integro-differential regularization term. Such a function is assumed to belong to a Cartesian product of Sobolev spaces of finite order (although extensions to the infinite-order case are possible, modeling, e.g., the case of a Gaussian kernel), and is searched for as a solution to the Euler-Lagrange equations induced by the problem formulation.

The work makes use of variational calculus to show that the optimal solution to the constrained learning problem has an additive decomposition into several terms, each related to a specific constraint and given as the convolution between the free-space Green function associated to the differential regularization operator and a so-called *constraint-reaction function*, which, roughly speaking, takes the role of the SVM's dual variable associated with the constraint. The paper also investigates in detail contexts in which the convolution can be carried out easily (e.g., in the case of SVMs) and situations for which it cannot. The latter cases are actually more interesting as they open a new research direction, in which the optimal solution to the learning problem has to be numerically found as the solution of a system of partial differential equations.

In our approach, an agent performs an induction process under a regularization mechanism that goes beyond the classical regularization based on supervised examples only. To give an insight into the way we measure the smoothness of a feasible solution, let us consider, as an instance, an open set $\mathcal{X} \subseteq \mathbb{R}$, scalar-valued admissible functions $f : \mathcal{X} \to \mathbb{R}$ of one variable and, for $b_0, b_1 \geq 0$, the minimization of the parsimony index

$$\| f \|^2 := b_0 \int_{\mathcal{X}} f^2(x)dx + b_1 \int_{\mathcal{X}} \left(\frac{df}{dx}\right)^2 dx, \tag{3}$$

which is the square of a seminorm in the Sobolev space of functions on $\mathcal{X}$ that are square-integrable together with their partial derivatives up to the order k (it is a norm when $b_0, b_1 > 0$). In Section 2.1, we extend this setting to more general norms of the form $\| Pf \|$, where P is a linear differential operator, and also to vectorial functions f. Basically, in this case the learning problem that we address consists in discovering appropriate functions f in a Sobolev space that are required to minimize $\| f \|^2$, while fulfilling a given set of hard/soft constraints. 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, which are included in the objective of the optimization problem together with the parsimony index. Likewise for kernel machines, it is of special interest the situation in which there is a unique optimal solution to the learning problem, which is proven to hold for a very large class of regularization operators, regardless of specific assumptions on the shape of the optimal solution.

To fully appreciate the difference between this kind of regularization operators with respect to classical kernels, we start noting that any problem of learning from examples can directly be formulated in our framework, by replacing the traditional loss with one that measures the fulfillment of the constraints. This usually requires collecting unsupervised data to check the constraints. It represents indeed a general way of facing learning from constraints, that in this paper is fully covered and referred to as *learning from pointwise constraints* and has already been the subject of investigation (see, e.g., (Diligenti et al. (2012))).

As it will be shown, a significant class of constraints can be expressed as $\forall x \in \mathcal{X} : \quad \phi(x, f(x)) = 0$ (or a generalization of this case, in which the set $\mathcal{X}$ is replaced by its subset $\mathcal{X}_i$) and, in one instance of the proposed learning paradigm, a parsimonious agent is required to fulfill such constraints, while minimizing $\| f \|$. Notice, on passing, that unlike the case of supervised learning in which the constraints are clearly imposed on the training set, this class of constraints is expressed by quantifiers on infinite domains. In the paper we prove that, for several problems of learning from constraints, the optimal solution is fully representable by using the following notion of *constraint reaction*:

$$\begin{aligned} &i. \quad \omega^{\text{hard}}(x) := -\lambda(x) \cdot \nabla_f \phi(x, f(x)) \,, \\ &ii. \quad \omega^{\text{soft}}(x) := -p(x) \cdot \nabla_f \phi(x, f(x)), \end{aligned} \tag{4}$$

where ∇_f denotes the gradient with respect to the second vector argument of ϕ, computed at $f(x)$, $\lambda(x)$ is a Lagrange multiplier function associated with the constraint, and $p(x)$ is a probability density (see Fig. 1). The case $i.$ refers to a hard constraint, whereas the case $ii.$ to a soft one. Interestingly, the two constraint reactions have exactly the same dependence on the gradient of the constraint. However, whereas in the case of a hard constraint the Lagrange multiplier function associated with the constraint ϕ needs to be determined so as to impose the hard fulfillment of the constraint, for soft constraints the role of $\lambda(x)$ is replaced by the probability density $p(x)$. Both functions $\lambda(\cdot)$ and $p(\cdot)$ play crucial roles in determining the magnitude of the constraint reactions. For a given point x, in the case of hard fulfilment the weight $\lambda(x)$ needs to be discovered in a way not to violate the constraint at x, whereas in case of soft fulfillment $p(x)$ is the typical weight that is large in regions of high probability. An important result given in the paper concerns the representation of an optimal solution to the problem of learning from constraints. Recall that the free-space Green function g associated to a linear differential operator O is a solution to the distributional differential equation $Og = \delta$, where δ is the Dirac distribution, centered on the origin. In the paper we prove that, if g is the (free-space) Green function of the operator $L := (P^\star)'P$, being P a (vectorial) differential regularization operator and $P^\star$ its formal adjoint (the symbol $'$ denotes transposition), then under suitable conditions the functional representation of an optimal solution[4] is[5]

$$f^\star(x) = \sum_{\kappa=1}^{m} (g * \omega_\kappa)(x) \,, \tag{5}$$

where "$*$" denotes the convolution operator, m is the number of constraints, and ω_κ is the *reaction of the k-th constraint*. Interestingly, the Fourier transforms[6] of both sides yield

$$\hat{f}^\star(\xi) = \sum_{\kappa=1}^{m} \hat{g}(\xi) \cdot \hat{\omega}_\kappa(\xi) \,. \tag{6}$$

The basic results on the representation of the optimal solution are given in Sections 3.1 - 3.3, for various classes of constraints. Under some assumptions, g is also the kernel of a Reproducing Kernel Hilbert Space (RKHS) Berlinet and Thomas-Agnan (2004).

In equations (5) and (6), we can easily recognize a structure that very much resembles the one arising from the representer theorems of kernel machines (see, e.g., Schölkopf and Smola (2002); Dinuzzo and Schölkopf (2012)). Compared to other classical representer theorems, those obtained in this paper provide necessary optimality conditions expressed in the form of (distributional) partial differential equations. This is due to the choice of a Sobolev space in the definition of the ambient space of the optimization problem, and of a regularization term expressed in integro-differential form. So, one can exploit in this framework Green functions and other tools from functional analysis and variational calculus, which are not standardly used in machine learning problems. A similar approach was adopted in (Poggio and Girosi (1989)), but only for soft pointwise constraints, and in the absence of hard constraints.

4. We denote by $f^\star$ a globally optimal solution to the constrained learning problem ($f^\star$ is a particular instance of a locally optimal solution, denoted by f^o, which is actually the case for which most of the equations presented in this section are derived in the paper). Moreover, under certain circumstances (in particular, for convex problems), any f° is also a globally optimal solution.
5. For the sake of simplicity, the linear differential operator herein considered is not scaled by a factor γ as in the general definition (11), i.e., equation (5) refers to the case $\gamma = 1$.
6. We use the unitary definition of the Fourier transform in terms of the frequency vector $\xi \in \mathbb{R}^d$, i.e., $\hat{g}(\xi) := \int_{\mathbb{R}^d} g(x) \exp(-\iota\langle 2\pi\xi, x\rangle) dx$ and $g(x) := \int_{\mathbb{R}^d} \hat{g}(\xi) \exp(\iota\langle 2\pi\xi, x\rangle) d\xi$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^d$.

In Section 4.1, it is proven that the functional representation provided by equation (5) collapses perfectly to the kernel expansion of kernel machines, where the constraints are restricted to supervised pairs (x_κ, y_κ). Indeed, we prove that in such a case the global constraint reaction (i.e., the sum of the reactions of all the constraints) can be written as

$$\omega(x) = \sum_{\kappa=1}^{m} \alpha_\kappa \delta(x - x_\kappa), \tag{7}$$

where the α_κ's are real coefficients, and we have made the simplified assumption of dealing with a scalar-valued function f. Consequently, in the Fourier domain, we have

$$\hat{\omega}(\xi) = \sum_{\kappa=1}^{m} \alpha_\kappa e^{-2\pi\imath x_\kappa \xi}\,. \tag{8}$$

Hence, it turns out that the weights of the obtained kernel machine can be thought of as the coefficients α_κ of the Fourier transform of the global constraint reaction. The constraint reaction associated with a single supervised pair is shown in Fig. 2-a.

In Fig. 2-c we can see the response to the κ-th constraint dictated by the Green function g of the regularization operator. Such a response has the form (or is proportional to) $g(x) * \delta(x - x_\kappa) = g(x - x_\kappa)$. We call such a function *plain kernel*. In general, when quantifiers are involved, the Fourier transform does not get the simple structure of equation (8), which reveals an intriguing peculiarity of classical supervised learning (more generally, of learning with pointwise constrants). The connection with classical kernel machines vanishes when considering other constraints (i.e., holonomic constraints), since in such cases the plain kernel that arises as the response to the impulse $\delta(x - x_\kappa)$ is replaced by $(g * \omega_k)(x)$. Basically, this new response comes out from the "marriage" of the plain kernel with the reaction of the constraint, which contributes to defining the class of functions to be used for the representation of the optimal solution to the constrained learning problem.

Let us consider, as an example, the soft relaxation of the so-called *box constraint*, i.e., $\forall x \in \mathcal{B} \subset \mathcal{X}: \;\; f(x) - 1 = 0$ and, for the sake of simplicity[7], let $\mathcal{B} := [a, b] \subset \mathbb{R}$ for some $a < b$. As depicted in Fig. 2-b, in this case the reaction of the constraint is a rectangular impulse, instead of a Dirac distribution. If, for $\sigma > 0$, we use the differential regularization operator

$$L = \sum_{\kappa=0}^{\infty} (-1)^\kappa \frac{\sigma^{2\kappa}}{\kappa!\, 2^\kappa} \nabla^{2\kappa} \tag{9}$$

(associated with a suitable operator P), whose Green function is the Gaussian, then for a single box the general functional representation given by equation (5) yields for the optimal solution the expression

$$(g * 1_{\mathcal{B}})(x) \propto \operatorname{erf}((x - a)/\sigma) - \operatorname{erf}((x - b)/\sigma),$$

where $1_{\mathcal{B}}$ is the characteristic function of $\mathcal{B}$,

$$\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

is the error function, g is the Green function of the operator $L := (P^\star)'P$, and $P^\star$ is the formal adjoint of P.

The two examples above clearly indicate that the functional representation of the optimal solution can be thought of as the response of a system with a certain Green function g (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 often the representation of the optimal solution is not based on the plain kernel anymore, but on a function that arises from its "marriage" with the reaction of the constraint. In (Melacci and Gori (2013)), it is proven that, under certain conditions, also $g * 1_{\mathcal{B}}$ is in fact the kernel of a RKHS and, therefore, all the related mathematical and algorithmic apparatus of kernel machines can be directly re-used. It is also shown therein that significant experimental results are achieved in cases where prior knowledge of this form is naturally available (e.g., for the case of medical applications, rules provided by physicians). Basically, the emergence of the plain kernel as the Green function of the operator L is just the consequence of the degeneration of the reaction of the constraint to a Dirac distribution, which happens in the case of a supervised pair.

7. In Section 4, the results are established in general.

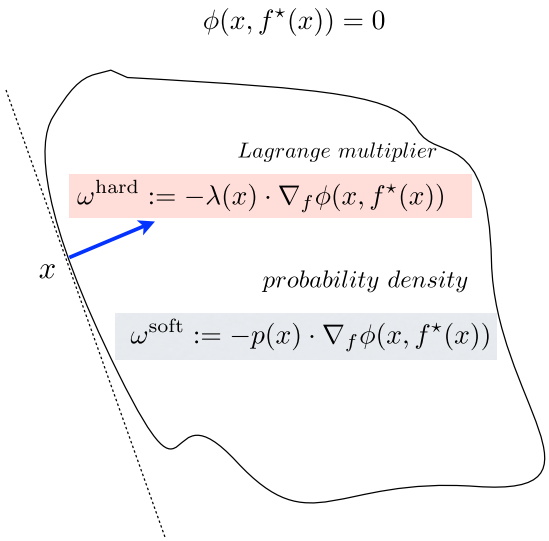


Figure 1: Constraint reactions in the generic point $x \in \mathcal{X}$ corresponding to hard and soft constraints, where one can see the roles of the Lagrange multiplier function and of the probability density in the same point x for hard and soft constraints, respectively. For illustrative purposes, the case $n = 2$ is considered here. The reaction of the constraint ϕ in x is a vector orthogonal to the level curve $\phi(x, f^\star(x)) = 0$, where, in the definition of the level curve, $\phi(x, \cdot)$ is interpreted as a function of its second vector-valued argument only.

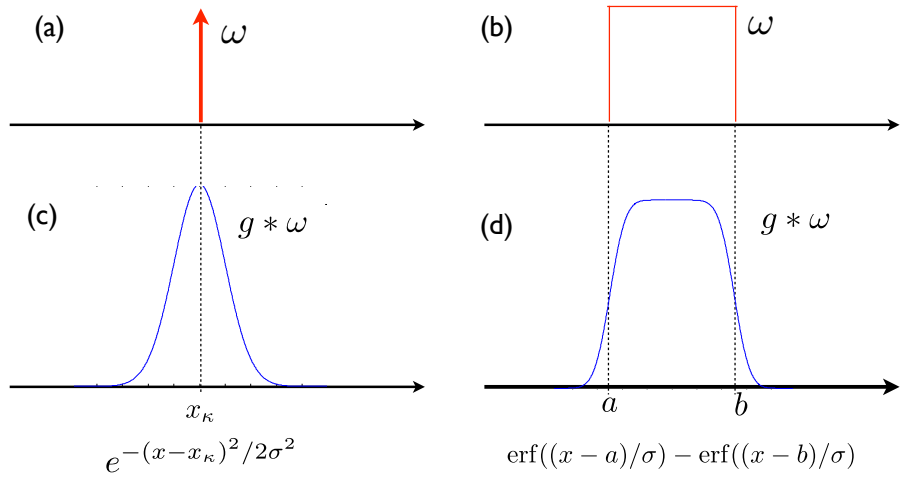


Figure 2: Constraint reactions corresponding (a) to a classical supervised pair and (b) to the soft constraint $\forall x \in [a, b] : \ f(x) = 1$ (box constraint). In (c) and (d) we can see the emergence of the plain and the box kernels, respectively. The regularization operator of equation (9) is used, which for supervised learning yields the classical plain Gaussian kernel. The representation (5) (here, for $m = 1$) leads to a new kernel-based optimal solution and prescribes its form (box kernel).

Constraint Reactions and Kernels		
type of constraint	*reaction*	*optimal solution*
supervised pair (x_κ, y_κ)	$\delta(x - x_\kappa)$	$g(x) * \delta(x - x_\kappa) \propto e^{-(x-x_\kappa)^2/2\sigma^2}$
box constraint	$1_{[a,b]}$	$g(x) * 1_{[a,b]}(x) \propto \mathrm{erf}((x-a)/\sigma) - \mathrm{erf}((x-b)/\sigma)$
$[Af(x) = b(x)]_i$	$a_i \left([AA']^{-1}Lb(x)\right)_i$	$g(x) * a_i \left([AA']^{-1}Lb(x)\right)_i \propto e^{-x^2/2\sigma^2} * a_i \left([AA']^{-1}Lb(x)\right)_i$

Table 3: Different types of constraints with the corresponding reactions and optimal solutions. The cases refer to the differential regularization operator of equation (9), which for supervised learning yields the classical plain kernel.

As already pointed out, the general representation (5) of the optimal solution does not hold only for soft constraints, but also for hard constraints (see also Table 2, item v). In Section 4.3.2 we prove that in this case, for linear hard constraints of the form $Af(x) = b(x)$ (where A is a given matrix and $b(x)$ is a given smooth function and with compact support), the reaction of the i-th constraint has the expression

$$\omega_i(x) \propto a_i \left([AA']^{-1}Lb(x)\right)_i ,$$

where the transpose a_i' of the column vector a_i is the i-th row of the matrix A, and $\left([AA']^{-1}Lb(x)\right)_i$ denotes the i-th element of the column vector $[AA']^{-1}Lb(x)$.

For the three examples above, the reactions of the constraints and the corresponding functional representations of the optimal solutions are shown in Table 3. In the case of supervised learning, the reaction of the constraint associated with a single supervised pair (x_κ, y_κ) is proportional to $\delta(x-x_\kappa)$ and the corresponding Fourier transform $\hat{\omega}_\kappa$ has constant absolute value. In the second case, the reaction is a rectangle function, and its Fourier transform is a window-like function (more precisely, a sinc function). Intuitively, in this case, the more we stress the universal quantification over $\mathcal{X}$, the more the spectrum of the constraint reaction is peaked. Finally, in the last case, the constraint reaction depends on $Lb(x)$ (which is approximated by a constant when $b(x)$ is "nearly" constant in some portion of interest of the domain[8], and in such a case its Fourier transform is approximated by a Dirac distribution). The proofs of these and related results are given in Section 4, where we apply the notion of constraint reaction.

Summing up, the proposed framework can be given a twofold interpretation. First, any constraint can be thought of yet another stimulus, just like a supervised input-target pair. Hence, the development of an agent in such a learning environment consists in finding a parsimonious solution compatible with the given constraints. Second, we can think of the constraints as a way to provide an additional prior with respect to the one introduced by the classical smoothness assumptions on the hypothesis space. Unlike supervised examples, any constraint quantified on an infinite domain turns out to represent a sort of rule derived from the environment, which offers a semantic statement on the tasks. Hence, an agent learning in such a framework is subjected to priors that give rise to a sort of *semantic-based regularization*.

1.5 Algorithmic issues

The results that we obtain on the structure of the optimal solution of the constrained learning problem ("body of the agent") are not only important in themselves but they also play a crucial role for the construction of learning algorithms. Basically, whenever we determine the structure of the optimal solution, we often identify a dependence on parameters that leads to bridge most classical machine learning algorithms. To sum up, the cases of Table 3, even when we consider a mixture of constraints, can be attacked by using the classical kernel machine mathematical and algorithmic apparatus (see Section 4 for details). Unfortunately, an explicit expression of the constraint reaction likewise in Table 3 is not always possible, and numerical approximations are needed in that case.

The general solution of the given constrained learning problem leads to representations that stimulate the development of learning algorithms. This is summarized in Fig. 3, where we distinguish between kernel-based approaches and solutions based on equation (5). As already mentioned, for some learning problems, the most

8. See Section 4.3 for a discussion about an extension to the case of a function $b(x)$ that is constant on the whole input space (hence, its support is not compact).

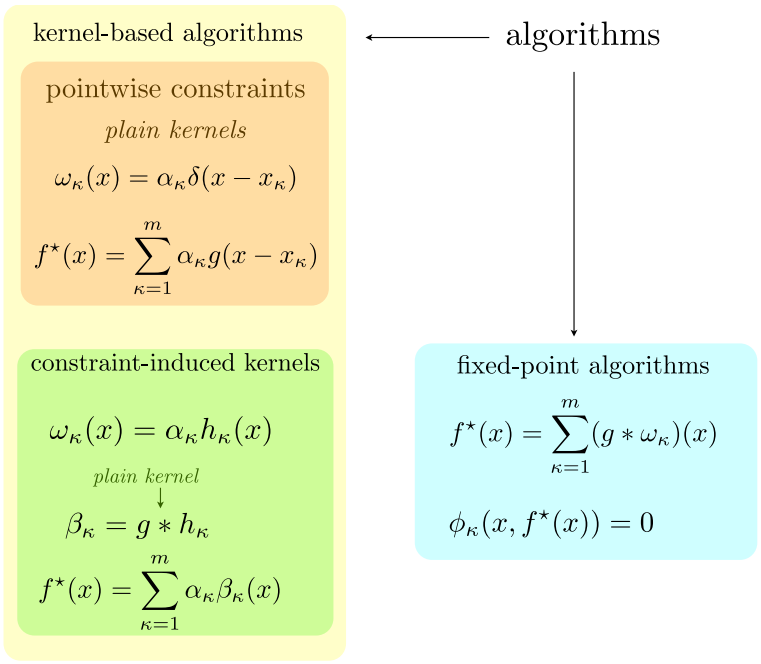


Figure 3: Three different approaches leading to the development of learning algorithms. Besides the case of kernel-based algorithms, this paper gives hints to devise algorithms based on constraint-induced kernels and on the direct solution of fixed-point functional equations.

straightforward approximation of the optimal solution can be gained by quantizing the constraints over a given set of unsupervised data, which yields a representation based on an expansion in terms of the plain kernel (see Section 3.2, Theorem 22). This approach, which has already been the subject of in-depth investigation, can be experimented also by a software simulator that is available at

`https://sites.google.com/site/semanticbasedregularization/home/software`

where any pointwise soft constraint can be considered. From an algorithmic point of view, our results suggest also the two approaches depicted in the lower part of Fig. 3.

- For the case in which each constraint reaction takes on the form $\omega_\kappa(x) = \alpha_\kappa h_\kappa(x)$, where α_κ is a coefficient (to be determined) and $h_\kappa(\cdot)$ is a known function, we can construct the new *constraint-induced kernel* $g * h_\kappa$, which, likewise the box kernel, comes from the "marriage" of the plain kernel and the κ-th constraint. In so doing, one can still rely on the usual kernel machine apparatus for the concrete algorithmic development.
- In other cases, one can also learn the optimal solution $f^\star$ as a *fixed point* of equation (5). This is a more general approach, as it applies also to the case in which the reaction ω_κ to the κ-th constraint has an unknown structure. Such a case will be briefly discussed in Section 6, along with a new perspective of on-line learning.

1.6 Related literature

The variational formulation of learning proposed in this work is mostly inspired by Poggio and Girosi in (Poggio and Girosi (1989)). They offered a clear variational formulation of learning single tasks, which opens the doors to fruitful developments in the area of kernel machines. Under Gaussian assumptions on the probability density of the data, they also nicely pointed out that their formulation of learning, which is extended in this paper, has some intriguing connections with expressing a prior in Bayesian learning (see also (Kaipio and Somersalo (1994), pp. 79-89) for Gaussian smoothness priors). Basically, the presence of a generalized version of the Dirichlet integral in the performance index of the agent, which favors the development of "simple solutions", can be thought of as a prior within the Bayesian framework. In general, however, the smoothing of the solution is not interchangeable with priors and it is somehow related to luckiness functions (Herbrich and Williamson (2002)).

The notion of simplicity in our agents emerges from the classical way of facing inverse problems, which can be traced back to (Wahba (1975); Tikhonov and Arsenin (1977)). Interestingly, the extended formulation that we develop to encompass the concept of constraints requires to solve a Fredholm equation of the II kind, which has been massively investigated in the context of inverse problems (see, e.g., (Kaipio and Somersalo (1994), Chapter 2)). The same issue is central in inverse imaging, where it is required to reconstruct an object from its image (see, e.g., (Bertero and Boccacci (1998), Chapter 8)).

The context in which our agents operate is the one of multi-task learning, brought to the attention of the machine learning community in (Caruana (1997)), and is related to recent studies in such a field; see, e.g., (Caponnetto et al. (2008); Argyriou et al. (2007); Evgeniou et al. (2005)). However, following (Poggio and Girosi (1989)), here we adopt a measure of parsimony that is based on linear differential operators instead than directly on kernels of RKHSs, which allows us to deal naturally with quantifiers on infinite sets. To this end, we extend the studies on approximation and learning presented in (Poggio and Girosi (1989); Girosi et al. (1995)) to the case in which the agent interacts with general hard/soft constraints instead of the classical interaction restricted to supervised examples (see also Girosi et al. (2000); Chen and Haykin (2002)). The two approaches of inducing parsimony via differential operations vs via kernels are connected as follows: the optimal solutions to certain learning problems containing regularization terms that depend on linear differential operators can be written in terms of Green functions of related differential operators and, under certain assumptions, such Green functions are also kernels of RKHSs[9] (see, e.g., (Smola et al. (1998); Schölkopf and Smola (1998); Gnecco et al. (2013b))). When invoking this connection, we see that, unlike the above-mentioned approaches to multi-task kernels, our agents do not capture cross-dependencies amongst different tasks at the level of the choice of the kernel, but they express any such dependence via constraints. More precisely, the constraint-induced kernels, which emerge from the marriage of the plain kernel with the reactions of the constraints, are entrusted the role of capturing the cross-dependencies. This kind of dependence between tasks and, in general, the idea of injecting prior knowledge into the learning process, goes into the direction of probabilistic inductive logic programming (De Raedt et al. (2008)).

This paper is organized as follows. In Section 2 we introduce our formulation of learning from constraints in a variational framework. In Section 3, first we investigate the structures of the optimal solutions to the problem of learning hard, soft, and mixed constraints via suitable Euler-Lagrange equations, by providing the corresponding representer theorems. Then, we introduce the basic concepts of constraint reactions and support constraints, which open the doors to the new learning paradigm of Support Constraint Machines (SCMs). Section 4 applies the results to some relevant case studies, while Section 5 addresses the algorithmic side of the theory. Finally, Section 6 contains a perspective view and a discussion on the application of the theory. Some technical lemmas are collected in the Appendix.

2. Learning from constraints

In this section we introduce a formulation of learning from constraints based on a variational formulation of the parsimony principle, which aims at keeping small a functional that involves the function to be learned and its derivatives up to some order, via suitable linear differential operators. Interestingly, it is different - yet related - to the classical approach of kernel machines. The bridge between the proposed approach and the classical learning theory in RKHSs is represented by the fact that, as already mentioned, the Green functions of certain linear differential operators are also kernels of certain RKHSs (this issue is addressed in Gnecco et al. (2013b); see Section 7, Section 11, and Table 1 therein).

2.1 Task space and parsimonious agents

Definition 2 [AGENT, TASKS, AND TASK SPACE] *We think of an intelligent agent acting on a subset $\mathcal{X}$ of the* perceptual space $\mathbb{R}^d$ *as one implementing a vectorial function $f := [f_1, \ldots, f_n]' \in \mathcal{F}$, where $\mathcal{F}$ is a space of functions from $\mathcal{X}$ to $\mathbb{R}^n$. The function f_j, $j = 1, \ldots, n$, is called the j-th* task of the agent *and $\mathcal{F}$ the* task space.

9. However, it is not true that any kernel of a RKHS can be expressed as the Green function of a linear differential operator. For instance, a polynomial kernel cannot be a Green function even for an infinite-order differential operator, as it has nonzero partial derivatives only up to a finite order. We refer the interested reader to (Gnecco et al. (2013b), Section 11 and Table 1) for examples of kernels that either can or cannot be expressed as Green functions of linear differential operators.

For technical reasons, related to the theory of Sobolev spaces (Adams and Fournier (2003)), in the following 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 (Adams and Fournier, 2003, 4.9, p. 83). In particular, we consider the case in which, $\forall j \in \mathbb{N}_n := \{1, \ldots, n\}$ and some positive integer k, the function $f_j : \mathcal{X} \to \mathbb{R}$, belongs to the Sobolev space $\mathcal{W}^{k,2}(\mathcal{X})$, i.e., the subset of $\mathcal{L}^2(\mathcal{X})$ whose elements f_j have weak partial derivatives up to the order k with finite $\mathcal{L}^2(\mathcal{X})$-norms, and we require $k > d/2$. We express this choice in the next definition, which, unless stated otherwise, is supposed to hold throughout the paper. By the superscript $'$ we denote transposition.

Definition 3 [HARD/SOFT CONSTRAINT SATISFACTION] *The interaction between the agent and the environment is modeled by constraints that have to be strictly satisfied (hard constraints) and/or constraints that can be violated, at the cost of some penalization quantified by a loss (soft constraints).*

In Definition 3, we slightly abuse the terminology. Although the qualifiers "hard" and "soft" regard the way in which a constraint is interpreted, depending on the application context, for the sake of simplicity we refer to "hard constraints" and "soft constraints", respectively.

Definition 4 [CHOICE OF THE TASK SPACE] *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be open, $f_j : \mathcal{X} \to \mathbb{R}$ the j-th task of the agent, $j = 1, \ldots, n$, and $f := [f_1, \ldots, f_n]'$. Then $f \in \mathcal{F}$, where the task space is chosen as*

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

with $k > d/2$.

By the Sobolev Embedding Theorem (see, e.g., (Adams and Fournier (2003), Chapter 4)), the requirement $k > d/2$ implies that each element of $\mathcal{W}^{k,2}(\mathcal{X})$ has a unique bounded and continuous representative, on which the constraints are well-defined[10]. Moreover, by the Sobolev Embedding Theorem, for $k > d/2$ the space $\mathcal{F}$ is a RKHS, too (see also (Berlinet and Thomas-Agnan (2004), Chapter 6) for a proof)[11].

To define our learning model, we introduce a *parsimony index*, defined via a differential operator. We linear consider differential operators that are invariant under spatial shift and have constant coefficients, as summarized in the next definition. We use the following notation. For a function u and a multi-index α with n non-negative components α_j, we write $D^\alpha u$ to denote $\frac{\partial^{|\alpha|}}{\partial\alpha_1 \ldots \partial\alpha_n} u$, where $|\alpha| := \sum_{j=1}^n \alpha_j$.

Definition 5 [FINITE-ORDER LINEAR DIFFERENTIAL OPERATOR] *We define the* vectorial finite-order linear differential operator $P := [P_0, \ldots, P_{l-1}]'$ *as the l-tuple of operators P_i, $i = 0, \ldots, l-1$, acting on the Sobolev space $\mathcal{W}^{k,2}(\mathcal{X})$ and such that*

$$P_i := \sum_{|\alpha| \le k_i} b_{i,\alpha} D^\alpha \,,$$

where $k_i \le k$ and the $b_{i,\alpha}$'s are suitable real coefficients. The formal adjoint of P is defined as the operator $P^\star = [P_0^\star, \ldots, P_{l-1}^\star]'$ whose i-th component $P_i^\star$ has the form

$$P_i^\star := \sum_{|\alpha| \le k_i} (-1)^{|\alpha|} b_{i,\alpha} D^\alpha \,.$$

We also define the operator

$$L := (P^\star)' P \,.$$

10. Indeed, each element of $\mathcal{W}^{k,2}(\mathcal{X})$ is an equivalence class of functions, which differ on sets of zero Lebesgue measure. Knowing that each such equivalence class has a unique bounded and continuous representative allows one to define the constraints unambiguously, i.e., to evaluate them on such a representative element (otherwise, some constraints, such as pointwise ones, might be ambiguously defined, since it would not be clear on which representative element they should be evaluated).
11. When the condition $k > d/2$ is violated, one can easily construct functions belonging to $\mathcal{W}^{k,2}(\mathcal{X})$ and containing peaks of infinite amplitude (see, e.g., (Stein (1970), p. 132), and (Girosi and Anzellotti (1993))). This shows that in such cases $\mathcal{W}^{k,2}(\mathcal{X})$ is not a RKHS since, by one of its equivalent definitions, a RKHS is a Hilbert space of functions for which all evaluation functionals are bounded (see, e.g., (Berlinet and Thomas-Agnan (2004))).

For two functions $u_1, u_2 : \mathcal{X} \to \mathbb{R}$ such that the right-hand side of (10) is well-defined and finite, we let

$$\langle u^{(1)}, u^{(2)} \rangle := \int_{\mathcal{X}} u^{(1)}(x) u^{(2)}(x)\, dx\,. \tag{10}$$

Definition 6 [PARSIMONY INDEX] *Let $P := [P_0, \ldots, P_{l-1}]'$ be a finite-order linear differential operator, $\| f_j \|_P^2 := \langle Pf_j, Pf_j \rangle := \sum_{r=0}^{l-1} \int_{\mathcal{X}} (P_r f_j(x))^2 dx$, and $\gamma \in \mathbb{R}^n$ a vector of positive components. We endow the task space $\mathcal{F} := \underbrace{\mathcal{W}^{k,2}(\mathcal{X}) \times \ldots \times \mathcal{W}^{k,2}(\mathcal{X})}_{n \text{ times}}$ with a seminorm*

$$\| f \|_{P,\gamma} := \left(\sum_{j=1}^{n} \gamma_j \| f_j \|_P^2 \right)^{1/2} = \left(\sum_{j=1}^{n} \gamma_j \sum_{r=0}^{l-1} \int_{\mathcal{X}} (P_r f_j(x))^2 dx \right)^{1/2}.$$

The parsimony index is given by the functional

$$\mathcal{E}(\cdot) := \| \cdot \|_{P,\gamma}^2\,. \tag{11}$$

Equation (3) provides a simple example of a parsimony index of the form (11).

Definition 7 [PARSIMONIOUS AGENT] *A parsimonious agent that interacts with the environment aims at minimizing over the task space the functional $\mathcal{E}$. More specifically, in the case of hard constraints the functional (11) has to be minimized on the subset of $\mathcal{F}$ of the task space that satisfies the hard constraints, whereas in the soft case one has to minimize on $\mathcal{F}$ the sum of (11) and a suitable penalty term, which quantifies the violation of the given set of soft constraints. In the case of hard constraints mixed with soft ones, the sum above is minimized on the subset of $\mathcal{F}$ of the task space that satisfies the hard constraints.*

Notice that, in principle, any constraint can be regarded as hard or soft, depending on the way it is encoded in the optimization problem that models learning with constraints.

Our formulation is a generalization to multi-task learning of what was proposed in (Poggio and Girosi (1989)) for regularization networks with soft constraints associated with supervised examples only, whose definition is rooted in Tikhonov's regularization theory (Tikhonov and Arsenin (1977)) and related studies on spline functions (Wahba (1975, 1990)); see also Hadamard's seminal paper on well-posedness (Hadamard (1902)). This relies on a generalized version of the *Dirichlet integral*, which plays a fundamental role in many classical physic laws. In the case $n = 1$, the operator P has been related to the notion of kernel in a RKHS, too (see, e.g., Schölkopf and Smola (1998); Smola et al. (1998); Gnecco et al. (2013b)), and in (Evgenious et al. (1999)) there are some relevant links between regularization networks and kernel machines.

The representer theorems that we derive in Section 3 are based on the existence of a free-space Green function for the operator $L := (P^\star)'P$ introduced in Definition 5. It has to be remarked that, when P is a finite-order linear differential operator, as stated in Definition 5, a Gaussian cannot be the free-space Green function of L. Indeed, if g were a Gaussian, then the first hand-side of the distributional differential equation $Lg = \delta$ would be smooth (differently from the right-hand side). So, our theory does not cover directly the Gaussian kernel, which has been used, nevertheless, in the introductory example of Section 1.4. In Section 3.6, we discuss a way to circumvent this problem.

If we choose for the operator P the form used in Tikhonov's stabilizing functionals (Tikhonov and Arsenin (1977)), for $n = 1$ (i.e., there is only one task $f : \mathcal{X} \to \mathbb{R}$) and $l = k + 1$ we get

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

where the function $\rho_r(x)$ is non-negative, $P_r := \sqrt{\rho_r(x)} D_r$, and D_r denotes a linear differential operator with constant coefficients and containing only partial derivatives of order r. An interesting case is the one in which $\rho_r(x) \equiv \rho_r \geq 0$ for every $x \in \mathcal{X}$ and the differential operator satisfies

$$D_{2r} = \Delta^r = \nabla^{2r} \tag{12}$$

and

$$D_{2r+1} = \nabla \Delta^r = \nabla \nabla^{2r}\,, \tag{13}$$

where $\Delta := \nabla^2$ denotes the Laplacian operator and ∇ the gradient, with the additional condition $D_0 f = f$ (see Poggio and Girosi (1989); Yuille and Grzywacz (1989)). For instance, $D_1 = \nabla$, $D_2 = \Delta = \nabla^2$, and $D_3 = \nabla\Delta = \nabla\nabla^2$. According to (11), when $n > 1$ the operator P acts separately[12] on all the components of f, i.e., $Pf := [Pf_1, Pf_2, \ldots, Pf_n]'$. Then, unlike what is done in (Micchelli and Pontil (2005); Argyriou et al. (2007); Evgeniou et al. (2005)), the interaction among the components $f_j^\star$ of an optimal solution $f^\star$ to the learning problem is modeled by the fulfillment of the constraints, not by the presence of the differential operator P in the parsimony index.

2.2 Admissible tasks and constraint transformations

The kinds of constraints that we consider in this paper have been summarized in Definition 1; see Table 2 for several instances. We use the symbol $\mathcal{C}$ to denote a collection of constraints. For instance, $\mathcal{C} := \left\{\check{\phi}_i,\ i \in \mathbb{N}_{\check{m}_H}\right\}$ is a collection of unilateral holonomic constraints.

Definition 8 [ADMISSIBLE TASKS] *The set $\mathcal{F}_i \subseteq \mathcal{F}$ of the functions belonging to the task space $\mathcal{F}$ and compatible with the i-th constraint is called the* set of admissible tasks wrt the i-th constraint. *For a constraint collection $\mathcal{C}$, the set $\mathcal{F}_\mathcal{C} \subseteq \mathcal{F}$ of the functions belonging to the task space $\mathcal{F}$ and compatible with all the constraints in $\mathcal{C}$ is called the* set of admissible tasks *and its functions are the* admissible tasks.

A specially interesting case of the isoperimetric bilateral constraints introduced in Definition 1 (iii) is

$$\Phi_i(f) = \int_\mathcal{X} \psi_i(x, f(x))dx. \tag{14}$$

where $\psi_i : \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$, i.e., when $\Phi_i(\cdot)$ is an integral functional. A similar remark holds for $\check{\Phi}_i(\cdot)$ in the case of isoperimetric unilateral constraints (see Definition 1 (iv)). When the set $\mathcal{X}_i$ is made up of a single point, under mild smoothness assumptions any pointwise constraint defined in that point can be expressed in terms of an equivalent isoperimetric one by replacing ψ_i in (14) by a Dirac delta. For instance, the bilateral pointwise constraint $\phi_i(x_i, f(x_i)) = 0$ is equivalent to

$$\int_\mathcal{X} \phi_i(x, f(x))\delta(x - x_i)dx = 0,$$

provided that ϕ_i and f are continuous. Often, pointwise constraints can be interpreted as discretizations of constraints of holonomic type. In machine learning applications, they may be associated, e.g., with both supervised and unsupervised training examples.

Any holonomic bilateral constraint $\phi_i(x, f(x)) = 0$ can be expressed in terms of the pair of unilateral constraints $(\phi_i(x, f(x)) \geq 0, -\phi_i(x, f(x)) \geq 0)$. Likewise, unilateral constraints can be expressed in terms of an appropriate choice of bilateral constraints. To see this, let $(u)_+ := \max\{0, u\}$. Then, the unilateral constraint $\check{\phi}_i(x, f(x)) \geq 0$ is equivalent to $(-\check{\phi}_i(x, f(x)))_+ = 0$ and also to $\left((-\check{\phi}_i(x, f(x)))_+\right)^2 = 0$. This equivalence should be treated with care when applying the classical theory of Lagrange multipliers (see Sections 3.1 - 3.3), since it requires some properties that might be lost in making such a transformation. For instance, the reduction of a unilateral constraint $\check{\phi}_i(x, f(x)) \geq 0$ to the corresponding bilateral one $(-\check{\phi}_i(x, f(x)))_+ = 0$ may cause the loss of the differentiability. So, the direct extension of theoretical results from the case of bilateral constraints to unilateral ones is not always feasible in a plain way.

2.3 Learning from hard and soft constraints

We denote by $1_{\mathcal{F}_i} : \mathcal{F}_i \to \{0, 1\}$ the characteristic function(al) of the set of functions $\mathcal{F}_i$, i.e., $f \in \mathcal{F}_i \Leftrightarrow 1_{\mathcal{F}_i}(f) = 1$. Likewise, we use $1_{\mathcal{X}_i}(\cdot)$ to denote the characteristic function of the set $\mathcal{X}_i$ when $\mathcal{X}_i$ is open. In order to keep the notation uniform for pointwise constraints, we let $1_{\mathcal{X}_i}(\cdot) := \sum_{j=1}^{|\mathcal{X}_i|} \delta(\cdot - x_{(i,j)})$ for a finite set $\mathcal{X}_i$ made up of the points $x_{(i,j)}$ $(j = 1, \ldots, |\mathcal{X}_i|)$, where δ denotes the Dirac's delta; so, for a set $\mathcal{X}_i$ made up only of one point x_i, one has $1_{\mathcal{X}_i}(\cdot) := \delta(\cdot - x_i)$.

The following definition formalizes the problems of learning from hard constraints.

12. Basically, in this case we overload the notation and use the symbol P both for the (matrix) linear differential operator acting on f and for the (vector) differential operator acting on its components.

Definition 9 [LEARNING FROM HARD CONSTRAINTS] *Let $\mathcal{F}_{\mathcal{C}} \subseteq \mathcal{F}$ be the subset of the functions that belong to the given function space $\mathcal{F}$ and are compatible with a given collection $\mathcal{C}$ of constraints. The problem of determining a (local or gobal) minimizer of the functional $\mathcal{E}(\cdot) := \| \cdot \|^2_{P,\gamma}$ over $\mathcal{F}_{\mathcal{C}}$ is called* learning from the hard constraint collection $\mathcal{C}$.

Now we turn our attention to the soft case, focusing on holonomic constraints. The softness of pointwise and isoperimetric constraints can be directly understood once we have the notion for holonomic constraints. Whereas isoperimetric constraints yield directly a measure of their violation (given, for instance, by $|\Phi_i(f)|$ and $|(-\check{\Phi}_i(f))_+|$), for holonomic constraints we can express a global degree of mismatch in terms of some given (possibly generalized[13]) data probability density $p(\cdot)$ (more generally, a weighted average of their degree of violation). For example, for a continuous holonomic constraint ϕ_i associated with an open subset $\mathcal{X}_i$ of $\mathcal{X}$ and $q \in \mathbb{N}^+$, one can express the global degree of mismatch as

$$\int_{\mathcal{X}_i} |\phi_i(x, f(x))|^q p(x) dx \tag{15}$$

and, for a continuous unilateral holonomic constraint $\check{\phi}_i$, as

$$\int_{\mathcal{X}_i} |(-\check{\phi}_i(x, f(x)))_+|^q p(x) dx \,. \tag{16}$$

However, when different kinds of constraints are involved, the quantities in equations (15) and (16) might not satisfactorily represent the interactions of the agent with the environment. Basically, the constraints come with their own specificity and the agent might be willing to express a belief on them. Hence, whereas (15) and (16) represent inherent degrees of mismatch of ϕ_i, which depend on the probability density, it is useful to the belief of a soft constraint.

Definition 10 [BELIEF OF A SOFT CONSTRAINT] *Given the i-th constraint of a collection $\mathcal{C}$ of soft constraints, its* belief *is defined as follows.*

- *If isoperimetric: a non-negative constant β_i.*
- *If holonomic: either a function from $\mathcal{X}_i$ to $\mathbb{R}^+$ or a linear combination of Dirac's deltas with positive coefficients.*
- *If pointwise: a vector of $|\mathcal{X}_i|$ non-negative constants.*

For the pointwise case, the belief can be reduced to the one for the holonomic case, when the latter is expressed by a linear combination of Dirac's deltas with positive coefficients. So, in the following definition, we deal only with isoperimetric and holonomic constraints.

Definition 11 [DEGREE OF MISMATCH OF A SOFT CONSTRAINT] *Given the i-th constraint of a collection $\mathcal{C}$ of soft constraints (possibly of different kinds), its q*th-order degree of mismatch *is defined as follows.*

- *If isoperimetric bilateral: $\mu^{(q)}_{\Phi_i}(f) := |\Phi_i(f)|^q \, \beta_i$.*
- *If isoperimetric unilateral: $\mu^{(q)}_{\check{\Phi}_i}(f) := |(-\check{\Phi}_i(f))_+|^q \, \beta_i$.*
- *If holonomic bilateral: $\mu^{(q)}_{\phi_i}(f) := \int_{\mathcal{X}_i} |\phi_i(x, f(x))|^q \, \beta_i(x) \, p(x) \, dx$.*
- *If holonomic unilateral: $\mu^{(q)}_{\check{\phi}_i}(f) := \int_{\mathcal{X}_i} |(-\check{\phi}_i(x, f(x)))_+|^q \, \beta_i(x) \, p(x) \, dx$.*

The qth-order degree of mismatch of $\mathcal{C}$, denoted by $\mu^{(q)}_{\mathcal{C}}(f)$, is the summation of the degrees of mismatch of each constraint in $\mathcal{C}$. In this case, by $\mu^{(q)}_{i,\mathcal{C}}(f)$ we denote the qth-order degree of mismatch of the i-th constraint in $\mathcal{C}$.

Of course, $\mu^{(q)}_{\mathcal{C}}(f) = 0 \Leftrightarrow \mathcal{C}$ is a collection of constraints that are strictly satisfied by f. Notice that, in the holonomic case, it is reasonable to expect $\beta_i(x) \equiv c_i$ for every $i \in \mathbb{N}_m$, which expresses a *global belief* on the holonomic constraints. It might be the case that $c_i = c$ for every $i \in \mathbb{N}_m$, when one has no reason to express different beliefs on different constraints. In other cases, the choice of the belief is not obvious, since it may actually involve local properties of the constraints.

To get an insight into the joint role of the probability density and the belief of the constraints, we provide the following example.

13. E.g., expressed in terms of Dirac's deltas.

Example 1 *Let us consider the following holonomic constraints, along with their beliefs:*

$$\forall x \in \mathcal{X}: \;\; \phi_1(f_1(x), f_2(x)) := f_1(x)(1 - f_2(x)) = 0; \;\; \beta_1(x) = \frac{1}{2};$$
$$\phi_2(f_1(x), f_2(x)) = f_1(x) - y_1 = 0; \;\; \beta_2(x) = \frac{1}{4}\delta(x - \overline{x});$$
$$\phi_3(f_1(x), f_2(x)) = f_2(x) - y_2 = 0; \;\; \beta_3(x) = \frac{1}{4}\delta(x - \overline{x}).$$

The overall 2-nd degree of mismatch is

$$\frac{1}{4}\underbrace{\left((y_1 - f_1(\overline{x}))^2 + (y_2 - f_2(\overline{x}))^2\right)}_{supervised\ pairs,\ \beta_i(x)=\frac{1}{4}\delta(x-\overline{x})} + \frac{1}{2}\underbrace{\int_{\mathcal{X}} \left(f_1(x)(1 - f_2(x))\right)^2 dx}_{logic\ constraint,\ \beta_1(x)=\frac{1}{2}}.$$

While the first part involves supervised pairs on the same $\overline{x}$, the second models a logic-type constraint. Clearly, their soft fulfillment requires to express their belief, since it may be qualitatively different on different points of the domain. Basically, the belief can be thought of a weight to judge the subsequent constraint verification.

In the rest of the paper we shall consider the case $q = 1$. So, to simplify the notation we shall merely write $\mu_{\phi_i}(f)$ instead of $\mu^{(1)}_{\phi_i}(f)$ and similar notations.

Now, we consider the following functional.

$$\mathcal{E}^{\text{soft}}_{\mathcal{C}}(f) := \frac{1}{2} \parallel f \parallel^2_{P,\gamma} + \mu_{\mathcal{C}}(f). \tag{17}$$

By Definitions 1 and 4, the functional (17) is well-defined when $f \in \mathcal{F}$. In equation (17), we can clearly realize the twofold role of probability in the fuzziness arising from soft constraints. For instance, in the case of a soft holonomic constraint represented by the function $\breve{\phi}$, the degree of mismatch $\mu_{\breve{\phi}}(f)$ depends on each point $x \in \mathcal{X}$ via its belief and the probability density $p(\cdot)$ and, at the same time, the penalty $(-\breve{\phi})_+$ is a way of weighing the fuzziness of the set $\mathcal{F}_{\breve{\phi}} := \left\{\breve{\phi}: \; \breve{\phi}(x, f(x)) \geq 0 \;\; \forall x \in \mathcal{X}\right\} \subseteq \mathcal{F}$ that has already been defined in the hard case. In the case of a soft constraint, the penalty term $(-\breve{\phi})_+$ yields a measure connected with a certain membership functional of $\mathcal{F}_{\breve{\phi}}$.

Definition 12 [LEARNING FROM SOFT CONSTRAINTS] *Let $\mathcal{C}$ be a collection of constraints. The problem of determining a (local or global) minimizer of the functional $\mathcal{E}^{\text{soft}}_{\mathcal{C}}(\cdot) := \frac{1}{2} \parallel \cdot \parallel^2_{P,\gamma} + \mu_{\mathcal{C}}(\cdot)$ over $\mathcal{F}$ is called* learning from the soft constraint collection *$\mathcal{C}$.*

In the following, in order not to burden the notations and without loss of generality, we shall omit β_i and $\beta_i(x)$ and write merely 1 instead of β_i and $p(x)$ instead of $\beta_i(x)\, p(x)$.

2.4 Existence and uniqueness of optimal solutions

Let us investigate the existence of optimal solutions to the problems of learning from hard and soft constraints. The following theorem provides sufficient conditions for the existence of global minimizers in the hard case.

Theorem 13 [EXISTENCE AND UNIQUENESS FOR THE PROBLEM OF LEARNING FROM HARD CONSTRAINTS] *Let $\mathcal{C}$ be a hard constraint collection. If $\parallel \cdot \parallel_{P,\gamma}$ is a Hilbert-space norm on $\mathcal{F}$ and the set $\mathcal{F}_{\mathcal{C}}$ is nonempty, closed, and convex, then there exists a unique solution to the problem of learning from $\mathcal{C}$.*

Proof. As $\parallel \cdot \parallel_{P,\gamma}$ is a Hilbert-space norm on $\mathcal{F}$ and the norm is convex and continuous, it follows by Lemma 32 (iv) in the Appendix that the functional $\parallel \cdot \parallel_{P,\gamma}$ is weakly lower semicontinuous on $\mathcal{F}$. Moreover, since for any sufficiently large $M \in \mathbb{R}$ the set

$$S_M := \left\{ f \in \mathcal{F}_{\mathcal{C}} \;\middle|\; \parallel f \parallel_{P,\gamma} \leq M \right\}$$

is nonempty, closed, bounded and convex, by Lemma 32 (iii) S_M is nonempty and weakly compact. Finally, since any Hilbert space norm is strictly convex, by Lemma 32 (v) $\arg\min_{f \in \mathcal{F}_{\mathcal{C}}} \parallel f \parallel_{P,\gamma}$ is nonempty and contains only one element. ■

For $n = 1$, examples of problems for which $\| \cdot \|_{P,\gamma}$ is a Hilbert-space norm on $\mathcal{F}$ are provided in (Gnecco et al. (2013b)) for the class of rotationally-symmetric linear differential operators as defined in that paper (see also Section 4.3.1 for some details on such operators). 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 $\mathcal{F}_{\mathcal{C}}$ is nonempty, closed, and convex arise when the constraints are bilateral and holonomic and the constraint functions $\phi_i(\cdot, \cdot)$ are linear with respect to the second vector-valued argument. For unilateral and holonomic constraints, the assumption of closedness and convexity of $\mathcal{F}_{\mathcal{C}}$ holds when the associated constraint functions $\check{\phi}_i(\cdot, \cdot)$ are concave with respect to the second vector-valued argument and continuous. A similar remark holds for the other kinds of constraints.

The next theorem addresses the case of soft constraints.

Theorem 14 [EXISTENCE AND UNIQUENESS FOR THE PROBLEM OF LEARNING FROM SOFT CONSTRAINTS] *Let $\mathcal{C}$ be a soft constraint collection. If $\| \cdot \|_{P,\gamma}$ is a Hilbert-space norm on $\mathcal{F}$ and the penalty term $\mu_{\mathcal{C}}(\cdot)$ is convex and continuous, then there exists a unique solution to the problem of learning from $\mathcal{C}$.*

Proof. Likewise in the proof of Theorem 14, by Lemma 32 (iv) in the Appendix the functionals $\| \cdot \|_{P,\gamma}$ and $\mu_{\mathcal{C}}(\cdot)$ are weakly lower semicontinuous on $\mathcal{F}$, such is their sum. Moreover, as for any sufficiently large $M \in \mathbb{R}$ the set

$$S_M := \left\{ f \in \mathcal{F} \;\middle|\; \frac{1}{2} \| f \|^2_{P,\gamma} + \mu_{\mathcal{C}}(f) \leq M \right\}$$

is nonempty, closed, bounded and convex, it follows by Lemma 32 (iv) that S_M is nonempty and weakly compact. Finally, since any Hilbert space norm is strictly convex and the sum of a strictly convex functional and a convex one is strictly convex, by Lemma 32 (v), $\arg\min_{f \in \mathcal{F}} \left(\frac{1}{2} \| f \|^2_{P,\gamma} + \mu_{\mathcal{C}}(f) \right)$ is nonempty and contains only one element. ■

When the constraints are holonomic and bilateral, problems for which $\mu_{\mathcal{C}}(\cdot)$ is convex and continuous arise when each $\phi_i(\cdot, \cdot)$ is non-negative, continuous, and convex with respect to the second vector-valued argument (this implies the convexity of $\phi_i^2(\cdot, \cdot)$ with respect to the same argument, too). For unilateral holonomic constraints, the convexity and continuity of $\mu_{\mathcal{C}}(\cdot)$ hold when the constraint functions $\check{\phi}_i(\cdot, \cdot)$ are concave with respect to the second vector-valued argument and continuous. Again, a similar remark holds for the other kinds of constraints.

Remark 15 *In Sections 4 and 5, we shall see that for many important learning tasks the conditions of Theorem 14 (e.g., the convexity and continuity ones) hold true. However, when departing from the convexity hypothesis of the constraints, the optimization problem can become hard to face.*

3. Representer theorems and Support Constraint Machines (SCMs)

The investigations of this section are based on results established in the areas of unconstrained and constrained variational calculus (see e.g., (van Brunt (2003)), (Giaquinta and Hildebrand (1996), Chapter 2), (Ernestovic (1970), Chapter IX), (Gelfand and Fomin (1963), pp. 42–50)).

3.1 Representer theorems for hard constraints

We start considering hard holonomic constraints. The following theorem prescribes the functional representation of an optimal solution to the corresponding learning problem. Given a set of m holonomic constraints (defined, in general, on possibly different open subsets $\mathcal{X}_i$), we denote by $m(x)$ the number of constraints that are defined in the same point x of the domain. For each set $\mathcal{X}_i$, we denote by $\mathrm{cl}(\mathcal{X}_i)$ its closure in the Euclidean topology. 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., $v_j = (u^{(1)} * u^{(2)})_j = u^{(1)}_j * u^{(2)}_j$, for each index j. We denote by $\hat{\mathcal{X}} \subset \mathcal{X}$ an open set in which the same subset of constraints is defined in all its points, in such a way that $m(x)$ is constant on the same $\hat{\mathcal{X}}$. We recall that we denote by f^o a locally-optimal solution to the problem of learning from (hard or soft) constraints, and by $f^\star$ a globally-optimal solution. The following two definitions recall some classical concepts.

Definition 16 [ACTIVE CONSTRAINTS] *A constraint $\check{\phi}_i(x, f(x)) \geq 0$ is said to be* active *in x_0 at local optimality iff $\check{\phi}_i(x_0, f^o(x_0)) = 0$, otherwise is* inactive *in x_0 at local optimality.*

Definition 17 [FREE-SPACE GREEN FUNCTION] *The free-space Green function g associated to a linear differential operator O is a solution to the distributional differential equation $Og = \delta$, where δ is the Dirac distribution, centered on the origin.*

Theorem 18 [REPRESENTER THEOREM FOR HARD HOLONOMIC CONSTRAINTS] *Let us consider the learning problem formulated in Definition 9 in the case of $m < n$ hard bilateral constraints of holonomic type, which define the subset*

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

of the function space $\mathcal{F}$, where $\forall i \in \mathbb{N}_m : \phi_i \in \mathcal{C}^{k+1}(\mathrm{cl}(\mathcal{X}_i) \times \mathbb{R}^n)$. Let f^o be any constrained local minimizer of class $\mathcal{C}^{2k}(\mathcal{X}, \mathbb{R}^n)$ of the functional (11). Let us assume that for every $\hat{\mathcal{X}}$ and every x_0 in the same set $\hat{\mathcal{X}}$ one 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, respectively, 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 $\lambda_i : \hat{\mathcal{X}} \to \mathbb{R}$, $i \in \mathbb{N}_m$, such that f^o satisfies on $\hat{\mathcal{X}}$ the Euler-Lagrange equations

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

where $\gamma L := [\gamma_1 L, \ldots, \gamma_n L]'$ is a spatial-invariant operator[14] 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 $\mathcal{X}_i = \mathcal{X} = \mathbb{R}^d$, L is invertible on $\mathcal{W}^{k,2}(\mathcal{X})$, and there exists[15] a free-space Green function g of L that belongs to $\mathcal{W}^{k,2}(\mathcal{X})$, then f^o has the representation

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

where $\omega_i(\cdot) := -\lambda_i(\cdot) 1_{\mathcal{X}_i}(\cdot) \nabla_f \phi_i(\cdot, f^o(\cdot))$.

(iii) For the case of $m < n$ unilateral constraints of holonomic type, which define the subset

$$\mathcal{F}_{\check{\phi}} := \left\{f \in \mathcal{F} : \forall i \in \mathbb{N}_m, \forall x \in \mathcal{X}_i \subseteq \mathcal{X} : \check{\phi}_i(x, f(x)) \geq 0\right\}$$

of the function space $\mathcal{F}$, (i) and (ii) still hold if the nonsingularity of (18) is required when restricting the constraints defined in x_0 to the ones that are active in x_0 at local optimality (of course, replacing the ϕ_i's by the $\check{\phi}_i$'s). Moreover, each Lagrange multiplier function $\lambda_i(x)$ is non-positive and equal to 0 when the correspondent constraint is inactive in x at local optimality.

Proof. (i) Let f^o be a constrained local minimizer of the functional $\mathcal{E}(f) := \| f \|^2_{P,\gamma}$ over $\mathcal{F}$. Fix $x_0 \in \hat{\mathcal{X}}$ and a compact subset $\mathcal{X}_C \subset \hat{\mathcal{X}}$ contained in an open ball of sufficiently small radius containing x_0, and, after performing the permutations σ_ϕ and σ_f, reorder the constraints and the components of f in such a way that the ones with indexes $\sigma_\phi(1), \ldots, \sigma_\phi(m(x_0))$ and $\sigma_f(1), \ldots, \sigma_f(m(x_0))$, respectively, are the first $m(x_0)$ ones. By an application of Lemma 33, if we fix arbitrarily the functions $\eta_i \in \mathcal{C}^k_0(\mathcal{X}_C)$ 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:

$$f_1 = f_1^o + \varepsilon\eta_1 + \mathcal{O}(\varepsilon^2)\,,$$
$$f_2 = f_2^o + \varepsilon\eta_2 + \mathcal{O}(\varepsilon^2)\,,$$
$$\ldots$$

14. Here we use an overloaded notation, as made for the operator P.
15. Existence and uniqueness of g and invertibility of L are discussed in (Gnecco et al. (2013b)).

$$
\begin{aligned}
f_{m(x_0)} &= f^o_{m(x_0)} + \varepsilon\eta_{m(x_0)} + \mathcal{O}(\varepsilon^2)\,,\\
f_{m(x_0)+1} &= f^o_{m(x_0)+1} + \varepsilon\eta_{m(x_0)+1}\,,\\
f_{m(x_0)+2} &= f^o_{m(x_0)+2} + \varepsilon\eta_{m(x_0)+2}\,,\\
&\dots\\
f_n &= f^o_n + \varepsilon\eta_n\,,
\end{aligned}
$$

where the functions $\eta_i \in \mathcal{C}^k_0(\mathcal{X}_C)$, for $i = 1,\dots,m(x_0)$, are still determined by Lemma 33. In particular, by setting $y(x) := [f^o_{m(x_0)+1}(x), f^o_{m(x_0)+2}(x),\dots,f^o_n(x)]'$, $z(x) := [f^o_1(x),\dots,f^o_{m(x_0)}(x)]'$, $\phi := [\phi_1,\dots,\phi_{m(x_0)}]'$, $\eta_y := [\eta_{m(x_0)+1},\eta_{m(x_0)+2},\dots,\eta_n]'$, and $\eta_z = [\eta_1,\dots,\eta_{m(x_0)}]'$, we have[16]

$$
\eta_z(x) = -(\nabla_3\phi(x,y(x),z(x)))^{-1}(\nabla_2\phi(x,y(x),z(x)))\eta_y(x)\,. \tag{21}
$$

Moreover, the partial derivatives, up to the order k, of the first $m(x_0)$ components of f, have similar expressions, obtained from (71) in the Appendix, which contain terms of order $\mathcal{O}(\varepsilon^2)$. This implies that $\mathcal{E}(f)$ can be written as

$$
\begin{aligned}
\mathcal{E}(f) &= \sum_{j=1}^n \gamma_j\langle P(f^o+\varepsilon\eta)_j, P(f^o+\varepsilon\eta)_j\rangle + \mathcal{O}(\varepsilon^2)\\
&= \sum_{j=1}^n \gamma_j\langle Pf^o_j, Pf^o_j\rangle + 2\varepsilon\sum_{j=1}^n \gamma_j\langle Pf^o_j, P\eta_j\rangle + \varepsilon^2\sum_{j=1}^n \gamma_j\langle P\eta_j, P\eta_j\rangle + \mathcal{O}(\varepsilon^2)\\
&= \sum_{j=1}^n \gamma_j\langle Pf^o_j, Pf^o_j\rangle + 2\varepsilon\sum_{j=1}^n \gamma_j\langle Pf^o_j, P\eta_j\rangle + \mathcal{O}(\varepsilon^2)\,.
\end{aligned}
$$

Moreover, by an application of Green formula (see, e.g., (Attouch et al. (2006), Proposition 5.6.2)), we have

$$
\langle Pf^o_j, P\eta_j\rangle = \langle (P^\star)'Pf^o_j, \eta_j\rangle = \langle Lf^o_j, \eta_j\rangle\,,
$$

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

$$
\lambda(x) := -[\gamma_1(Lf^o)_1(x),\dots,\gamma_{m(x_0)}(Lf^o)_{m(x_0)}(x)](\nabla_3\phi(x,y(x),z(x)))^{-1}\,. \tag{22}
$$

Then, with such a definition, and exploiting equation (21), we get

$$
\sum_{j=1}^{m(x_0)} \gamma_j\langle Pf^o_j, P\eta_j\rangle = \sum_{j=1}^{m(x_0)} \gamma_j\langle Lf^o_j, \eta_j\rangle = \int_{\mathcal{X}} \lambda(x)(\nabla_2\phi(x,y(x),z(x)))\eta_y(x)dx\,.
$$

Summing up, we have

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

Now, since $\mathcal{E}(f)-\mathcal{E}(f^o) \geq 0$ for $|\varepsilon| > 0$ sufficiently small due to the local optimality of f^o, and $\eta_y \in \mathcal{C}^k_0(\mathcal{X}_C, \mathbb{R}^{n-m(x_0)})$ is arbitrary, by applying the Fundamental Lemma of the Calculus of Variations (see, e.g., (Giaquinta and Hildebrand (1996), Section 2.2)) we conclude that

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

16. 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.

on $\mathcal{X}_C$. This, together with the definition (22) of $\lambda(x)$, shows that (19) holds on $\mathcal{X}_C$ (setting also $\lambda_i(x) = 0$ for the constraints that are not defined in x). Finally, by varying the point x_0, we obtain (19) on the whole $\hat{\mathcal{X}}$.

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

(iii) For the case of unilateral constraints, the constraints 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 (restricting the analysis to the active constraints in x_0 at optimality and replacing the ϕ_i's by the $\check{\phi}_i$'s), since, for every sufficiently small $|\varepsilon| > 0$, the 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 non-positive. 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 $\hat{x}_0 \in \hat{\mathcal{X}}$ 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 in $\hat{x}_0$ are active in $\hat{x}_0$ at local optimality. Then, by the nonsingularity of the Jacobian matrix, there exists a vector $u = [u_1, \ldots, u_{m(\hat{x}_0)}]'$ such that $\nabla_3 \check{\phi}(\hat{x}_0, y(\hat{x}_0), z(\hat{x}_0))u = e_i$, where e_i is a column vector of all 0's, with the exception of the i-th component, which is 1. By an application of the Implicit Function Theorem (likewise in the proof of Lemma 33), for every sufficiently small $\varepsilon > 0$ (but in this case, not for every sufficiently small $\varepsilon < 0$) we can construct a feasible smooth perturbation $f(x)$ of $f^o(x)$ such that its components f_j satisfy

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

for suitable functions $\eta_1, \ldots, \eta_{m(\hat{x}_0)} \in \mathcal{C}_0^k(\mathcal{X}_C)$ such that $\eta_1(\hat{x}_0) = u_1$, $\eta_2(\hat{x}_0) = u_2$, …, $\eta_{m(\hat{x}_0)}(\hat{x}_0) = u_{m(\hat{x}_0)}$, and such that $\mathcal{E}(f) - \mathcal{E}(f^o)$, apart from an infinitesimal of order $\mathcal{O}(\varepsilon^2)$, 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, we have $\lambda_i(\hat{x}_0) \leq 0$. ■

When $\mathcal{X}_i \neq \mathbb{R}^d$, the constraints considered in Theorem 18 may include conditions on the borders $\partial\mathcal{X}_i$ which have to be taken into account when solving the correspondent Euler-Lagrange equations (see equation (19)). In particular, they are required to join the solutions to the Euler-Lagrange equations associated to two adjacent sets $\hat{\mathcal{X}}$'s.

In the expression

$$
f^o = \sum_{i=1}^{m} \gamma^{-1} g * \omega_i\,,
$$

derived for the case $\mathcal{X}_i = \mathcal{X} = \mathbb{R}^d$ for all i, we can recognize both the ingredients of a parsimonious knowledge-based solution, i.e., the Green function g and the functions ω_i, mixed by convolution. Note also that, by defining $\omega := \sum_{i=1}^m \omega_i$, we have $f^o = \gamma^{-1} g * \omega$ (the functions ω_i and $\omega := \sum_{i=1}^m \omega_i$ play a special role, which will be discussed in Section 3.4). Denoting by $\circ$ the Hadamard (entrywise) product, we can express the Fourier transform (by Fourier transform of a vector-valued function we mean the vector of Fourier transforms of each component) $\hat{f}^o$ of f^o as follows:

$$
\hat{f}^o = \gamma^{-1}\hat{g} \circ \hat{\omega}\,,
$$

which promptly shows the filtering role of $g(\cdot)$, although we should also take into account that ω itself depends on f^o through the collection of the ϕ_i's. The frequency-vector distribution $\hat{\omega}_{i,j}(\xi)$ (i.e., the j-th component of the Fourier transform $\hat{\omega}_i$ of ω_i) is referred to as the *weight* of the constraint ϕ_i in the representation of the function f_j^o. These results represent a formal statement of the general results concerning the notion of *constraint reaction* (equation (4-i)) and the representation of an optimal solution to the constrained learning problem (equations (5) and (6)).

Since the operator L is invertible on $\mathcal{W}^{k,2}(\mathcal{X})$ and has $g*$ as its inverse, by $f^o = \gamma^{-1} g * \omega$ we get $\omega = \gamma L f^o$, which is just a compact expression of the solution (20) to the Euler-Lagrange equations (19).

Corollary 19 *Under the assumptions of Theorem 18 (ii), one has* $\| f^o \|^2_{P,\gamma} = \langle \omega, \gamma^{-1} g * \omega \rangle$.

Proof. Directly from (19), when considering the definition of the function ω, we get

$$\| f^o \|^2_{P,\gamma} = \langle \gamma L f^o, f^o \rangle = \langle \omega, \gamma^{-1} g * \omega \rangle .$$

■

By the Parseval Theorem and the Mean Value Theorem, we can find $\check{\xi}_{\omega_{i,j}} \in \mathbb{R}^d$ $(i = 1, \ldots, m, j = 1, \ldots, n)$ such that (as $L = (P^*)'P$ and g solves $Lg = \delta$, $\hat{g}$ is real)

$$\begin{aligned} \| f^o \|^2_{P,\gamma} &= \langle \omega, \gamma^{-1} g * \omega \rangle = \sum_{j=1}^{n} \int_{\mathbb{R}^d} \gamma_j^{-1} \hat{g}(\xi) \cdot |\hat{\omega}|^2 (\xi) d\xi \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \int_{\mathbb{R}^d} \gamma_j^{-1} \hat{g}(\xi) \cdot |\hat{\omega}_{i,j}|^2 (\xi) d\xi = \sum_{i=1}^{m} \sum_{j=1}^{n} \gamma_j^{-1} \hat{g}(\check{\xi}_{\omega_{i,j}}) \int_{\mathbb{R}^d} |\hat{\omega}_{i,j}|^2 (\xi) d\xi. \end{aligned} \tag{23}$$

Now we consider hard isoperimetric constraints. The structure of an optimal solution of the correspondent learning problem is investigated in the following theorem. By $\mathcal{C}_0^k(\mathcal{X}, \mathbb{R}^n)$ we denote the set of functions from $\mathcal{X}$ to $\mathbb{R}^n$ that are continuously differentiable up to the order k and have compact support. We say that the constraint $\int_{\mathcal{X}} 1_{\mathcal{X}_i}(x) \cdot \check{\psi}_i(x, f(x)) dx \geq 0$ is *active* at local optimality iff $\int_{\mathcal{X}} 1_{\mathcal{X}_i}(x) \cdot \check{\psi}_i(x, f^o(x)) dx = 0$, otherwise is *inactive* at local optimality.

Theorem 20 [REPRESENTER THEOREM FOR HARD ISOPERIMETRIC CONSTRAINTS] *Let f^o be a constrained local minimizer of (11). Consider a parsimonious agent that minimizes (11) on $\mathcal{F}$, consistently with the set*

$$\forall i \in \mathbb{N}_m, \quad \int_{\mathcal{X}} 1_{\mathcal{X}_i}(x) \cdot \psi_i(x, f(x)) dx = 0 ,$$

of bilateral isoperimetric constraints, where $\forall i \in \mathbb{N}_m : \psi_i \in \mathcal{C}^1(\mathrm{cl}(\mathcal{X}_i) \times \mathbb{R}^n)$ and assume that there exist functions $\eta^{(1)}, \ldots, \eta^{(m)} \in \mathcal{C}_0^k(\mathcal{X}, \mathbb{R}^n)$ such that the matrix

$$\begin{pmatrix} \int_{\mathcal{X}_1} (\nabla_f \psi_1(x, f^o(x)))' \eta^{(1)}(x) dx & \ldots & \int_{\mathcal{X}_1} (\nabla_f \psi_1(x, f^o(x)))' \eta^{(m)}(x) dx \\ \ldots & \ldots & \ldots \\ \int_{\mathcal{X}_m} (\nabla_f \psi_m(x, f^o(x)))' \eta^{(1)}(x) dx & \ldots & \int_{\mathcal{X}_m} (\nabla_f \psi_m(x, f^o(x)))' \eta^{(m)}(x) dx \end{pmatrix} \tag{24}$$

is nonsingular. Then the following hold.

(i) There exist m constants $\lambda_i \in \mathbb{R}$ such that f^o satisfies on $\mathcal{X}$

$$\gamma L f^o(x) + \sum_{i=1}^{m} \lambda_i 1_{\mathcal{X}_i}(x) \cdot \nabla_f \psi_i(x, f^o(x)) = 0.$$

(ii) If $\mathcal{X} = \mathbb{R}^d$, L is invertible on $\mathcal{W}^{k,2}(\mathcal{X})$, and there exists a free-space Green function g of L that belongs to $\mathcal{W}^{k,2}(\mathcal{X})$, then f^o has the representation

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) * \omega_i(\cdot) , \tag{25}$$

where $\omega_i(\cdot) := -\lambda_i 1_{\mathcal{X}_i}(\cdot) \nabla_f \psi_i(\cdot, f^o(\cdot))$.

(iii) Now, consider a parsimonious agent that minimizes (11) on $\mathcal{F}$ consistently with the set

$$\forall i \in \mathbb{N}_m, \quad \int_{\mathcal{X}} 1_{\mathcal{X}_i}(x) \cdot \check{\psi}_i(x, f(x))dx \geq 0$$

of m unilateral isoperimetric constraints, where $\forall i \in \mathbb{N}_m : \check{\psi}_i \in \mathcal{C}^1(\mathrm{cl}(\mathcal{X}_i) \times \mathbb{R}^n)$. Then (i) and (ii) still hold the nonsingularity of (24) is required on the active constraints at local optimality (of course, replacing the ψ_i's by the $\check{\psi}_i$'s). Moreover, each Lagrange multiplier λ_i is non-positive and equal to 0 when the correspondent constraint is inactive at local optimality.

Proof. (i) Let f^o be a constrained local minimizer of the functional $\mathcal{E}(f) := \| f \|^2_{P,\gamma}$ over $\mathcal{F}$, let the auxiliary functions $\eta^{(1)}, \ldots, \eta^{(m)} \in \mathcal{C}^k_0(\mathcal{X}, \mathbb{R}^n)$ be such that the matrix (24) is nonsingular, and fix $\eta^{(m+1)} \in \mathcal{C}^k_0(\mathcal{X}, \mathbb{R}^n)$ arbitrarily. Let $\varepsilon_1, \ldots, \varepsilon_{m+1} \in \mathbb{R}$ and consider the problem of minimizing the function

$$F(\varepsilon_1, \ldots, \varepsilon_{m+1}) := \mathcal{E}\left(f^o + \sum_{i=1}^{m+1} \varepsilon_i \eta^{(i)}\right)$$

subject to the m equality constraints given by

$$\forall i \in \mathbb{N}_m, \;\; W_i(\varepsilon_1, \ldots, \varepsilon_{m+1}) := \int_{\mathcal{X}} 1_{\mathcal{X}_i}(x) \cdot \psi_i(x, (f^o + \sum_{i=1}^{m+1} \varepsilon_i \eta^{(i)})(x))dx = 0\,. \tag{26}$$

Of course, being f^o a local constrained minimizer of $\mathcal{E}(f)$ implies that $(0, \ldots, 0)$ is a constrained local minimizer of $F(\varepsilon_1, \ldots, \varepsilon_{m+1})$ under the constraints (26). Since the nonsingularity of (24) provides the qualification[17] of the set of constraints (26) in $(0, \ldots, 0)$, we can apply the standard theory of Lagrange multipliers in finite dimensional spaces (see, e.g., Chapter 3 in (Bertsekas (1999)) to conclude that there exists a vector $\lambda = [\lambda_1, \ldots, \lambda_m]' \in \mathbb{R}^m$ of Lagrange multipliers such that

$$\nabla_{\varepsilon_1,\ldots,\varepsilon_m} F(0, \ldots, 0) + (\nabla_{\varepsilon_1,\ldots,\varepsilon_m} W)'(0, \ldots, 0)\lambda = 0\,, \tag{27}$$

$$\nabla_{\varepsilon_{m+1}} F(0, \ldots, 0) + (\nabla_{\varepsilon_{m+1}} W)'(0, \ldots, 0)\lambda = 0\,, \tag{28}$$

where W is the column vector of components W_i, for $i = 1, \ldots, m$. Both terms $\nabla_{\varepsilon_1,\ldots,\varepsilon_m} F(0, \ldots, 0)$ and $(\nabla_{\varepsilon_1,\ldots,\varepsilon_m} G)'(0, \ldots, 0)$ do not depend on the arbitrary function $\eta^{(m+1)}$. Moreover, $\nabla_{\varepsilon_1,\ldots,\varepsilon_m} W(0, \ldots, 0)$ is equal to the matrix (24), which is invertible by assumption. Concluding, equation (27) allows us to compute the vector λ, regardless of the specific choice of $\eta^{(m+1)}$. Finally, redefining the Lagrange multipliers up to the common multiplicative constant $1/2$, equation (28) is equivalent to

$$\int_{\mathcal{X}} \left[\gamma L f^o(x) + \sum_{i=1}^{m} \lambda_i 1_{\mathcal{X}_i}(x) \cdot \nabla_f \psi_i(x, f^o(x))\right]' \eta^{(m+1)}(x)dx = 0. \tag{29}$$

from which we obtain (29), by applying the Fundamental Lemma of the Calculus of Variations.

(ii) is obtained likewise in the proof of Theorem 18 (ii).

(iii) Proceeding likewise in the proof of part (i), we consider the problem of minimizing the function

$$F(\varepsilon_1, \ldots, \varepsilon_{m+1}) := \mathcal{E}\left(f^o + \sum_{i=1}^{m+1} \varepsilon_i \eta^{(i)}\right)$$

subject to the m inequality constraints given by

$$\forall i \in \mathbb{N}_m, \;\; \check{W}_i(\varepsilon_1, \ldots, \varepsilon_{m+1}) := \int_{\mathcal{X}} 1_{\mathcal{X}_i}(x) \cdot \check{\psi}_i(x, (f^o + \sum_{i=1}^{m+1} \varepsilon_i \eta^{(i)})(x))dx \geq 0\,.$$

17. In particular, the so-called *linear independence constraint qualification* (*LICQ*).

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

The next theorem states a similar result for the case of hard pointwise constraints. We denote by $x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,|\mathcal{X}_i|)}$ the elements of each finite set $\mathcal{X}_i$ (notice that in general there may be a nonempty intersection between different $\mathcal{X}_i$'s). We say that the constraint $\check{\phi}_i(x_{(i,j)}, f(x_{(i,j)})) \geq 0$ is *active* at local optimality iff $\check{\phi}_i(x_{(i,j)}, f^o(x_{(i,j)})) = 0$, otherwise is *inactive* at local optimality.

Theorem 21 [REPRESENTER THEOREM FOR HARD POINTWISE CONSTRAINTS] *Let us consider a parsimonious agent that minimizes (11) on $\mathcal{F}$ consistently with a given set of bilateral pointwise constraints given by*

$$\forall i \in \mathbb{N}_m \forall j \in \mathbb{N}_{|\mathcal{X}_i|}, \;\; \phi_i(x_{(i,j)}, f(x_{(i,j)})) = 0\,,$$

where $\forall i \in \mathbb{N}_m : \phi_i \in \mathcal{C}^1(\mathcal{X} \times \mathbb{R}^n)$. Let f^o be any constrained local minimizer of (11). Moreover, assume that for any x_0 in the finite set $\cup_{i=1}^m \mathcal{X}_i$ we have $m(x_0) \leq n$ and 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

$$\frac{\partial(\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))})}{\partial(f^o_{\sigma_f(1)}, \ldots, f^o_{\sigma_f(m(x_0))})}\,, \tag{30}$$

evaluated in x_0, is nonsingular. Then the following hold.

(i) There exist $\sum_{i=1}^m \sum_{j=1}^{|\mathcal{X}_i|}$ constants $\lambda_{(i,j)} \in \mathbb{R}$ such that f^o satisfies on $\mathcal{X}$

$$\gamma L f^o(x) + \sum_{i=1}^{m} \sum_{j=1}^{|\mathcal{X}_i|} \lambda_{(i,j)} \delta(x - x_{(i,j)}) \nabla_f \phi_i(x, f^o(x)) = 0. \tag{31}$$

(ii) If $\mathcal{X} = \mathbb{R}^d$, L is invertible on $\mathcal{W}^{k,2}(\mathcal{X})$, and there exists a free-space Green function g of L that belongs to $\mathcal{W}^{k,2}(\mathcal{X})$, then f^o has the representation

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) * \omega_i(\cdot)\,, \tag{32}$$

where $\omega_i(\cdot) := -\sum_{j=1}^{|\mathcal{X}_i|} \lambda_{(i,j)} \delta(\cdot - x_{(i,j)}) \nabla_f \phi_i(\cdot, f^o(\cdot))$.

(iii) For the case of m unilateral pointwise constraints given by

$$\forall i \in \mathbb{N}_m \forall j \in \mathbb{N}_{|\mathcal{X}_i|}, \;\; \check{\phi}_i(x_{(i,j)}, f(x_{(i,j)})) \geq 0\,,$$

where $\forall i \in \mathbb{N}_m : \check{\phi}_i \in \mathcal{C}^1(\mathcal{X} \times \mathbb{R}^n)$, (i) and (ii) still hold if the nonsingularity of (30) is required on the active constraints at local optimality (of course, replacing the ϕ_i's by the $\check{\phi}_i$'s). Moreover, each Lagrange multiplier $\lambda_{(i,j)}$ is non-positive and equal to 0 when the correspondent constraint is inactive at local optimality.

Proof. Items (i), (ii) and (iii) are obtained likewise the correspondent items in Theorem 20. In the following, we detail only the required changes with respect to the proof of item (i) of Theorem 20.

Let us denote by M the cardinality of the set $\cup_{i=1}^m \mathcal{X}_i$ and by $x_0^{(1)}, \ldots, x_0^{(M)}$ its elements. For $l = 1, \ldots, M$, let the auxiliary functions $\eta^{(l,1)}, \ldots, \eta^{(l,m(x_0^{(l)}))} \in \mathcal{C}_0^k(\mathcal{X}, \mathbb{R}^n)$ be chosen according to the following rules:

- the supports of $\eta^{(l,1)}, \ldots, \eta^{(l,m(x_0^{(l)}))}$ are contained in an open ball of sufficiently small radius, centered in $x_0^{(l)}$;
- the auxiliary functions correspondent to different indexes l (hence to different points $x_0^{(l)}$) have disjoint supports;

- the $m(x_0^{(l)}) \times m(x_0^{(l)})$ matrix

$$\frac{\partial(\phi_1, \ldots, \phi_{m(x_0^{(l)})})}{\partial(f_1^o, \ldots, f_n^o)}[\eta^{(l,1)}, \ldots, \eta^{(l,m(x_0^{(l)}))}],$$

evaluated in $x_0^{(l)}$, is nonsingular[18], where $[\eta^{(l,1)}, \ldots, \eta^{(l,m(x_0^{(l)}))}]$ denotes the $n \times m(x_0^{(l)})$ matrix obtained by concatenating the column vectors $\eta^{(l,1)}, \ldots, \eta^{(l,m(x_0^{(l)}))}$.

Now, let us choose an arbitrary function $\eta \in \mathcal{C}_0^k(\mathcal{X}, \mathbb{R}^n)$ and, for $\varepsilon, \varepsilon_{(l,r)} \in \mathbb{R}$ ($l = 1, \ldots, M, r = 1, \ldots, m(x_0^{(l)})$), consider the problem of minimizing the function

$$F(\varepsilon, \{\varepsilon_{(l,r)}\}) := \mathcal{E}\left(f^o + \varepsilon\eta + \sum_{l=1}^{M}\sum_{r=1}^{m(x_0^{(l)})} \varepsilon_{(l,r)}\eta^{(l,r)}\right)$$

subject to the $\sum_{h=1}^{M} m(x_0^{(h)})$ equality constraints[19] given by

$$\begin{aligned}
&\forall h \in \mathbb{N}_M \forall i \in \mathbb{N}_{m(x_0^{(h)})},\\
&G_{h,i}(\varepsilon, \{\varepsilon_{(l,r)}\}) := \phi_i\left(x_0^{(h)}, \left(f^o + \varepsilon\eta + \sum_{l=1}^{M}\sum_{r=1}^{m(x_0^{(l)})} \varepsilon_{(l,r)}\eta^{(l,r)}\right)(x_0^{(h)})\right) = 0\,.
\end{aligned}$$

By the construction of the auxiliary functions $\eta^{(l,r)}$, the qualification of the constraints holds and the Lagrange multipliers can be chosen independently of the function η. So, we can proceed likewise in the remaining of the proof of item (i) of Theorem 20. The Dirac delta terms in (31) arise from the fact that $(\nabla_\varepsilon G_{h,i})(0, \ldots, 0)$ depends only the value assumed by the vector $\eta(x)$ for $x = x_0^{(h)}$, $h = 1, \ldots, M$. ■

Example 2 *Let us consider the classical supervised learning in which we want to enforce hard fulfillment of the constraints*

$$\forall \kappa \in \mathbb{N}_{l_s} : \quad f(x_\kappa) - y_\kappa = 0\,,$$

where the subscript "s" in l_s stands for "supervised". We can promptly see that conditions (30) concerning the qualification of the constraints holds true whenever we deal with distinct points, since the Jacobian matrix becomes the diagonal matrix.

3.2 Representer theorems for soft constraints

In this section we do not consider the case of soft isoperimetric constraints, which, however, can be dealt with similarly to holonomic and pointwise ones.

We can associate any holonomic or pointwise unilateral constraint $\check{\phi}_i(x, f(x)) \geq 0$ with $\phi_i^{\geq}(x, f(x)) = 0$, where $\phi_i^{\geq}(\cdot, \cdot)$ is a suitable non-negative function. Similarly, each holonomic or pointwise bilateral constraint can be reformulated as a pair of unilateral constraints. Hence, the learning problem amounts at minimizing

$$\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f) = \frac{1}{2} \parallel f \parallel_{P,\gamma}^2 + \sum_{i=1}^{m} \mu_{i,\mathcal{C}}(f) = \frac{1}{2} \parallel f \parallel_{P,\gamma}^2 + \sum_{i=1}^{m} \int_{\mathcal{X}} 1_{\mathcal{X}_i}(x)\phi_i^{\geq}(x, f(x))\, p(x)\, dx\,, \tag{33}$$

where each set $\mathcal{X}_i$ is either open or made up of a finite number of points.

The next result is a representer theorem for an optimal solution. Note that the classical supervised learning is a degenerate case in which each set $\mathcal{X}_i$ is made up of a single point, and in this case we set $p(x)\, 1_{\mathcal{X}_i}(x) = p(x)\delta(x - x_i)$.

Theorem 22 (REPRESENTER THEOREM FOR SOFT HOLONOMIC AND SOFT POINTWISE CONSTRAINTS). *Let $\mathcal{C}$ be a collection of soft constraints, p be continuous, non-negative and in $\mathcal{L}^1(\mathcal{X})$, and consider the problem of minimizing over $\mathcal{F}$ the functional $\mathcal{E}_{\mathcal{C}}^{\text{soft}}$ (see (33)). Let f^o be a local minimizer, then the following holds true.*

18. Choosing the functions $\eta^{(l,1)}, \ldots, \eta^{(l,m(x_0^{(l)}))}$ in such a way is always possible, due to the assumed nonsingularity of the Jacobian matrix (30).

19. We have implicitly reordered the constraints in such a way that the first $m(x_0^{(h)})$ ones are those defined in $x_0^{(h)}$.

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

$$\gamma L f^o(x) + \sum_{i=1}^{m} p(x) 1_{\mathcal{X}_i}(x) \cdot \nabla_f \phi_i^{\geq}(x, f^o(x)) = 0. \tag{34}$$

Under the same assumptions on $\phi_i^{\geq}$, the same result holds if the sets $\mathcal{X}_i$ are made up of a finite number of points.

(ii) Suppose that the sets $\mathcal{X}_i$ are disjoint, each set $\mathcal{X}_i$ 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 $\mathcal{X}$

$$\gamma_j L f_j^o(x) + \sum_{i=1}^{m} p(x) 1_{\mathcal{X}_i}(x) \cdot \overline{\partial_{f_j}} \phi_{i,j}^{\geq}(x, f_j^o(x)) = 0 \,,\ j = 1, \ldots, n \,, \tag{35}$$

where $\overline{\partial_{f_j}} \phi_{i,j}^{\geq}(x, f_j^o(x))$ denotes a suitable element of the subdifferential[20] $\partial_{f_j} \phi_{i,j}^{\geq}(x, f_j^o(x))$, which is equal to the subdifferential $\partial_{f_j} \phi_i^{\geq}(x, f^o(x))$.

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

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) * \omega_i^{\geq}(\cdot) \,, \tag{36}$$

where $\omega_i^{\geq}(\cdot)$, under the assumptions of item (i), has the expression

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

whereas, under the assumptions of item (ii), its n components $\omega_{i,j}^{\geq}(\cdot)$ are given by

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

Proof. (i) Let us first consider the case in which $\mathcal{X}_i \subseteq \mathcal{X}$ is open. Then, (i) is proved by fixing arbitrarily $\eta \in \mathcal{C}_0^k(\mathcal{X}, \mathbb{R}^n)$, then computing

$$\lim_{\varepsilon \to 0} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon\eta) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} = \int_{\mathcal{X}} \left(\gamma L f^o(x) + \sum_{i=1}^{m} p(x) 1_{\mathcal{X}_i}(x) \cdot \nabla_f \phi_i^{\geq}(x, f^o(x)) \right)' \eta(x) dx = 0 \,, \tag{37}$$

and finally applying the Fundamental Lemma of the Calculus of Variations. The first equality in (37) is obtained by exploiting the assumption that $\forall x \in \mathcal{X}_i$ there is an open neighborhood $\mathcal{N}$ of $(x, f^o(x))$ for which $\phi_i^{\geq} \in \mathcal{C}^1(\mathcal{N})$, and the second is derived by the local optimality of f^o. The case in which the sets $\mathcal{X}_i$ are made up of a finite number of points is similar and is proved in (Gnecco et al. (2014)).

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

$$\lim_{\varepsilon \to 0} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon\eta) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} = \int_{\mathcal{X}} (\gamma L f^o(x))' \eta(x) dx = 0 \,. \tag{38}$$

Since, for every $j = 1, \ldots, n$, $\gamma_j L f_j^o$ is a distribution, (38) implies that the support of $\gamma_j L f_j^o$ is a subset of $\{x_1, \ldots, x_m\}$, which is a set of finite cardinality. By (Schwartz (1978), Theorem XXXV in Chapter 3), $\gamma_j L f_j^o$ is made up of a finite

20. Let $\Omega \subseteq \mathbb{R}^d$ be a convex set. 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 $f(x) - f(x_0) \geq v'(x - x_0)$.

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. This can be checked by choosing a function $\eta \in \mathcal{C}_0^\infty(\mathcal{X}, \mathbb{R}^n)$ such that only its j-th component η_j is different from 0, and $\eta_j(x) = 0$ for all $x \in \cup_{i=1}^m \mathcal{X}_i$ (although some partial derivatives of some order of η_j may be different from 0 for some $x \in \cup_{i=1}^m \mathcal{X}_i$). Concluding, $\gamma_j L f_j^o$ satisfies on $\mathcal{X}$

$$\gamma_j L f_j^o(x) = \sum_{i=1}^m B_i \delta(x - x_i)\,, \tag{39}$$

where the B_i's are constants. Notice that (39) is of the same form as (35).

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 following arguments hold also for the case $m > 1$. Let η^{j+} denote any function in $\mathcal{C}_0^k(\mathcal{X}, \mathbb{R}^n)$ 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.

1. $(1 - y_{1,j} \cdot f_j^o(x_1))_+ < 0$;
2. $(1 - y_{1,j} \cdot f_j^o(x_1))_+ > 0$;
3. $(1 - y_{1,j} \cdot f_j^o(x_1))_+ = 0$ and $y_{1,j} = -1$;
4. $(1 - y_{1,j} \cdot f_j^o(x_1))_+ = 0$ and $y_{1,j} = 1$.

- In the case (1), we get

$$\lim_{\varepsilon \to 0^+} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon \eta^{j+}) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} = \int_{\mathcal{X}} \gamma_j L f_j^o(x) \eta_j^{j+}(x) dx = B_1 \eta_j^{j+}(x_1) \geq 0 \tag{40}$$

  and

$$\lim_{\varepsilon \to 0^+} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon \eta^{j-}) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} = \int_{\mathcal{X}} \gamma L_j f_j^o(x) \eta_j^{j-}(x) dx = -B_1 \eta_j^{j+}(x_1) \geq 0\,, \tag{41}$$

  then $B_1 = 0$ (since $\eta_j^{j+}(x_1) > 0$). Notice that the inequalities in (40) and (41) follow by the local optimality of f^o, whereas the equalities by the left/right differentiability[21] of the function $(\cdot)_+$.

- Similarly, in the case (2), we have

$$\begin{aligned} \lim_{\varepsilon \to 0^+} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon \eta^{j+}) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} &= \int_{\mathcal{X}} \left(\gamma_j L f_j^o(x) - y_{1,j} p(x) 1_{\mathcal{X}_1}(x)\right) \eta_j^{j+}(x) dx \\ &= (B_1 - y_{1,j} p(x_1)) \eta_j^{j+}(x_1) \geq 0 \end{aligned}$$

  and

$$\begin{aligned} \lim_{\varepsilon \to 0^+} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon \eta^{j-}) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} &= \int_{\mathcal{X}} \left(\gamma_j L f_j^o(x) - y_{1,j} p(x) 1_{\mathcal{X}_1}(x)\right) \eta_j^{j-}(x) dx \\ &= -(B_1 - y_{1,j} p(x_1)) \eta_j^{j+}(x_1) \geq 0\,. \end{aligned}$$

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

- In the case (3), we obtain

$$\begin{aligned} \lim_{\varepsilon \to 0^+} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon \eta^{j+}) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} &= \int_{\mathcal{X}} \left(\gamma_j L f_j^o(x) - y_{1,j} p(x) 1_{\mathcal{X}_1}(x)\right) \eta_j^{j+}(x) dx \\ &= (B_1 - y_{1,j} p(x_1)) \eta_j^{j+}(x_1) \geq 0 \end{aligned}$$

  and

$$\lim_{\varepsilon \to 0^+} \frac{\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o + \varepsilon \eta^{j-}) - \mathcal{E}_{\mathcal{C}}^{\text{soft}}(f^o)}{\varepsilon} = \int_{\mathcal{X}} \gamma_j L f_j^o(x) \eta_j^{j-}(x) dx = -B_1 \eta_j^{j+}(x_1) \geq 0\,.$$

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

21. Depending on the sign of $y_{1,j}$.

- Finally, in the case (4) we get

$$\lim_{\varepsilon\to 0^+}\frac{\mathcal{E}_{\mathcal{C}}^{\rm soft}(f^o+\varepsilon\eta^{j+})-\mathcal{E}_{\mathcal{C}}^{\rm soft}(f^o)}{\varepsilon}=\int_{\mathcal{X}}\gamma_j Lf_j^o(x)\eta_j^{j+}(x)dx=B_1\eta_j^{j+}(x_1)\geq 0$$

and

$$\begin{aligned}\lim_{\varepsilon\to 0^+}\frac{\mathcal{E}_{\mathcal{C}}^{\rm soft}(f^o+\varepsilon\eta^{j-})-\mathcal{E}_{\mathcal{C}}^{\rm soft}(f^o)}{\varepsilon} &= \int_{\mathcal{X}}\left(\gamma_j Lf_j^o(x)-y_{1,j}p(x)1_{\mathcal{X}_1}(x)\right)\eta_j^{j-}(x)dx\\ &= -(B_1-y_{1,j}p(x_1))\eta_j^{j+}(x_1)\geq 0\,.\end{aligned}$$

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

Finally, summarizing the results of the analysis of cases (1)-(4) and applying the definition of subdifferentiability to the function $(\cdot)_+$, we get[22] (35).

(iii) is obtained likewise in the proof of Theorem 18 (ii). ■

The item (i) of Theorem 22 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 this case, we can approximate such a function by a continuously-differentiable approximation or, for some $\phi_i^{\geq}$, we can deal directly with the nondifferentiable case, as shown in Theorem 22 (ii).

A remarkable difference with the case of hard constraints is that the solution provided by Theorem 22 is based on the assumption of knowing the probability density of the data, whereas for hard constraints we need to compute the Lagrange multipliers associated with the constraints and also to check the satisfaction of the constraints. We also remark that the classical supervised learning with a smooth loss is a degenerate case of Theorem 22 (i), in which each set $\mathcal{X}_i$ is made up of a single element x_i, and one sets $p(x)1_{\mathcal{X}_i}(x)=p(x)\delta(x-x_i)$. Such a degenerate case is extended in Theorem 22 (ii) to the case of a particular nondifferentiable function $\phi_i^{\geq}$. Examples of applications of Theorem 22 are provided in Section 4.

When the probability density p is not known but a finite set $\mathcal{U}:=\{\tilde{x}_\kappa\in\mathbb{R}^d,\,\kappa=1,\ldots,l_u\}$ of unsupervised examples is given (where the subscript "u" in "l_u" stands for "unsupervised"), we can exploit them to estimate p according to a mixture of kernel functions, i.e., $p(x)=\sum_{\kappa\in\mathbb{N}_{l_u}}\pi_\kappa\,\vartheta_\kappa(x-\tilde{x}_\kappa)$ for suitable coefficients π_κ's and kernel functions ϑ_κ's. For example, we can take the kernel functions equal to the Green function g of the differential operator L; in such a way, the Euler-Lagrange equations (34) become

$$\gamma Lf^o(x)+\sum_{i=1}^{m}\sum_{\kappa=1}^{\ell_u}\pi_\kappa\cdot 1_{\mathcal{X}_i}(x)\cdot\nabla_f\phi_{i,\kappa}^{\geq}(x,f^o(x))=0\,,$$

where

$$\phi_{i,\kappa}^{\geq}(x,f^o(x)):=g(x-x_\kappa)\cdot\phi_i^{\geq}(x,f^o(x)).$$

If we plug the expression of $\omega_i^{\geq}(\cdot)$ into the representational equation for $f^o(\cdot)$ given by Theorem 22 (iii) we get

$$\begin{aligned}f^o(x) &= -\sum_{i=1}^{m}\gamma^{-1}g(x)*\left(p(x)1_{\mathcal{X}_i}(x)\nabla_f\phi_i^{\geq}(x,f^o(x))\right)\\ &= -\sum_{i=1}^{m}\gamma^{-1}g(x)*\left(\sum_{\kappa=1}^{\ell_u}\pi_\kappa g(x-x_\kappa)1_{\mathcal{X}_i}(\cdot)\nabla_f\phi_i^{\geq}(x,f^o(x))\right)\\ &= -\sum_{i=1}^{m}\sum_{\kappa=1}^{\ell_u}\gamma^{-1}\pi_\kappa g(x)*\left(g(x-x_\kappa)1_{\mathcal{X}_i}(x)\nabla_f\phi_i^{\geq}(x,f^o(x))\right)\,.\end{aligned}\tag{42}$$

In Section 5 we will provide algorithmic issues for the concrete solution of the learning problem, which are based on the notion of dimensionality collapse.

22. A different method to prove item (ii) in Theorem 22 consists in reducing the soft constraints to hard unilateral constraints, then applying a slight modification of Theorem 21 (iii).

3.3 Representer theorems for mixed constraints

The proof techniques used to derive the results in Sections 3.1 and 3.2 can be applied to the case in which the constraints are of mixed type, too; for instance, when there are simultaneously a bilateral isoperimetric constraint and a bilateral holonomic constraint. While the extension to such a case in the soft case is straightforward, in the hard case it requires in addition the satisfaction and the qualification of the whole set of mixed constraints, which are both required to apply the theory of Lagrange multipliers. A problem that arises when dealing with mixed hard and soft constraints, together with the associated Euler-Lagrange equations, is reported and solved in Section 4.3.2.

As an example, the following theorem combines hard holonomic and soft pointwise constraints. The learning problem amounts at minimizing, for some $c > 0$, the functional

$$\mathcal{E}_{\mathcal{C}}^{\text{mixed}}(f) := \frac{1}{2} \parallel f \parallel_{P,\gamma}^2 + \frac{c}{2l_s} \sum_{\kappa=1}^{l_s} \sum_{j=1}^{n} (y_{\kappa,j} - f_j(x_\kappa))^2 \,, \tag{43}$$

in the presence of a collection $\mathcal{C}$ of hard holonomic constraints and of l_s supervised examples $(x_\kappa \in \mathbb{R}^d, y_{\kappa,j} \in \mathbb{R})$, dealt with in a soft way, where $\kappa = 1, \ldots, l_s$ and $j = 1, \ldots, n$.

Theorem 23 [REPRESENTER THEOREM FOR MIXED HARD HOLONOMIC AND SOFT POINTWISE CONSTRAINTS] *Let us consider the minimization of the functional (43) in the case of $m < n$ hard bilateral constraints of holonomic type, which define the subset*

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

of the function space $\mathcal{F}$, where $\forall i \in \mathbb{N}_m : \phi_i \in \mathcal{C}^\infty(\text{cl}(\mathcal{X}_i) \times \mathbb{R}^n)$. Let f^o be any constrained local minimizer of the functional (43), and let the holonomic constraints be defined in such a way that either $Lf^o \in \mathcal{C}^0(\mathcal{X}, \mathbb{R}^n)$ or they are of the form $Af(x) = b(x)$, where $A \in \mathbb{R}^{m,n}$ with $m < n$ and $\text{rank}(A) = m$, and $b \in \mathcal{C}_0^{2k}(\mathcal{X}, \mathbb{R}^m)$. Let us assume that for any $\hat{\mathcal{X}}$ and for every x_0 in the same $\hat{\mathcal{X}}$ 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, respectively, 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

$$\frac{\partial(\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))})}{\partial(f^o_{\sigma_f(1)}, \ldots, f^o_{\sigma_f(m(x_0))})} \,, \tag{44}$$

evaluated in x_0, is not singular. Suppose also that (44) is of class $\mathcal{C}^\infty(\hat{\mathcal{X}}, \mathbb{R}^n)$. Then, the following hold.

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

$$\gamma L f^o + \sum_{i=1}^{m} \lambda_i 1_{\mathcal{X}_i}(\cdot) \cdot \nabla_f \phi_i(\cdot, f^o(\cdot)) + \frac{c}{l_s} \sum_{\kappa=1}^{l_s} (f^o(\cdot) - y_\kappa) \delta(\cdot - x_\kappa) = 0 \,.$$

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

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1} g(\cdot) * \omega_i^{\text{hard}}(\cdot) + \sum_{\kappa=1}^{l_s} \gamma^{-1} g(\cdot) * \omega_\kappa^{\text{soft}}(\cdot) \,, \tag{45}$$

where $\omega_i^{\text{hard}}(\cdot) := -\lambda_i 1_{\mathcal{X}_i}(\cdot) \nabla_f \phi_i(\cdot, f^o(\cdot))$, and $\omega_\kappa^{\text{soft}}(\cdot) := -\frac{c}{l_s}(f^o(x_\kappa) - y_\kappa)\delta(\cdot - x_\kappa)$.

(iii) For the case of $m < n$ unilateral constraints of holonomic type, which define the subset

$$\mathcal{F}_{\check{\phi}} := \left\{f \in \mathcal{F} : \ \forall i \in \mathbb{N}_m, \ \forall x \in \mathcal{X}_i \subseteq \mathcal{X}, \ \check{\phi}_i(x, f(x)) \geq 0\right\}$$

of the function space $\mathcal{F}$, (i) and (ii) still hold (with every occurrence of ϕ_i replaced by $\check{\phi}_i$) if the nonsingularity of the Jacobian matrix (see (44)) is required when restricting the constraints defined in x_0 to those active in x_0 at local optimality. Moreover, each Lagrange multiplier λ_i is non-positive and locally equal to 0 when the correspondent constraint is locally inactive at local optimality.

Proof. For the sake of conciseness, we merely summarize the differences with respect to the proof of Theorem 18. First, in the Euler-Lagrange equations there is an additional term $\frac{c}{l_s} \sum_{\kappa=1}^{l_s} (f^o(x) - y_\kappa)\delta(x - x_\kappa)$, due to the presence

of the supervised examples. Then, in general the Lagrange multipliers λ_i are not functions, likewise in Theorem 18, but distributions, obtained by a variation of equation (22), which is well-defined in a distributional sense since the Jacobian matrix (44) is locally invertible and infinitely smooth, and since either $Lf^o \in \mathcal{C}^0(\mathcal{X}, \mathbb{R}^n)$ or $Af(x) = b(x)$ hold (with the stated assumptions on A and $b(x)$). More precisely, equation (22) is replaced by

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

where now λ is a row vector distribution. Finally, additional smoothness of f^o is not required, since only (44) has to be infinitely smooth. ■

3.4 Constraint reactions

The following definition introduces a concept that plays a central role in the theory developed in this paper.

Definition 24 [CONSTRAINT REACTIONS] *The function ω_i in Theorems 18 and 20 and the function $\omega_i^{\geq}$ in Theorem 22 is called* reaction of the i-th constraint, *while $\omega := \sum_{i=1}^m \omega_i$ (respectively, $\omega^{\geq} := \sum_{i=1}^m \omega_i^{\geq}$) is the overall* reaction of the constraints. *A similar definition holds for the distributions ω_i and $\omega := \sum_{i=1}^m \omega_i$ in Theorem 21, the distributions $\omega_i^{\geq}$ and $\omega^{\geq} := \sum_{i=1}^m \omega_i^{\geq}$ in Theorem 22, and the distributions ω_i^{hard}, $\omega_\kappa^{\text{soft}}$ and $\omega^{\text{mixed}} := \sum_{i=1}^m \omega_i^{\text{hard}} + \sum_{\kappa=1}^{l_s} \omega_\kappa^{\text{soft}}$ in Theorem 23.*

We emphasize the fact that the constraint reaction is a concept associated with the (constrained) local minimizer f^o. In particular, two different 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 the respective representer theorems, the reaction of the i-th constraint provides the way under which such a constraint contributes to the expansion of f^o. For instance, under the assumptions of Theorem 18 (ii), we have the expansion

$$f^o = \sum_{i=1}^m \gamma^{-1} g * \omega_i = \gamma^{-1} g * \omega\,,$$

which shows the roles of ω_i and ω in the representation of f^o. Hence, in such cases the problem of learning from constraints is reduced to finding the reactions of the constraints.

The following uniqueness property that involves constraint reactions and Lagrange multipliers plays an important role for concrete development of algorithms. Indeed, although in general a local minimizer is a-priori unknown, this is still a structural property of local minimizers, which can be useful when searching for them.

Proposition 25 *Both for soft and hard constraints, under the hypothesis of the respective representer theorems, the constraint reactions and the Lagrange multipliers are uniquely determined by the local minimizer f^o.*

Proof. For the soft case, the property concerning the constraint reaction is trivial, and follows directly from the definition of the constraint reaction, as no Lagrange multiplier has to be determined. Let us prove the uniqueness of Lagrange multipliers for hard constraints. We detail the proof in the holonomic bilateral case, i.e., under the hypotheses of Theorem 18 (ii). Similar proofs can be given for the other cases. Let us proceed by contradiction and assume that there exist two different sets $\{\lambda_i,\ i = 1 \ldots, m\}$ and $\{\overline{\lambda}_i,\ i = 1 \ldots, m\}$ of Lagrange multipliers associated with the same constrained local minimizer f^o, with at least one index i such that $\lambda_i \neq \overline{\lambda}_i$. According to Theorem 18 (i), f^o satisfies the Euler-Lagrange equations (19). Without loss of generality, for each $x \in \hat{\mathcal{X}}$, we 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 $x \in \hat{\mathcal{X}}$, 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. By the assumption of distinct Lagrange multipliers, we have

$$\begin{aligned}\gamma Lf^o(x) + \sum_{i=1}^m \lambda_i(x) \cdot \nabla_f \phi_i(x, f^o(x)) = 0\,,\\ \gamma Lf^o(x) + \sum_{i=1}^m \overline{\lambda}_i(x) \cdot \nabla_f \phi_i(x, f^o(x)) = 0,\end{aligned}$$

from which we get

$$\sum_{i=1}^{m}(\lambda_i - \overline{\lambda}_i)\nabla_f \phi_i = (\lambda_{(D)} - \overline{\lambda}_{(D)})' \frac{\partial(\phi_1, \ldots, \phi_{m(x)})}{\partial(f_1, \ldots, f_{m(x)})} = 0\,,$$

where $\lambda_{(D)} := [\lambda_1, \ldots, \lambda_m(x)]'$ and $\overline{\lambda}_{(D)} := [\overline{\lambda}_1, \ldots, \overline{\lambda}_{m(x)}]'$. Distinct multipliers are only compatible with the singularity of the Jacobian matrix, which contradicts the assumption on the invertibility of (18). ■

3.5 Support constraints and Support Constraint Machines

Starting from the concept of the reaction of a constraint, we introduce the following definition.

Definition 26 [SUPPORT CONSTRAINT] *A* support constraint *is a constraint associated with a reaction that is different from* 0 *at least in one point of the domain* $\mathcal{X}$*.*

Under the assumptions of the representer theorems of Sections 3.1 - 3.3, a local optimal solution f^o to a problem of learning from hard or soft constraints can be obtained by the knowledge of the Lagrange multipliers (in the hard case) and the reactions associated merely with the support constraints. This motivates the use of the terminology "support constraints" as an extension of the concept of "support vector" in kernel methods to problems of learning from constraints. The connection with kernel methods arises also because, as already mentioned, under quite general conditions the free-space Green function g associated with the operator L is the kernel of a RKHS (see, e.g., Gnecco et al. (2013b) 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, with soft pointwise constraints expressed by the quadratic loss and in the absence of hard constraints, one obtains kernel ridge regression. In the case of soft pointwise constraints expressed by the hinge loss and no hard constraints, one gets support vector machines. In Section 4, a precise connection will be given with support vectors that are associated with the correspondent support constraints.

The emergence of constraints whose reaction is identically 0 at local optimality (thus, constraints that are not support constraints) is particularly evident for the case of hard holonomic unilateral constraints. For instance, under the assumptions of Theorem 18 (iii), a hard holonomic unilateral constraint that is inactive at local optimality for all $x \in \mathcal{X}$ is associated with a Lagrange multiplier function $\lambda_i(\cdot)$ that is identically 0, so its reaction is identically 0, too. Therefore, such a constraint is *not* a support constraint.

It is interesting to discuss the case of an instance of a problem of learning from hard constraints in which one of the constraints is entailed by the collection of the remaining ones, in the sense that the fulfillment of all the other hard constraints guarantees its fulfillment, too. Without any loss of generality, such a redundant constraint[23] can be discarded from the problem formulation and, provided that the assumptions of one of Theorems 18, 20, or 21 hold, we still have the representations (20), (25), or (32) for the constrained local minimizer f^o, where the Lagrange multiplier associated with the redundant constraint is 0, hence also the reaction from that constraint is 0. Therefore, we can say that the redundant constraint is *not* a support constraint.

The concept of support constraints arises also in the soft case. Indeed, although in such a context the Lagrange multipliers - which make their appearance in the hard case - are replaced by fixed (possibly generalized) probability densities, the reaction of the i-th constraint can be still identically 0 - hence the constraint is *not* a support constraint - provided that the term $\nabla_f \phi_i^{\geq}(x, f^o(x))$ in equation (34) or - for all indexes j - the one $\overline{\partial_{f_j}} \phi_{i,j}^{\geq}(x, f_j^o(x))$ in equation (35) - is identically 0 on $\mathcal{X}_i$. Examples of such a behavior are provided in Section 4. It is also shown therein that the classical concept of *support vector* and the one of *support set* are two instances of the more general concept of support constraint.

The role played by support constraints in the proposed learning paradigm is emphasized in the next definition.

Definition 27 [SUPPORT CONSTRAINT MACHINE (SCM)] *A* Support Constraint Machine *(SCM) is any computational machinery capable of finding a (local or global) optimal solution to the problem of learning from constraints for which one of the representer theorems provided in Sections 3.1, 3.2, and 3.3 holds, and such an optimal solution can be expressed in terms of the correspondent constraint reactions.*

23. Of course, redundant hard constraints can appear only if the assumption on the invertibility of the Jacobian matrix made in the correspondent represent theorem of Section 3.1 is violated.

3.6 Approximating the Gaussian kernel

The Gaussian is often used as a kernel function in classical kernel methods and SVMs. However, when P is a finite-order linear differential operator (see Definition 5), a Gaussian cannot be the Green function of $L = (P^\star)'P$. Indeed, if g were a Gaussian, the first hand-side of the distributional differential equation $Lg = \delta$ would be smooth (differently from the right-hand side). So, our theory does not cover directly the Gaussian kernel. 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 function. Indeed, in (Yuille and Grzywacz (1988)) it was shown that the Gaussian kernel with mean 0 and variance σ^2 is a free-space Green function of the linear differential operator of infinite order[24]

$$L = L_{(\infty)} := \sum_{i=0}^{\infty} (-1)^i a_i \nabla^{2i} \,, \tag{46}$$

where $a_i := \frac{\sigma^{2i}}{i! 2^i}$ (see also Poggio and Girosi (1989), Section 5.1.2). This can be proved via Fourier transforms, observing that the Fourier transforms of $(-1)^i \nabla^{2i}$ and δ are, respectively, the function $\|2\pi\xi\|^{2i}$ and the constant function 1. Then, with such coefficients a_i, by exploiting the Taylor-series expansion

$$\exp(t) = \sum_{i=0}^{\infty} \frac{t^i}{i!} \,,$$

one obtains

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

whose inverse Fourier transform is the Gaussian function

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

The simplest way to apply the results of the previous sections to this situation consists in replacing, for some positive integer k, the infinite-order operator (46) with its truncation

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

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

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

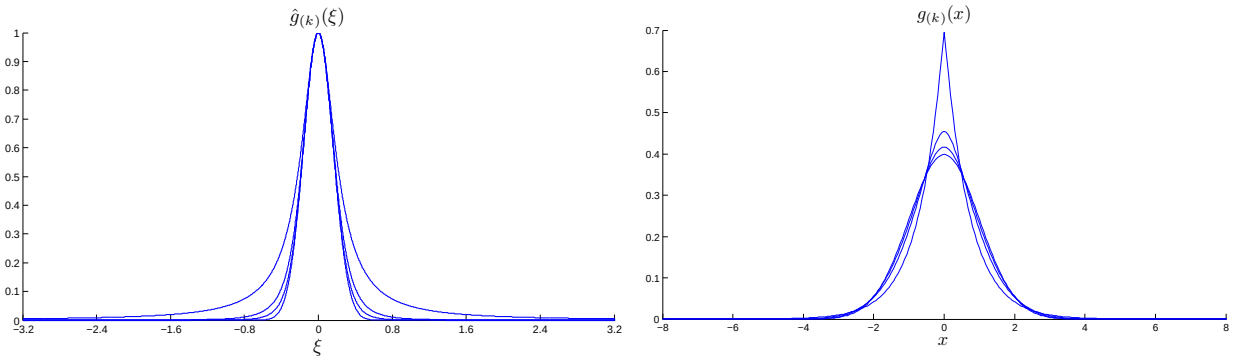
which, for sufficiently large values of k, is a good approximation of the Gaussian kernel (see Fig. 4 for a numerical example of computation of the inverse Fourier transforms of $\hat{g}(\xi)$ and $\hat{g}_{(k)}(\xi)$ via the Inverse Fast Fourier Transform (IFFT), implemented in Matlab 7.7).

A similar method can be used for other infinitely-smooth kernels that are free-space Green functions of infinite-order linear differential operators. Otherwise, direct extensions of the results to infinite-order differential operators may be obtained by setting the problem of learning from constraints on Sobolev spaces of infinite order (Dubinskij (1986)).

4. Case studies

In this section we discuss some instances of learning problems that involve hard and soft constraints. We show that the application of the representations given in Sections 3.1 - 3.3 leads in some cases to problems that can be attacked by using the mathematical and algorithmic apparatus of kernel machines. In particular, for the reactions of the constraints we provide expressions that clearly show the connections with classical kernel methods. As the examples refer to convex problems, there is no distinction between global and local minimizers $f^\star$ and f^o.

24. Of course, differential operators P associated with L via $L = (P^\star)'P$ can be constructed in several ways.



(a) (b)

Figure 4: For $\sigma^2 = 1$ and $d = 1$, (a) the functions $\hat{g}_{(k)}(\xi) := \left(\sum_{i=0}^{k} a_i \|2\pi\xi\|^{2i}\right)^{-1}$ with $a_i := \frac{\sigma^{2i}}{i!2^i}$, and (b) their inverse Fourier transforms $g_{(k)}(x)$ for several values of the positive integer k. As $k \to \infty$, $g_{(k)}(x) \to g_{(\infty)}(x) := \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$ in the $\mathcal{L}^2$ norm.

4.1 Supervised learning

The classical framework of supervised learning from examples is a particular case of learning from constraints. Interestingly, this classical framework can be reproduced either by pointwise or isoperimetric constraints. In the hard context, the first case is trivial and produces an interpolation of the supervised examples. However, in learning from supervised examples the constraints are typically interpreted in a soft sense. Given $\mathcal{Y} \subseteq \mathbb{R}^n$ and the training set $\mathcal{E}_L := \{(x_\kappa, y_\kappa) \in \mathcal{X} \times \mathcal{Y},\ \kappa \in \mathbb{N}_{\ell_s}\}$, a possible transcription by a single hard isoperimetric constraint of a collection of given hard pointwise constraints imposed on the training set is

$$\check{\Phi}(f) = \int_{\mathcal{X}} \sum_{j \in \mathbb{N}_n} \sum_{\kappa \in \mathbb{N}_{\ell_s}} V(y_{\kappa,j}, f_j(x)) \delta(x - x_\kappa) dx = 0, \tag{47}$$

where $V(\cdot,\cdot)$ is a continuous loss function, i.e., a continuous non-negative function $V : \mathbb{R}^2 \mapsto [0, +\infty)$ such that $V(z, z) = 0$ for each $z \in \mathbb{R}$. This can be translated into a soft constraint for the function space $\mathcal{F}$. Of course, V has to be properly chosen in dependence of the specific learning problem (e.g., classification or regression). Note that, for the case in which $V(z^{(1)}, z^{(2)}) \neq 0$ for every $z^{(1)}, z^{(2)} \in \mathbb{R}$ with $z^{(1)} \neq z^{(2)}$, as long as hard constraints are considered, the specific form of V is not important, since for this class of loss functions $\int_{\mathcal{X}} \sum_{j \in \mathbb{N}_n} \sum_{\kappa \in \mathbb{N}_{\ell_s}} V(y_{\kappa,j}, f_j(x)) \delta(x - x_\kappa) dx = 0$ implies $y_{\kappa,j} = f_j(x_\kappa)$, for each $j \in \mathbb{N}_n$ and $\kappa \in \mathbb{N}_{\ell_s}$. Instead, the form of the loss function becomes important when considering the correspondent soft constraints, in the sense that in the soft case different loss functions may lead to different optimal solutions to the correspondent learning problems. Finally, it is worth mentioning that in general the transcription by a single hard isoperimetric constraint of several pointwise constraints causes the loss of the constraint qualification, required to apply the technique of Lagrange multipliers.

Let us now focus on the soft framework. We have to find $f^\star \in \operatorname{argmin}_{f\in\mathcal{F}}\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f)$, where $\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f)$ is given in equation (33) with $p(x)1_{\mathcal{X}_i}(x) = p(x)\delta(x-x_i)$, assuming that each set $\mathcal{X}_i$ is made up of a single point[25]. 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. There are different possible choices for $\phi_i^{\geq}(f(x))$, which typically depend mostly on whether we face regression or classification problems. Here, we consider the following two cases.

- The *quadratic loss* $V_Q(u) := \frac{1}{2}u^2$, associated with the hard bilateral constraints $\phi_{i,j}(f_j(x)) = (y_{i,j} - f_j(x)) = 0$, which originates $\phi_i^{\geq}(f(x)) = \sum_{j\in\mathbb{N}_n} V_Q \circ \phi_{i,j}(f_j(x)) = \frac{1}{2}\sum_{j\in\mathbb{N}_n}(y_{i,j} - f_j(x))^2$. The quadratic loss is used for both classification and regression problems.
- The *hinge loss* $V_H(u) := (u)_+$, associated with the hard unilateral constraints[26] $\breve{\phi}_{i,j}(f(x)) = (1 - y_{i,j}\cdot f_j(x)) \leq 0$, which gives rise to $\phi_i^{\geq}(f(x)) = \sum_{j\in\mathbb{N}_n} V_H \circ \breve{\phi}_{i,j}(f_j(x)) = \sum_{j\in\mathbb{N}_n}(1 - y_{i,j}\cdot f_j(x))_+$. The hinge loss is used for classification, although related functions can be exploited in regression problems.

As we shall see in the following, the reactions of the constraints are related to the values $f_j(x_i)$ with respect to the targets $y_{i,j}$. Here, each $y_{i,j}$ denotes a given real number for regression problems and a given element of the set $\{-1, 1\}$ for classification problems.

4.1.1 QUADRATIC LOSS

For every $j \in \mathbb{N}_n$ and $x \in \mathcal{X}$, by Theorem 22 (i), (iii) the j-th component of the reaction of the generic i-th constraint is given by

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

Then

$$\begin{aligned}f_j^o(x) &= \frac{1}{\gamma_j}\sum_{i=1}^m g * \omega_{i,j}^{\geq}(x) = \frac{1}{\gamma_j}\sum_{i=1}^m g * \big(p(x)(y_{i,j} - f_j^o(x))\cdot\delta(x-x_i)\big) \\ &= \sum_{i=1}^m p(x_i)\frac{y_{i,j} - f_j^o(x_i)}{\gamma_j}g(x-x_i) = \sum_{i=1}^m \alpha_{i,j}^{(ql)}g(x-x_i), \qquad (48)\end{aligned}$$

where $\alpha_{i,j}^{(ql)} := p(x_i)\frac{y_{i,j} - f_j^o(x_i)}{\gamma_j}$. The computation of f^o from Theorem 22 shows how to compute the coefficients $\alpha_{i,j}^{(ql)}$: for $p(x_i) \neq 0$, they can be obtained by solving

$$\left[\frac{\gamma_j}{p(x_i)}\alpha_{i,j}^{(ql)} + \sum_{\kappa\in\mathbb{N}_m} g(x_i - x_\kappa)\alpha_{\kappa,j}^{(ql)}\right] = y_{i,j}\,. \qquad (49)$$

In the case in which all the $p(x_i)$'s are equal to $1/m$, equation (49) can be compactly re-written as

$$[m\gamma_j I_n + G]\,\alpha_{\cdot,j}^{(ql)} = y_{\cdot,j}\,. \qquad (50)$$

Here, $I_n \in \mathbb{R}^{n,n}$ is the $n \times n$ identity matrix and G is the Gram matrix associated with the input data x_i and g which, under suitable conditions (see, e.g., Gnecco et al. (2013b)) is the kernel of a RKHS. So, in this case we get the classical representer theorem used in kernel machines. This emerged in a very similar way in (Poggio and Girosi (1989)). Interestingly, we also find that

$$\omega_{i,j}^{\geq}(x) = p(x)(y_{i,j} - f_j^o(x))\delta(x-x_i) = \gamma_j\alpha_{i,j}^{ql}\cdot\delta(x-x_i).$$

This means that $\hat{\omega}_{i,j}^{\geq}(\xi) \propto \alpha_{i,j}^{ql}e^{-2\pi\iota\langle x_i,\xi\rangle}$, which is just the general form of equation (8) in case of vectorial functions.

25. As discussed in Section 2, this comes from merging the belief on the constraint with the probability distribution.
26. Here, we consider hard unilateral constraints of the form $\breve{\phi}_{i,j}(f(x)) \leq 0$, because this formulation is more natural than the equivalent one $-\breve{\phi}_{i,j}(f(x)) \geq 0$.

4.1.2 HINGE LOSS

Here, we consider the case $y_{i,j} \in \{-1, 1\}$. For every $j \in \mathbb{N}_n$ and $x \in \mathcal{X}$, by Theorem 22 (ii), (iii) the j-th component of the reaction of the generic constraint i is

$$\begin{aligned} \omega^{\geq}_{i,j}(x) &= -p(x)1_{\mathcal{X}_i}(x)\overline{\partial_{f_j}}\phi^{\geq}_i(f^o(x)) \\ &= -p(x)\delta(x - x_i)\overline{\partial_{f_j}}\left((1 - y_{i,j} \cdot f^o_j(x))_+\right) , \end{aligned}$$

where

$$-\overline{\partial_{f_j}}\left((1 - y_{i,j} \cdot f^o_j(x))_+\right)$$

is equal to 0 if $(1 - y_{i,j} \cdot f^o_j(x)) < 0$ and to $y_{i,j}$ if $(1 - y_{i,j} \cdot f^o_j(x)) > 0$, whereas if $(1 - y_{i,j} \cdot f^o_j(x)) = 0$, it denotes an element (to be found) either of the set $[0, 1]$, when $y_{i,j} = 1$, or $[-1, 0]$, when $y_{i,j} = -1$.

Then we get

$$\begin{aligned} f^o_j(x) &= \frac{1}{\gamma_j}\sum_{i=1}^{m} g * \omega^{\geq}_{i,j}(x) = \frac{1}{\gamma_j}\sum_{i=1}^{m} g * \left(-p(x)\overline{\partial_{f_j}}\left((1 - y_{i,j} \cdot f^o_j(x))_+\right) \cdot \delta(x - x_i)\right) \\ &= -\frac{1}{\gamma_j}\sum_{i=1}^{m} p(x_i)\overline{\partial_{f_j}}\left((1 - y_{i,j} \cdot f^o_j(x_i))_+\right) g(x - x_i) \\ &= \sum_{i=1}^{m} y_{i,j}\alpha^{(hl)}_{i,j} g(x - x_i), \end{aligned} \tag{51}$$

where, recalling that $y_{i,j} \in \{-1, 1\}$, we have $\alpha^{(hl)}_{i,j} := -p(x_i)\frac{y_{i,j}}{\gamma_j}\overline{\partial_{f_j}}\left((1 - y_{i,j} \cdot f^o_j(x))_+\right)$. Of course, we may find points x_i defined by indexes $i \in \mathbb{N}_m$ for which $\alpha^{(hl)}_{i,j} \neq 0$, for at least one choice of the other index j. We denote by $\mathcal{S}$ the set of such indexes i. Such points x_i correspond to the *support vectors* of the classical kernel machines. It is worth remarking that $\alpha^{(hl)}_{i,j} \in [0, p(x_i)/\gamma_j]$. In the case $p(x_i) = 1/m$, this is the classical result on the range of the weights in support vector machines (see, e.g., (Bishop (2006), Chapter 6)).

Following the same computations made above for the hinge loss, one can also derive the classical results for the ϵ-insensitive loss used for regression, using an adaptation of Theorem 22 (ii), (iii) to this case.

4.2 Learning from propositional descriptions

In this section we consider the problem of learning from propositional descriptions, like those given in Table 2 (vii). This extends naturally learning from supervised examples with the hinge loss to the situation in which the points are replaced by open sets $\mathcal{X}_i$ defined by the corresponding characteristic functions $1_{\mathcal{X}_i}(\cdot)$, and associated with the volumes $\mathrm{vol}(\mathcal{X}_i)$. The open sets can degenerate into single points and we suppose to be given m_d points and m_o open sets. Hence, $m = m_d + m_o$ (the subscripts "d" and "o"stand for "discrete" and "open", respectively)[27].

Likewise in Section 4.1.2, we need to express the (generalized) probability density of the data. A strong simplification arises when this is approximated by

$$\frac{1}{m}\sum_{i=1}^{m_o}\frac{1}{\mathrm{vol}(\mathcal{X}_i)}1_{\mathcal{X}_i}(x) + \frac{1}{m}\sum_{\kappa=1}^{m_d}\delta(x - x_\kappa)\,,$$

which corresponds to assuming a uniform distribution within the ordinary open sets and Dirac distributional degeneration for points. So, we make the hypothesis of working in an environment with uniform probability distribution $1/m$, where the agent gives the same weight to the constraints, and we give a reasonable answer to the question whether more importance should be given to a rule with respect to an example, or vice-versa. In addition, we assume that $1 - y_{i,j}f^o_j$ is nonzero on the sets $\mathcal{X}_i$ and does not change its sign within each of such sets. We formalize this as an assumption.

Assumption 4.1 [SIGN CONSISTENCY] *On each set $\mathcal{X}_i$, $i = 1, \ldots, m$, the quantities $1 - y_{i,j}f^o_j$, $j = 1, \ldots, n$, are nonzero and have constant signs.*

27. Note that m_d corresponds to the notation l_s used in Section 3.3 for the number of supervised examples.

We take $y_{i,j} \in \{-1, 1\}$ and invoke[28] Theorem 22. For every $x \in \mathcal{X}$, the reaction of the generic constraint is different depending on whether we are considering a degenerate set (point) or an ordinary one (open set). In the first case, by the previous analysis we get for every $\kappa \in \mathbb{N}_{m_d}$ and $j \in \mathbb{N}_n$

$$\omega_{\kappa,j}^{\geq,d}(x) = y_{\kappa,j}\alpha_{\kappa,j}^{(hl)}\delta(x - x_\kappa)\,.$$

For non-degenerate sets, under the sign consistency hypothesis we have

$$\omega_{i,j}^{\geq,o}(x) = -\frac{1}{m \cdot \mathrm{vol}(\mathcal{X}_i)} 1_{\mathcal{X}_i}(x) \frac{\partial}{\partial f_j} \left((1 - y_{i,j} f_j^o(\hat{x}_i))_+\right),$$

where $\hat{x}_i$ denotes here any point in $\mathcal{X}_i$. Let

$$\hat{\alpha}_{i,j}^{(hl)} := -\frac{y_{i,j}}{m \cdot \gamma_j} \frac{\partial}{\partial f_j} \left((1 - y_{i,j} f_j^o(\hat{x}_i))_+\right).$$

By the sign-consistency hypothesis, we get $\hat{\alpha}_{i,j}^{(hl)} \in \{0, \frac{1}{m \cdot \gamma_j}\}$ and such a value does not depend on the choice of $\hat{x}_i \in \mathcal{X}_i$. Likewise for points, we call *support set* every set $\mathcal{X}_i$ whose index $i \in \mathbb{N}_{m_o}$ satisfies $\hat{\alpha}_{i,j}^{(hl)} \neq 0$ for at least one index j. We denote by $\mathcal{S}_o$ the set of such indexes i. Interestingly, because of the sign-consistency hypothesis, the reaction of each support set comes out from the whole contribution of the set, as each point inside the set provides an equal contribution to the constraint reaction. Similarly, in the degenerate case, we denote by $\mathcal{S}_d$ the set of indexes of support vectors (which corresponds to the set $\mathcal{S}$ of Section 4.1.2). Now, if we define[29]

$$\beta(x; \mathcal{X}_i) := \left[\frac{1}{\mathrm{vol}(\mathcal{X}_i)} g(\cdot) * 1_{\mathcal{X}_i}(\cdot)\right]_x$$

as the *set kernel* associated to the pair $(x, \mathcal{X}_i)$, we end up with the representation

$$f_j^o(x) = \sum_{\kappa \in \mathcal{S}_d} y_{\kappa,j}\alpha_{\kappa,j}^{(hl)} g(x - x_\kappa) + \sum_{i \in \mathcal{S}_o} y_{i,j}\hat{\alpha}_{i,j}^{(hl)} \beta(x, \mathcal{X}_i)\,, \tag{52}$$

which is limited to the support sets and support vectors.

As already discussed at the end of Section 3.2, another useful approximation can be given when a set $\mathcal{U} := \{\tilde{x}_\kappa \in \mathbb{R}^d,\ \kappa = 1, \ldots, \ell_u\}$ of unsupervised examples is given, too, and used to estimate $p(x)$ according to a mixture of kernel functions[30] $g(x - \tilde{x}_\kappa)$, i.e.,

$$p(x) = \sum_{\kappa \in \mathbb{N}_{\ell_u}} \pi_\kappa g(x - \tilde{x}_\kappa)$$

for suitable coefficients π_κ's (to be determined, by using the information coming from the unsupervised examples). For the sake of simplicity, let us assume that we are given only non-degenerate sets, i.e., open sets[31]. Let us assume again the sign-consistency hypothesis. Similar arguments as before provide

$$\begin{aligned}
f_j^o(x) &= -\frac{1}{\gamma_j}\left[g(\cdot) * \sum_{i=1}^{m_o}\left(\sum_{\kappa=1}^{\ell_u}\pi_\kappa g(\cdot - \tilde{x}_\kappa)\right) 1_{\mathcal{X}_i}(\cdot) \frac{\partial}{\partial f_j}\left((1 - y_{i,j} f_j^o(\tilde{x}_i))_+\right)\right]_x \\
&= \left[g(\cdot) * \sum_{i=1}^{m_o} \check{\alpha}_{i,j}^{(hl)} y_{i,j}\left(\sum_{\kappa=1}^{\ell_u}\pi_\kappa g(\cdot - \tilde{x}_\kappa)\right) 1_{\mathcal{X}_i}(\cdot)\right]_x \\
&= \sum_{i \in \mathbb{N}_{m_o}} \check{\alpha}_{i,j}^{(hl)} \sum_{\kappa \in \mathbb{N}_{\ell_u}} \pi_\kappa y_{i,j} \cdot [g(\cdot) * (g(\cdot - \tilde{x}_\kappa) 1_{\mathcal{X}_i}(\cdot))]_x \\
&= \sum_{(i,\kappa) \in \mathbb{N}_{m_o} \times \mathbb{N}_{\ell_u}} \pi_\kappa \check{\alpha}_{i,j}^{(hl)} y_{i,j} \beta(x; \tilde{x}_\kappa, \mathcal{X}_i) = \sum_{(i,\kappa) \in \check{\mathcal{S}}_o \times \mathbb{N}_{\ell_u}} \check{\alpha}_{i,j,\kappa}^{(hl)} y_{i,j} \beta(x; \tilde{x}_\kappa, \mathcal{X}_i)\,,
\end{aligned} \tag{53}$$

28. Actually, here we are applying an extension of Theorem 22 to the case under consideration, which differs slightly from the latter but can be proved similarly (it combines the assumptions of Theorem 22 (i) and (ii) by considering both kinds of soft constraints, i.e., those on open sets and those on single points).

29. The notation $\left[\frac{1}{\mathrm{vol}(\mathcal{X}_i)} g(\cdot) * 1_{\mathcal{X}_i}(\cdot)\right]_x$ means that the argument $\cdot$ is evaluated at x.

30. For the sake of notational simplicity, we suppose that the kernel in this expansion coincides with the Green function g of the linear differential operator L.

31. The case in which also supervised points are given can be treated as in the previous case. Notice that, since the constraints are of the same kind, in this context it is tacitly assumed that their beliefs are the same.

where $\check{\alpha}^{(hl)}_{i,j,\kappa} := \pi_\kappa \check{\alpha}^{(hl)}_{i,j}$, $\check{x}_i$ denotes any point in $\mathcal{X}_i$,

$$\check{\alpha}^{(hl)}_{i,j} := -\frac{y_{i,j}}{\gamma_j} \frac{\partial}{\partial f_j} \left((1 - y_{i,j} f^o_j(\check{x}_i))_+ \right) ,$$

$$\beta(x; \tilde{x}_\kappa, \mathcal{X}_i) := \int_{\mathcal{X}} g(x - \zeta) g(\zeta - \tilde{x}_\kappa) 1_{\mathcal{X}_i}(\zeta) d\zeta ,$$

and the last equality in equation (53) follows by restricting the summation to the set $\check{\mathcal{S}}_o \subseteq \mathbb{N}_{m_o}$ of indexes i for which $\check{\alpha}^{(hl)}_{i,j} > 0$ for at least one index j (i.e., the set of indexes i associated to the support sets). Moreover, by the sign-consistency assumption one has $\check{\alpha}^{(hl)}_{i,j} \in \{0, \frac{1}{\gamma_j}\}$. Of course, once the coefficients $\check{\alpha}^{(hl)}_{i,j,\kappa}$ have been discovered assessing to the (unsupervised) data and to the prescribed supervisions on the sets $\mathcal{X}_i$, the sign-consistency assumption has to be checked on the solution f^o obtained in such a way.

Again, we have derived for f^o a representation that depends on a finite number of parameters. However, notice that, whereas the representations (48) and (51) are kernel expansions in which the kernel is the Green function of the differential operator L, in the case considered in this section the functions $\beta(x; \mathcal{X}_i)$ and $\beta(x; \tilde{x}_\kappa, \mathcal{X}_i)$ inherit a structure that, in addition to L, closely depends on the correspondent sets $\mathcal{X}_i$. For this reason, such functions are called *constraint-induced kernels*. More details on learning from propositional descriptions can be found in (Melacci and Gori (2013)), where the methodology is applied to the case in which the sets $\mathcal{X}_i$ are boxes, thus originating a particular case of constraint-induced kernels called therein *box kernels.*

4.3 Learning from hard bilateral holonomic constraints

First we deal with linear constraints without supervised examples, then we consider the situation in which also the latter are available.

4.3.1 Linear constraints and no supervised examples

Let $\mathcal{X} = \mathbb{R}^d$ and $\forall i \in \mathbb{N}_m,\ \forall x \in \mathcal{X}$ define

$$\phi_i(f(x)) := a_i' f(x) - b_i = 0 , \tag{54}$$

where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ are given. We consider problems of learning from hard constraints like those in Table 2 (v and vi). Basically, we have hard holonomic bilateral constraints that can be written in the form (54), where a_i' is the i-th row of a given constraint matrix $A \in \mathbb{R}^{m,n}$. The constraints can be compactly written as $Af(x) = b$, where $b \in \mathbb{R}^m$. We also suppose $b \neq 0$. In the following, we assume $n > m$ and $\mathrm{rank}(A) = m$. We discuss the solution for the class of so-called *rotationally-symmetric linear differential operators*

$$P := [\sqrt{\rho_0} D_0, \sqrt{\rho_1} D_1, \ldots, \sqrt{\rho_\kappa} D_\kappa, \ldots, \sqrt{\rho_k} D_k]' ,$$

as defined in (Gnecco et al. (2013b)), where each operator D_κ satisfies (12) and (13), $\rho_0, \rho_1, \ldots, \rho_\kappa, \ldots, \rho_k \geq 0$, and $\rho_k > 0$. Such operators correspond via $L = (P^\star)' P$ to $L = \sum_{\kappa=0}^k (-1)^\kappa \rho_\kappa \nabla^{2\kappa}$, which is invertible on $\mathcal{W}^{k,2}(\mathbb{R}^d)$ (see, e.g., (Gnecco et al. (2013b), Lemma 1)). In addition, we assume that all the components of γ are equal to some constant $\bar{\gamma} > 0$.

We first address the case $\rho_0 \neq 0$, for which we show by a counterexample that being a solution of the Euler-Lagrange equations (19) is not a sufficient condition to solve the associated problem of learning from hard constraints, although in this case such a problem is convex. Inspired by Theorem 18, we verify that the Euler-Lagrange equations (19) are satisfied for a constant function $\bar{f}(\cdot)$. We have $L\bar{f} = \sum_{\kappa=0}^k (-1)^\kappa \rho_\kappa \nabla^{2\kappa} \bar{f} = \rho_0 \bar{f}$ and $\nabla_f \phi_i(\bar{f}) = a_i$. Hence, from (19) we get $\bar{\gamma}\rho_0 \bar{f} + A'\lambda = 0$, where the Lagrange multipliers are elements of the constant vector $\lambda \in \mathbb{R}^m$. Now, every constant solution $\bar{f}$ to the algebraic equation above is also a solution to $\bar{\gamma}\rho_0 A\bar{f} + AA'\lambda = 0$ and, therefore, of $\bar{\gamma}\rho_0 b + AA'\lambda = 0$. Moreover, by the assumptions $n > m$ and $\mathrm{rank}(A) = m$, we have $\det[AA'] \neq 0$. So, we can determine the vector of Lagrange multipliers by

$$\lambda = -\bar{\gamma}\rho_0 [AA']^{-1} b$$

and, consequently, denoting by λ_i the i-th component of the vector λ, the reaction of the i-th constraint is given by $\omega_i = -a_i \lambda_i$. This, in turns, yields

$$\omega_i = \bar{\gamma}\rho_0 a_i \left([AA']^{-1} b \right)_i .$$

Hence, recalling that the overall reaction of the constraints is $\omega = \sum_{i=1}^{m} \omega_i$, the solution to the Euler-Lagrange equations (19) is given by

$$\bar{f} = \bar{\gamma}^{-1} g * \omega = \left(\rho_0 \int_{\mathcal{X}} g(\zeta) d\zeta \right) A'[AA']^{-1} b \,.$$

By $Lg = \delta$, we get $\sum_{\kappa=0}^{k} (-1)^{\kappa} \rho_{\kappa} \nabla^{2\kappa} g = \delta$. In terms of the Fourier transform $\hat{g}(\xi)$ of g, we have $\rho_0 \hat{g}(\xi) + \sum_{\kappa=1}^{k} \rho_{\kappa} \left(2\pi \|\xi\|\right)^{2\kappa} \hat{g}(\xi) = 1$. For $\xi = 0$ we obtain $\rho_0 \hat{g}(0) = \rho_0 \int_{\mathcal{X}} g(\zeta) d\zeta = 1$. Finally,

$$\bar{f} = A'[AA']^{-1} b. \tag{55}$$

Now, we check that the function $\bar{f}$ defined by (55) is *not* an optimal solution to the problem of learning from hard constraints, for which we have just solved the associated Euler-Lagrange equations. First of all, since $\bar{f}$ is a constant, each nonzero component $\bar{f}_j$ of $\bar{f}$ does not belong to the Sobolev space $\mathcal{W}^{k,2}(\mathbb{R}^d)$, so $\bar{f}$ does not belong to the ambient space $\mathcal{F}$. At a first look, this may be considered a minor point, since we may still replace the ambient space by a different one, for which the same form of the Euler-Lagrange equations still holds. The real issue is that, for the obtained $\bar{f}$, the value $\mathcal{E}(\bar{f}) = \|\bar{f}\|^2_{P,\gamma}$ assumed by the objective functional of the learning problem with hard constraints is not even finite[32], so such a function $\bar{f}$ cannot be an optimal solution even for other choices of the ambient space.

We now consider the case $\rho_0 = 0$. In such a situation, we can easily verify that $\bar{f} = A'[AA']^{-1} b$ solves the associated Euler-Lagrange equations with the constant choice $\lambda = 0$. Although such $\bar{f}$ does not belong to $\mathcal{F}$ (so, it is not an optimal solution to the original problem of learning from hard constraints when this is set on $\mathcal{F}$), its components $\bar{f}_j$ belong to the generalized Sobolev space $\mathcal{H}_P(\mathbb{R}^d)$ (Fasshauer and Ye (2011)), i.e., the set of functions $f_j : \mathbb{R}^d \to \mathbb{R}$ for which $\|f_j\|^2_P$ is finite. Finally, since $\mathcal{E}(f) = \|f\|^2_{P,\gamma} \geq 0$ for any admissible f and $\mathcal{E}(\bar{f}) = \|\bar{f}\|^2_{P,\gamma} = 0$, we can conclude a-posteriori that $\bar{f}$ is indeed an optimal solution $f^\star$ to the problem of learning from hard constraints above when this is set on

$$\bar{\mathcal{F}} = \underbrace{\mathcal{H}_P(\mathbb{R}^d) \times \ldots \times \mathcal{H}_P(\mathbb{R}^d)}_{n \text{ times}},$$

instead than simply on $\mathcal{F}$.

4.3.2 Linear constraints and supervised examples

Here we discuss the relevant in which a generalization of the previously-discussed linear constraints is combined with classical learning from supervised examples and the constant vector b is replaced by a function $b(\cdot)$. A preliminary version of these results has been presented without proofs in (Gori and Melacci (2010)). As a particular case, when no supervised examples are present, one also obtains an extension of the case discussed in Section 4.3.1 to the one of a vector-valued function $b(\cdot)$.

We carry out the analysis by assuming, as before, that $\forall i \in \mathbb{N}_m$ one has $\gamma_i = \bar{\gamma} > 0$. We suppose also that $\rho_0, \rho_k > 0$. We are now given a supervised learning set $\mathcal{L} := \left\{ (x_\kappa, y_\kappa),\ x_\kappa \in \mathbb{R}^d,\ y_\kappa \in \mathbb{R}^n,\ \kappa \in \mathbb{N}_{l_s} \right\}$ and the linear constraint $Af(x) = b(x)$, where $b \in \mathcal{C}_0^{2k}(\mathcal{X}, \mathbb{R}^m)$ is a smooth vector-valued function with compact support. Such a constraint is intended in the hard sense, whereas the given supervised pairs induce soft constraints expressed in terms of the quadratic loss[33].

32. Note that this is not in contrast with Theorem 13, about the existence of a global solution to the learning problem under hard constraints. Indeed, such a theorem cannot be applied here since the set $\mathcal{F}_{\mathcal{C}}$ is empty: any function $f(x)$ that satisfies the set of hard bilateral holonomic constraints $Af(x) = b$ for a constant vector $b \neq 0$ cannot belong to the ambient space $\mathcal{W}$. A possible way to solve this issue consists in reformulating the learning problem in a different ambient space and with a different functional. More precisely, we can consider the problem of finding $(c, u)^\star \in \arg\min_{(c,u) \in \tilde{\mathcal{F}}_{\mathcal{C}}} \|u\|^2_{P,\gamma}$, where $\tilde{\mathcal{F}}_{\mathcal{C}} := \{(c, u) : c \text{ is a constant vector in } \mathbb{R}^n, u \in \underbrace{\mathcal{W}^{k,2}(\mathcal{X}) \times \ldots \times \mathcal{W}^{k,2}(\mathcal{X})}_{n \text{ times}} \text{ and } A(c + u(x)) = b \text{ for all } x \in \mathcal{X}\}$, which has the optimal solution $(c, u)^\star = (\bar{f}, 0) = (A'[AA']^{-1} b, 0)$. Another possible way to attack the problem is illustrated in the next Section 4.3.2 and is obtained by replacing the constant vector b by a function $b(\cdot)$. To approximate the original problem, the function $b(\cdot)$ may be chosen "nearly constant" in some portion of interest of the domain.

33. This is reasonable in practice. For example, in the task of Table 2 $v.$, whereas the single asset functions can be learned in a soft way from supervised examples, a constraint on the assets, like the overall money available, must be intended in a hard sense.

Since this is a problem with mixed hard and soft constraints, we search for a solution $\bar{f}$ to the Euler-Lagrange equations

$$\bar{\gamma} L\bar{f}(x) + A'\lambda + \frac{1}{l_s}\sum_{\kappa=1}^{l_s}(\bar{f}(x) - y_\kappa)\delta(x - x_\kappa) = 0 \tag{56}$$

(see Theorem 23, with $c = 1$). Let us determine the vector of distributional Lagrange multipliers λ. We have

$$AL\bar{f} = A\sum_{\kappa=0}^{k}(-1)^\kappa \rho_\kappa \nabla^{2\kappa}\bar{f} = \sum_{\kappa=0}^{k}(-1)^\kappa \rho_\kappa A\nabla^{2\kappa}\bar{f} = \sum_{\kappa=0}^{k}(-1)^\kappa \rho_\kappa \nabla^{2\kappa}A\bar{f} = \sum_{\kappa=0}^{k}(-1)^\kappa \rho_\kappa \nabla^{2\kappa}b = Lb\,,$$

where $Lb \in \mathcal{C}_0^0(\mathcal{X}, \mathbb{R}^m)$ has compact support. Hence, from equation (56) we get

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

from which we find that the Lagrange multiplier distribution[34] λ is given by (see equation (45))

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

Now, if we plug this expression for λ into the Euler-Lagrange equations (56), we get

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

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

$$\bar{f}(x) = \bar{\gamma}^{-1}\int_{\mathcal{X}} g(\zeta)c(x - \zeta)d\zeta + \sum_{\kappa=1}^{l_s} Q^{\text{mixed}}\alpha_\kappa^{(ql)} g(x - x_\kappa). \tag{57}$$

Note that, in the absence of supervised examples and in the limit case $b(x) = b$ "nearly" constant (e.g., such that each of its components is a rectangular pulse of sufficiently large width), we get (55) as an approximation for $\|x\|$ not too large, since in that case $Lb \simeq \rho_0\, D_0\, b \simeq \rho_0 b$. So, the overall constraint reaction (of both hard and soft constraints) is

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

The coefficients $\alpha_\kappa^{(ql)}$ can be determined by following the same scheme as the one used for equation (48). Let $y = [y_1, \ldots, y_{l_s}] \in \mathbb{R}^{n,l_s}$ be the matrix of targets, where the κ-th column is associated with the corresponding sample x_κ and $\alpha^{(ql)} := [\alpha_1^{(ql)}, \ldots, \alpha_{l_s}^{(ql)}] \in \mathbb{R}^{n,l_s}$. By the definition of $\alpha^{(ql)}$, we have

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

where G is the Gram matrix of the input data and g, and $H : \mathcal{X} \to \mathbb{R}^{n,l_s}$ 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 13 (since for $\rho_0, \rho_k > 0$, $\|\cdot\|_{P,\gamma}$ is a Hilbert-space norm on $\mathcal{W}^{k,2}(\mathbb{R}^d)$ by (Gnecco et al. (2013b), Proposition 3), and the square loss is convex and continuous with respect to $\|\cdot\|_{P,\gamma}$, because under the stated conditions $\|\cdot\|_{P,\gamma}$ is a norm equivalent to the standard Hilbert-space norm of $\mathcal{W}^{k,2}(\mathbb{R}^d)$, which for $k > d/2$ is a RKHS) and the nonsingularity of the Jacobian matrix (44) associated with the set of hard constraints $Af(x) = b(x)$.

We conclude discussing the admissibility of the obtained solution (57). By an application of (Gnecco et al. (2013b), Theorem 3) about the smoothness properties of Green functions, it follows that $g \in \mathcal{W}^{k,2}(\mathbb{R}^d)$. This implies

34. Here, for uniformity of notation with the present section, a column vector is considered.

that $\bar{f} \in \mathcal{F}$, $\mathcal{E}_{\mathcal{C}}^{\text{mixed}}(\bar{f})$ (see equation (43)) is finite, and $\bar{f}$ is both a local and global minimizer (thanks to the convexity of the problem).

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.

4.4 Quadratic constraints

Let us consider the case of soft fulfillment of a holonomic quadratic constraint

$$\forall x \in \mathcal{X}_1 = \mathcal{X} = \mathbb{R}^d, \quad \phi_1(f(x)) = (f(x) - v(x))' Q (f(x) - v(x)) \,, \tag{58}$$

where $Q \in \mathbb{R}^{n,n}$ is a positive semi-definite matrix and $v(\cdot)$ is a given function. For simplicity, we take $v(x) := 0\,, \forall x \in \mathcal{X}_1$. This corresponds to a degree of mismatch $\mu_{\phi_1}(f) = \int_{\mathcal{X}_1} f'(x) Q f(x) dx$. We are also given a set (x_κ, y_κ), $\kappa = 1, \ldots, m$ of supervised examples, which are dealt with in a soft way through the quadratic loss. We assume that the belief of the holonomic constraint is uniform over $\mathcal{X}_1$ (e.g., including it in $p(x)$, we set[35] $p(x) \equiv \bar{p} > 0$). Recall that in the following, we use the symbol $\circ$ to denote the Hadamard (entrywise) product.

Proposition 28 *Let us consider the quadratic soft constraints (58) along with a collection of m supervised examples, dealt with in a soft way through the quadratic loss. Assume that:*

(i) $p(x) \equiv \bar{p} > 0\ ,\ \forall x \in \mathcal{X}_1 = \mathcal{X} = \mathbb{R}^d$;

(ii) $Q \in \mathbb{R}^{n,n}$ *is a symmetric positive definite matrix, for which we denote by* $\tilde{Q} := \operatorname{diag}[\tilde{q}_1, \ldots, \tilde{q}_n] \in \mathbb{R}^{n,n}$ *the full-rank diagonal matrix of its eigenvalues and by* $T \in \mathbb{R}^{n,n}$ *an invertible matrix such that* $T\tilde{Q}T^{-1} = Q$;

(iii) (a) L *is invertible on* $\mathcal{W}^{k,2}(\mathcal{X})$ *and (b) it has a free-space Green function* $g \in \mathcal{W}^{k,2}(\mathcal{X})$;

(iv) the operator $L^\omega := L + p_\gamma \circ \tilde{Q} I$ *is invertible*[36] *on* $\mathcal{F}$, *where* $p_\gamma := \gamma^{-1}\bar{p} := \left[\gamma_1^{-1}\bar{p}, \ldots, \gamma_n^{-1}\bar{p}\right]'$ *and* I *is the identity operator, and for every* $j \in \mathbb{N}_n$ *the operator* $L_j^\omega := L + p_{\gamma_j}\tilde{q}_j I$ *(where* $p_{\gamma_j} := \gamma_j^{-1}\bar{p}$*) is invertible on* $\mathcal{W}^{k,2}(\mathcal{X})$ *and has a free-space Green function* $g_j^\omega \in \mathcal{W}^{k,2}(\mathcal{X})$.

Then, for the components of any locally optimal solution f^o of this instance of the problem of learning with soft constraints, one has

$$\forall j \in \mathbb{N}_n : \quad f_j^o(x) = \sum_{\kappa=1}^{m} \sum_{l=1}^{n} \alpha_{\kappa,j,l}^\omega g_l^\omega (x - x_\kappa) \,, \tag{59}$$

and $g^\omega := [g_1^\omega, \ldots, g_n^\omega]'$ is the unique solution to

$$\forall j \in \mathbb{N}_n : \quad g_j^\omega + \breve{q}_j \cdot g * g_j^\omega = g, \tag{60}$$

where $\breve{q}_j := p_{\gamma_j}\tilde{q}_j$. When Q is diagonal, $\alpha_{\kappa,j,l}^\omega = 0$ for $l \neq j$, hence (59) reduces to the form

$$\forall j \in \mathbb{N}_n : \quad f_j^o(x) = \sum_{\kappa=1}^{m} \alpha_{\kappa,j}^\omega g_j^\omega (x - x_\kappa) \,. \tag{61}$$

Proof. By a slight extension of Theorem 22[37], following the arguments of Section 4.1, we have

$$Lf^o(x) = -p_\gamma \circ Q f^o(x) + \sum_{\kappa=1}^{m} \alpha_\kappa \delta(x - x_\kappa) \,,$$

35. Although this $p(x)$ is not a probability density as it does not belong to $\mathcal{L}^1(\mathcal{X}_1)$, it still provides a way to weigh the soft constraint uniformly on the whole domain (see also the concept of "improper prior" in Bayesian statistics).
36. Here, again, an overloaded notation is used for L and also for I, since we use the same notation for such operators both when acting on $\mathcal{F}$ and when acting on $\mathcal{W}^{k,2}(\mathcal{X})$.
37. The theorem can be extended straightforwardly to the case - considered here - of a function $p(\cdot)$ that does not belong to $\mathcal{L}^1(\mathcal{X}_1)$.

where $\alpha_\kappa \in \mathbb{R}^n$, and $\alpha_{\kappa,j} = \frac{1}{m}(y_{\kappa,j} - f_j^o(x))$. By hypothesis, there exists an invertible matrix $T \in \mathbb{R}^{n,n}$ such that $T\tilde{Q}T^{-1} = Q$, so $Lf^o(x) = -p_\gamma \circ T\tilde{Q}T^{-1}f^o(x) + \sum_{\kappa=1}^m \alpha_\kappa \delta(x - x_\kappa)$. Let $\tilde{f}^o(x) := T^{-1}f^o(x)$ and $\tilde{\alpha}_\kappa = T^{-1}\alpha_\kappa$ be. We have $LT\tilde{f}^o(x) = TL\tilde{f}^o(x)$. Thus $L\tilde{f}^o(x) + p_\gamma \circ \tilde{Q}\tilde{f}^o(x) = \sum_{\kappa=1}^m \tilde{\alpha}_\kappa \delta(x - x_\kappa)$. By introducing the linear differential operators $L^\omega := L + p_\gamma \circ \tilde{Q}I$ and $L_j^\omega := L + p_{\gamma_j}\tilde{q}_j I$, we can compactly write

$$L^\omega \tilde{f}^o(x) = \sum_{\kappa=1}^m \tilde{\alpha}_\kappa \delta(x - x_\kappa) \tag{62}$$

and

$$L_j^\omega \tilde{f}_j^o(x) = \sum_{\kappa=1}^m \tilde{\alpha}_{\kappa,j} \delta(x - x_\kappa).$$

Then, for every $j \in \mathbb{N}_n$, a Green function g_j^ω for the operator L_j^ω satisfies the equation

$$L_j^\omega g_j^\omega = Lg_j^\omega + p_{\gamma_j}\tilde{q}_j g_j^\omega = \delta. \tag{63}$$

Now we prove by contradiction the uniqueness of g_j^ω. Let $\overline{g}_j^\omega$ be another Green function such that $g_j^\omega \neq \overline{g}_j^\omega$. Then $L\left(g_j^\omega - \overline{g}_j^\omega\right) + p_{\gamma_j}\tilde{q}_j\left(g_j^\omega - \overline{g}_j^\omega\right) = 0$. By the hypothesis (iv), we get $g_j^\omega - \overline{g}_j^\omega = 0$ and, therefore, we end up with a contradiction. So, g_j^ω and also g^ω are unique. By the hypothesis (iii), L is invertible, so if we apply $g*$ to both sides of equation (63) we get the equation (60). Finally, from equation (62), one obtains

$$\tilde{f}^o(x) = \sum_{\kappa=1}^m \tilde{\alpha}_\kappa \circ g^\omega(x - x_\kappa)$$

and, from $f^o(x) = T\tilde{f}^o(x)$,

$$f^o(x) = T\left(\sum_{\kappa=1}^m \tilde{\alpha}_\kappa \circ g^\omega(x - x_\kappa)\right),$$

from which we get (59), for some coefficients $\alpha_{\kappa,j,l}^\omega$. In the particular case in which Q is diagonal, T can also be chosen to be diagonal, hence one obtains (61), for other coefficients $\alpha_{\kappa,j}^\omega$. ■

Let us consider the linear operator $\mathcal{T} := g*$ and define $\lambda_j := -\breve{q}_j^{-1} = (p_{\gamma_j}\tilde{q}_j)^{-1}$. We can promptly see that equation (60) can be re-written as

$$(\mathcal{T} - \lambda_j I)\, g_j^\omega = -\lambda_j g\,.$$

Hence, for every $j \in \mathbb{N}_n$ the Green function g_j^ω solves a classical Fredholm equation of the II kind (Kreyszig (1989)). For this reason, the kernels g_j^ω and g^ω derived from g and satisfying equation (60) are referred to as the *Fredholm kernels*. They will be addressed in Section 5.1.4.

5. Algorithmic framework and applications

The analysis carried out so far has been focused on functional representations of local or global optimal solutions to constrained learning problems. However, in order to have an experimental impact, the proposed investigation must be paired with a consequent algorithmic framework for the actual computation of such optimal solutions. In this section, we show that our approach is very well-suited to exploit constraints deriving from the availability of huge amounts of unsupervised data. Basically, since supervised data represent just special constraints, the viewpoint that emerges from our analysis is that of dismissing the difference between supervised and unsupervised data. While, in principle, the case of hard constraints can be attacked even without unsupervised data, the development of algorithms for soft constraints does rely on their massive availability. Intuitively, it is their availability that allows the constructions of penalties to check the satisfaction of each constraint.

The overall analysis carried out in the paper shows that learning from constraints is generally reduced to finding the constraint reactions. Whereas for both hard and soft constraints we get the same structure for the Euler-Lagrange equations, there is a strong difference since in the first case we need to determine the Lagrange multipliers. Fig. 5, which gives an overall view of algorithmic approaches for holonomic, pointwise, and isoperimetric constraints, shows that the hard (hr) and soft (sf) cases require different treatments. Moreover, whereas hard isoperimetric and hard pointwise constraints share a structure that is similar to the one obtained for hard holonomic ones, their

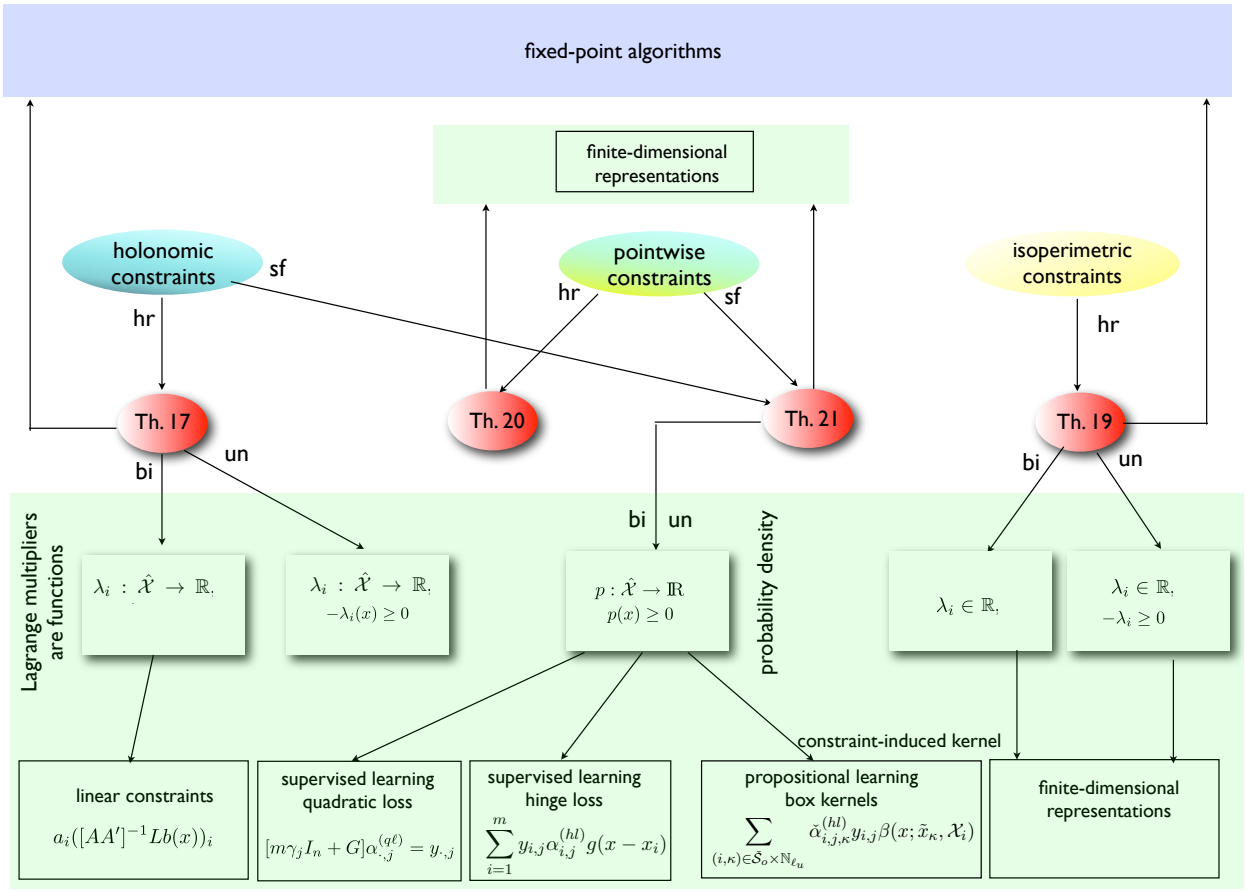


Figure 5: From representations to algorithms. The green blocks represent methods to reduce constrained learning to finite dimension and re-use the mathematical and algorithmic apparatus of classical kernel machines. We can either see plain kernels and constraint-induced kernels. An alternative possibility, represented by the blue block, is to use memory-based approaches along with fixed-point algorithms. While for isoperimetric and pointwise constraints the Lagrange multipliers are real numbers (non-positive for unilateral constraints), for holonomic constraints the Lagrange multipliers turn out to be functions (more generally, distributions). Finally, in the case of soft fulfillment, the (possibly generalized) probability density plays the role of the (sign-flipped) Lagrange multipliers used in the hard case.

Lagrange multipliers are constants instead of functions or distributions. Therefore, compared to the representation provided in Theorem (18) (ii), the ones given in Theorems 20 (ii) and 21 (ii) still give rise to non-linear forms of the Fredholm equation of the II kind, but the constraint reactions are easier to compute. The last row of blocks in Fig. 5 indicates that the infinite-dimensional functional representation of the optimal solution collapses to finite dimension in a number of relevant cases, which are discussed in detail in Section 5.1. While it is always possible to come up with the approximation of sampling the constraints (thus, obtaining pointwise constraints), the finite-dimensional representations obtained in such cases and presented in Sections 3 and 4 (see also the blocks associated with the pointwise constraints in Fig. 5) make it possible to determine the optimal solution, thanks to proper kernels induced by the constraints. A similar remark holds for isoperimetric constraints (see the last block in Fig. 5). Finally, Fig. 5 shows also that we might go beyond algorithmic schemes focusing on reduction to classical kernel machines and face directly the posed learning problem in the original ambient space (see Section 5.2).

5.1 Reduction to kernel machines

A straightforward way to apply the proposed approach consists in exploring the reduction to classical kernel machines. Under certain assumptions, the search for an optimal solution to the problem of learning from hard or soft constraints can be restricted to finite-dimensional spaces. We call this phenomenon the *collapse of the dimensionality*. The next proposition investigates cases for which it occurs. For simplicity of exposition, it refers to a globally-optimal solution, although some results can be extended to locally-optimal ones.

Proposition 29 *Suppose that a globally optimal solution $f^\star$ to a problem of learning from hard or soft constraints takes on the structure*

$$f^\star = \sum_{i=1}^{m} \gamma^{-1} g * \omega_i \,, \tag{64}$$

where each j-th component $\omega_{i,j}$ of ω_i ($j = 1, \ldots, n$) is of the form

$$\omega_{i,j}(\cdot) = \alpha_{i,j} \beta_{i,j}(\cdot) \tag{65}$$

for a given function $\beta_{i,j} : \mathbb{R}^d \to \mathbb{R}$ and the coefficients $\alpha_{i,j}$ are to be determined. Then the following hold.

(i) $\| f^\star \|^2_{P,\gamma}$ depends on the $\alpha_{i,j}$'s only, since it has the expression

$$\| f^\star \|^2_{P,\gamma} = \sum_{j_1,j_2=1}^{n} \sum_{i_1,i_2=1}^{m} \gamma_{j_1}^{-1} \gamma_{j_2}^{-1} \alpha_{i_1,j_1} \alpha_{i_2,j_2} \langle P(g * \beta_{i_1,j_1}), P(g * \beta_{i_2,j_2}) \rangle \,. \tag{66}$$

(ii) For the case of a problem with hard constraints, the optimal coefficients $\alpha^\star_{i,j}$'s are determined by minimizing (66) while imposing the hard satisfaction of the constraints.

(iii) For the case of a problem with soft constraints, the optimal coefficients $\alpha^\star_{i,j}$'s are determined by substituting the expressions (64) and (65) into the penalty term $\mu_{\mathcal{C}}(\cdot)$, then minimizing the sum of (66) and the resulting penalty term.

Proof. The proof of (66) is immediate. Items (ii) and (iii) follow directly from the structural properties of $f^\star$ and ω_i expressed by (64) and (65), respectively, and the optimality of $f^\star$. ■

It is also worth remarking that, when the collapse of the dimensionality occurs, standard finite-dimensional convex optimization algorithms (Boyd and Vandenberghe (2004)) can be used to find the reactions of the constraints (hence, to determine the support constraints) for the case of convex instances of the constrained learning problems.

5.1.1 Plain kernels

In the following discussion, we refer to soft pointwise constraints. Whereas Theorem 22 gives a general representation of an optimal solution, in the special case of supervised learning, equation (48) for the quadratic loss and its analogous (51) for the hinge loss provide directly the structural representation of the optimal body of the agent, which turns out to be defined by a finite number of parameters. More precisely, the representation of the optimal solution given by Theorem 22 collapses to the finite space of the coefficients $\alpha^{(ql)}_{i,j}$ for the quadratic loss and $\alpha^{(hl)}_{i,j}$ for

the hinge loss. A crucial consequence is that we are now ready to convert the infinite-dimensional problem attacked by Theorem 22 into finite dimension, simply by plugging equation (48) for the quadratic loss (respectively, equation (51) for the hinge loss) into $\mathcal{E}^{\text{soft}}(\cdot)$ and then by searching for the optimal parameters. In the case of quadratic loss with equal probabilities, this plugging leads directly to solving the linear system of equations (50), but with other loss functions we end up with non-linear equations. More generally, from $\phi_i^{\geq}(f(x_i)) = \int_{\mathcal{X}} \phi_i^{\geq}(f(x))\delta(x-x_i)dx$ we get the identity $\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f) = \frac{1}{2} \parallel f \parallel_{P,\gamma}^2 + \sum_{i=1}^m \phi_i^{\geq}(f(x_i))$, which can be combined with the representation of the optimal solution to obtain its coefficients. For instance, in the case of the hinge loss, any function f that satisfies the structural property for an optimal solution determined by equation (51) provides the following expression for $\parallel f \parallel_{P,\gamma}^2$:

$$
\begin{aligned}
\parallel f \parallel_{P,\gamma}^2 &= \sum_{j=1}^n \gamma_j \langle P \sum_{i=1}^m y_{i,j}\alpha_{i,j}^{(hl)} \cdot g(x-x_i), P \sum_{i=1}^m y_{i,j}\alpha_{i,j}^{(hl)} \cdot g(x-x_i)\rangle \\
&= \sum_{j=1}^n \gamma_j \langle \sum_{i=1}^m y_{i,j}\alpha_{i,j}^{(hl)} \cdot g(x-x_i), (P^\star)' P \sum_{i=1}^m y_{i,j}\alpha_{i,j}^{(hl)} \cdot g(x-x_i)\rangle \\
&= \sum_{j=1}^n \gamma_j \langle \sum_{i=1}^m y_{i,j}\alpha_{i,j}^{(hl)} \cdot g(x-x_i), L \sum_{i=1}^m y_{i,j}\alpha_{i,j}^{(hl)} \cdot g(x-x_i)\rangle \\
&= \sum_{j=1}^n \gamma_j \langle \sum_{h=1}^m y_{h,j}\alpha_{h,j}^{(hl)} \cdot g(x-x_h), \sum_{\kappa=1}^m y_{\kappa,j}\alpha_{\kappa,j}^{(hl)} \cdot Lg(x-x_\kappa)\rangle \\
&= \sum_{j=1}^n \gamma_j \langle \sum_{h=1}^m y_{h,j}\alpha_{h,j}^{(hl)} \cdot g(x-x_h), \sum_{\kappa=1}^m y_{\kappa,j}\alpha_{\kappa,j}^{(hl)} \cdot \delta(x-x_\kappa)\rangle \\
&= \sum_{j=1}^n \gamma_j \sum_{h=1}^m \sum_{\kappa=1}^m g(x_\kappa - x_h) y_{h,j}\alpha_{h,j}^{(hl)} y_{\kappa,j}\alpha_{\kappa,j}^{(hl)} \\
&= \sum_{j=1}^n \gamma_j (y_{\cdot,j} \circ \alpha_{\cdot,j}^{(hl)})' \, G \, (y_{\cdot,j} \circ \alpha_{\cdot,j}^{(hl)}),
\end{aligned}
$$

where G is the Gram matrix associated with the input data and g, and $\circ$ denotes the Hadamard product. Hence, we get

$$
\mathcal{E}_{\mathcal{C}}^{\text{soft}}(f) = \frac{1}{2} \sum_{j=1}^n \gamma_j (y_{\cdot,j} \circ \alpha_{\cdot,j}^{(hl)})' G (y_{\cdot,j} \circ \alpha_{\cdot,j}^{(hl)}) + \sum_{i=1}^m \phi_i^{\geq}(f(x_i)) \, .
$$

Finally, when considering that the penalty term $\sum_{i=1}^m \phi_i^{\geq}(f(x_i))$ depends only on the vector of parameters $\alpha^{(hl)}$, we end up with the optimization of an objective function of the form $\hat{\mathcal{E}}^{\text{soft}}(\alpha^{(hl)})$ in the finite-dimensional space of the parameters $\alpha^{(hl)}$. This is a classical scheme used in kernel machines and the perfect match arises when the Green function g of L is the kernel of a RKHS (Gnecco et al. (2013b)). As already mentioned, we call *g plain kernel*.

In both cases of quadratic and hinge losses, it is clear that there can be a non-zero reaction of a soft constraint whenever the associated hard constraint is not satisfied. In the case of quadratic loss, it 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[38] (apart from the case of an interpolating solution). On the opposite, when the hinge loss is used, the constraint reactions associated with some significantly large number of examples can be zero; this corresponds to the presence of *support* (*non-support*) *constraints*, of which the classical support (non-support, respectively) vectors are special cases.

5.1.2 SAMPLING-INDUCED KERNELS

There is a straightforward approach to collapse the dimensionality for the general case of holonomic constraints. The idea consists in sampling the constraints over a set of (unsupervised) examples $\mathcal{U} := \left\{\tilde{x}_\kappa \in \mathbb{R}^d, \ \kappa = 1, \ldots, \ell_u\right\}$,

38. Of course, the set $\mathcal{S}$ defined in Section 4.1.2 for the hinge loss can be also defined for the quadratic loss considered in Section 4.1.1.

so as to replace the original problem based on holonomic constraints into one based on pointwise constraints. When the constraints are sampled and dealt with in a soft way, we can promptly see that we end up exactly with the same scheme of dimensionality reduction discussed for supervised learning in Section 5.1.1, the only difference being that the accumulation of the loss is now over $\mathcal{U}$ instead than on the set of supervised pairs. As the corresponding plain kernel is obtained via a sampling process, we call it *sampling-induced kernel*. The consequent reduction nicely matches the analysis carried out in (Diligenti et al. (2012)), where the parsimony principle is imposed using the classical norm in a RKHS, expressed through its kernel. As remarked in Section 1.5, this approach is exploited by the software simulator available at

`https://sites.google.com/site/semanticbasedregularization/home/software`

which considers pointwise soft constraints and the classical plain kernel. Again, the plain kernel corresponds to the Green function of the operator L considered in this paper, and the learning problem is now framed into a finite-dimensional space of dimension $n \cdot |\mathcal{U}|$.

Interestingly, while the collapse of dimensionality with the correspondent reduction to the *plain kernel* comes out from sampling of the constraints, in general this is no longer true if soft holonomic constraints are used, even when one exploits the approximation of the probability density that leads to equation (42). This means that the collapse of dimensionality based on Proposition 29 is not always possible, since the reaction of the constraints might not be reducible to the structure expressed by equation (65). However, such a reduction is still possible when the approximation of the probability density is combined with soft pointwise constraints.

A final remark about sampling-induced kernels concerns the case in which the dimension of the input domain is high. In such a case, a large number of unsupervised examples may be needed to approximate the probability density satisfactorily. A similar problem may arise when the examples are generated from a probability density concentrated on an unknown manifold (this typically occurs, e.g., in machine-learning problems involving images). In such a case, the proposed approach can be complemented with tools and methods from manifold regularization (see (Belkin et al. (2006)) and (Melacci and Belkin (2011))).

5.1.3 Constraint-induced kernels

The collapse of dimensionality is not limited to pointwise constraints derived from supervised learning and constraint sampling. While in the discussion presented in Section 5.1.2 we have restricted the attention to holonomic constraints, related results can be derived for isoperimetric constraints. We also note that for the case of linear constraints discussed in Section 4.3 there exists a kernel-based representation theorem (Theorem 23) that is still based on the plain kernel. Hence, the analysis carried out in Sections 5.1.1 and 5.1.2 can be re-used so as to end-up into a finite-dimensional optimization problem. A finite-dimensional representation of the optimal solution is obtained also in Section 4.2 for the case of learning with propositional descriptions. Interestingly, such a representation is not expressed in terms of the plain kernel g, but via constraint-induced kernels. Moreover, the finite-dimensional representational structure that arises from the case studies presented in Section 4.2 can be plugged into $Lf^\star, \|f^\star\|^2_{P,\gamma}$ and the corresponding loss terms, exactly as shown in Section 5.1.1. So, also in this case learning with constraints is framed as a finite-dimensional optimization problem. The details of this computation, along with experimental analysis, can be found in (Melacci and Gori (2013)). In the next section we show how the same idea can be used for the case of quadratic constraints addressed in Section 4.4, which are associated with a particular class of constraint-induced kernels, i.e., the so-called the *Fredholm kernels*.

5.1.4 Fredholm kernels

In order to derive algorithms for the problem of learning under soft quadratic constraints and supervised examples (see Section 4.4), we need algorithms to compute g^ω (see equation (60)) and $\langle Lf^o, f^o \rangle$ when f^o is expanded according to (59). The following proposition indicates a method to compute g^ω. It is based on a classical iteration in normed spaces which, in this case, yields the Neumann series.

Proposition 30 *Let the hypotheses of Proposition 28 be satisfied, $\mathcal{T} := g*$, and assume that $\mathcal{T} : \mathcal{W}^{k,2}(\mathbb{R}^d) \rightarrow \mathcal{W}^{k,2}(\mathbb{R}^d)$. Consider the sequence $\{u^{(\kappa)}\}_{\kappa=0}^{\infty}$ defined as $u^{(0)} := g$ and $u^{(\kappa+1)} := \mathcal{T}(u^{(\kappa)})$, and assume that $\forall j \in \mathbb{N}_n\,, |\breve{q}_j| < 1/ \parallel \mathcal{T} \parallel$. Then, $\forall j \in \mathbb{N}_n$, one has*

$$g_j^\omega = \sum_{\kappa=0}^{\infty} (-1)^\kappa \breve{q}_j^\kappa \cdot u^{(\kappa)}, \tag{67}$$

where the series converges in $\mathcal{W}^{k,2}(\mathbb{R}^d)$.

Proof. We have

$$\| u^{(\kappa+1)} \| = \frac{\| \mathcal{T}(u^{(\kappa)}) \|}{\| u^{(\kappa)} \|} \| u^{(\kappa)} \| \leq \sup_{u \in \mathcal{W}^{k,2}(\mathbb{R}^d), u \neq 0} \left(\frac{\| \mathcal{T}(u) \|}{\| u \|} \right) \| u^{(\kappa)} \| = \| \mathcal{T} \| \cdot \| u^{(\kappa)} \| .$$

Since by assumption, $\forall j \in \mathbb{N}_n$ one has $|\breve{q}_j| < 1/ \| \mathcal{T} \|$, there exists $0 < \alpha < 1$ such that, $\forall j \in \mathbb{N}_n$ teh inequalities $\| \mathcal{T} \| \cdot |\breve{q}_j| < \alpha < 1$ hold. For every $\kappa \in \mathbb{N}$, let $M_\kappa := \| (-1)^\kappa \breve{q}_j^\kappa \cdot u^{(\kappa)} \|$. Then

$$M_{\kappa+1} = |\breve{q}_j|^{\kappa+1} \cdot \| u^{(\kappa+1)} \| \leq |\breve{q}_j|^{\kappa+1} \cdot \| \mathcal{T} \| \cdot \| u^{(\kappa)} \| \leq \| \mathcal{T} \| \cdot |\breve{q}_j| \left(|\breve{q}_j|^\kappa \| u^{(\kappa)} \| \right) < \alpha M_\kappa .$$

As $\sum_{\kappa=0}^{\infty} M_\kappa < M_0/(1-\alpha)$, by the Weierstrass M-test and the completeness of $\mathcal{W}^{k,2}(\mathbb{R}^d)$, we conclude that the series (67) converges in $\mathcal{W}^{k,2}(\mathbb{R}^d)$.

Finally, we can promptly check that the right-hand side of (67) satisfies equation (60) (hence, it coincides with g_j^ω, due to the uniqueness of the solution of (60)). Indeed, one has

$$-\breve{q}_j g * \left(\sum_{\kappa=0}^{\infty} (-1)^\kappa \breve{q}_j^\kappa u^{(\kappa)} \right) = \left(\sum_{\kappa=1}^{\infty} (-1)^\kappa \breve{q}_j^\kappa u^{(\kappa)} \right) = \left(\sum_{\kappa=0}^{\infty} (-1)^\kappa \breve{q}_j^\kappa u^{(\kappa)} \right) - g .$$

■

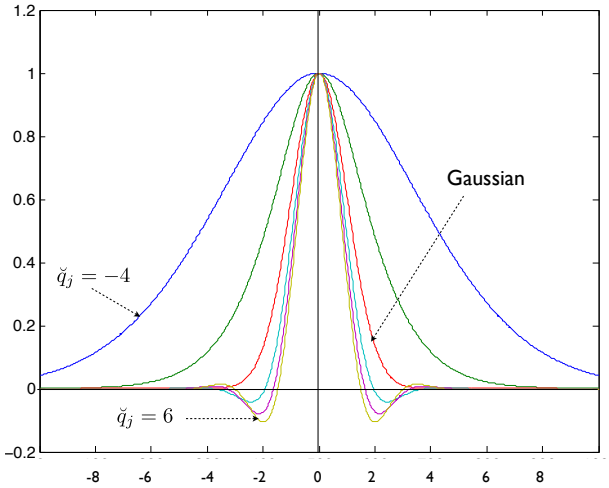


Figure 6: The Fredholm kernels corresponding to the Gaussian plain kernel g (with $d = 1, n = 1$, and $\sigma = 1$) for values of $\breve{q}_j$ ranging from -4 to 6. The plots have been obtained by using the Neumann series (67) and have been normalized to better assess the dependencies on $\breve{q}_j$. Although $\breve{q}_j$ has been defined in Section 4.4 as a non-negative real number, the series (67) can be formally defined also for negative values of $\breve{q}_j$, hence also such values are reported in the figure.

Notice that the condition $\forall j \in \mathbb{N}_n , |\breve{q}_j| < 1/ \| \mathcal{T} \|$ in Proposition 30 is merely sufficient for the convergence of the series.

To illustrate the result stated in Proposition 30, we consider the operator $\mathcal{T}$ obtained with the Gaussian kernel, which corresponds to the infinite-order linear differential operator $L := \sum_{i=0}^{\infty} (-1)^i \frac{\sigma^{2i}}{i! 2^i} \nabla^{2i}$. Although our theory has been developed only for the finite-order case, the operator above can be still dealt with as outlined in Section 3.6. For the case $d = 1$, the eigenvalues and eigenfunctions of the operator $\mathcal{T}$ corresponding to a Gaussian kernel are given by (Gradshteyn and Ryzhik (1980), p. 98)

$$\begin{aligned} \rho_{\kappa,\mathcal{T}} &= \sqrt{\frac{2a}{A}} B^\kappa , \\ \psi_{\kappa,\mathcal{T}}(x) &= \exp(-(c-a)x^2) \cdot H_\kappa(\sqrt{2c} \cdot x) , \end{aligned}$$

respectively, where $\kappa = 0, 1 \ldots,\ a^{-1} = 4\sigma^2$, $c = \sqrt{a^2 + a}$, $A = 1/2 + a + c$, $B = 1/(2A)$, and H_κ denotes the Hermite polynomial of order κ. The set of the eigenvalues is countable and its unique accumulation point is 0 (Kreyszig (1989), Theorem 8.3-1, pp. 421-422). Moreover,

$$\| \mathcal{T} \| = \sup_{\kappa=0,1,\ldots} \rho_{\kappa,\mathcal{T}} = \sqrt{\frac{4a}{1 + 2a + 2\sqrt{a^2 + a}}}.$$

In the case shown in Fig. 6 ($d = 1, n = 1$, and $\sigma = 1$), one has $\| \mathcal{T} \| = 2/(\sqrt{3 + 2\sqrt{3}})$, which guarantees the convergence of the Neumann series (67) for $\breve{q}_j < (\sqrt{3 + 2\sqrt{3}})/2 \simeq 1.271$, whereas the simulation of the series (67) indicates that the convergence holds also for larger values of $\breve{q}_j$. Notice that, in general, the application of the operator $\mathcal{T}$ is not feasible at high dimension. However, for most common plain kernels (i.e., Green functions of the operators L), we need not to use expensive numerical algorithms for the computation of the convolution. For instance, in the case of Gaussian plain kernels, the following proposition turns out to be useful.

Proposition 31 *Let $d = 1$ (i.e., $\mathcal{X} = \mathbb{R}$), $g_1(x) = c_1 e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$, $g_2(x) = c_2 e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$ and consider their convolution $g_{1*2}(x) = c_{1*2} e^{-\frac{(x-\mu_{1*2})^2}{2\sigma_{1*2}}}$ and their product $g_{1\times 2}(x) = c_{1\times 2} e^{-\frac{(x-\mu_{1\times 2})^2}{2\sigma_{1\times 2}}}$. Then*

$$i.\quad \mu_{1*2} = \mu_1 + \mu_2,\ \ \sigma_{1*2}^2 = \sigma_1^2 + \sigma_2^2,\ \ c_{1*2} = \sqrt{2\pi}\frac{c_1 c_2}{\sqrt{(\sigma_1^2)^{-1} + (\sigma_2^2)^{-1}}},$$

$$ii.\quad \mu_{1\times 2} = \frac{(\sigma_1^2)^{-1}\mu_1 + (\sigma_2^2)^{-1}\mu_2}{(\sigma_1^2)^{-1} + (\sigma_2^2)^{-1}},\ \ \sigma_{1\times 2}^2 = \frac{1}{(\sigma_1^2)^{-1} + (\sigma_2^2)^{-1}},\ \ c_{1\times 2} = c_1 c_2 e^{-\frac{(\mu_2-\mu_1)^2}{2(\sigma_1^2+\sigma_2^2)}}.$$

Proof. See (Bromiley (2003)). ■

Proposition 31 can be extended to the case $d > 1$. It allows one to avoid the numerical computation of the convolution in the Neumann series and makes the computations feasible in high-dimensional spaces for the case of the Gaussian kernel. The example clearly indicates that g^ω can be significantly different from the corresponding plain kernel g. In this case, as shown in Fig. 6, the values of $\breve{q}_j$ dramatically affect the structure of the kernel. Interestingly, g_j^ω is associated with the operator

$$L_j^\omega = \breve{q}_j I + \sum_{\kappa=0}^{\infty} (-1)^\kappa \frac{\sigma^{2\kappa}}{\kappa! 2^\kappa} \nabla^{2\kappa} = (1 + \breve{q}_j) I + \sum_{\kappa=1}^{\infty} (-1)^\kappa \frac{\sigma^{2\kappa}}{\kappa! 2^\kappa} \nabla^{2\kappa},$$

which differs from the above-considered infinite-order linear differential operator $L := \sum_{i=0}^{\infty} (-1)^i \frac{\sigma^{2i}}{i! 2^i} \nabla^{2i}$ only in the first term.

Now, recalling the form of an optimal solution for the case described by Proposition 28 when the matrix Q is diagonal, we show how to convert the infinite-dimensional optimization problem into a finite-dimensional one. We follow the scheme indicated in Section 5.1.1. From $g_j^\omega + \breve{q}_j g * g_j^\omega = g$ we get $L g_j^\omega = -\breve{q}_j g_j^\omega + \delta$. Then, for any f that satisfies the structural property for an optimal solution determined by equation (61), we get

$$
\begin{aligned}
\| f \|^2_{P,\gamma} &= \sum_{j=1}^{n} \gamma_j \langle Pf_j, Pf_j \rangle = \sum_{j=1}^{n} \gamma_j \langle f_j, Lf_j \rangle \\
&= \sum_{j=1}^{n} \gamma_j \langle \sum_{i=1}^{m} \alpha_{i,j}^{\omega} g_j^{\omega}(\cdot - x_i), L \sum_{\kappa=1}^{m} \alpha_{\kappa,j}^{\omega} g_j^{\omega}(\cdot - x_\kappa) \rangle \\
&= \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{\kappa=1}^{m} \gamma_j \alpha_{i,j}^{\omega} \alpha_{\kappa,j}^{\omega} \langle g_j^{\omega}(\cdot - x_i), L g_j^{\omega}(\cdot - x_\kappa) \rangle \\
&= \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{\kappa=1}^{m} \gamma_j \alpha_{i,j}^{\omega} \alpha_{\kappa,j}^{\omega} \langle g_j^{\omega}(\cdot - x_i), -\breve{q}_j g_j^{\omega}(\cdot - x_\kappa) + \delta(\cdot - x_\kappa) \rangle \\
&= -\sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{\kappa=1}^{m} \gamma_j \alpha_{i,j}^{\omega} \alpha_{\kappa,j}^{\omega} \breve{q}_j \langle g_j^{\omega}(\cdot - x_i), g_j^{\omega}(\cdot - x_\kappa) \rangle \\
&+ \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{\kappa=1}^{m} \gamma_j \alpha_{i,j}^{\omega} \alpha_{\kappa,j}^{\omega} g_j^{\omega}(x_\kappa - x_i). \qquad (68)
\end{aligned}
$$

To finalize the reduction to finite dimension, we need to express $G_{i,\kappa}^{j} := \langle g_j^{\omega}(\cdot - x_i), g_j^{\omega}(\cdot - x_\kappa) \rangle$. By the Neumann series, we get

$$
\begin{aligned}
G_{i,\kappa}^{j} &= \langle g_j^{\omega}(\cdot - x_i), g_j^{\omega}(\cdot - x_\kappa) \rangle = \langle \sum_{h=0}^{\infty} (-1)^h \breve{q}_j^h u^h, \sum_{\kappa=0}^{\infty} (-1)^\kappa \breve{q}_j^\kappa u^\kappa \rangle \\
&= \sum_{h=0}^{\infty} \sum_{\kappa=0}^{\infty} (-1)^{h+\kappa} \breve{q}_j^{h+\kappa} \langle u^h, u^\kappa \rangle.
\end{aligned}
$$

Whereas, in general, this is hard to compute at high dimension, there is a dramatic simplification with some plain kernels g. Again, we consider here the case of the Gaussian, for simplicity for $d = 1$. Then, from Proposition 31 we can directly express the product $u^h u^\kappa$ needed to compute $\langle u^h, u^\kappa \rangle$ by a Gaussian term with mean 0, variance $\sigma_u(h, \kappa) = \sigma^2 \cdot h\kappa/(h + \kappa)$, and constant multiplicative factor $\frac{1}{(\sqrt{2\pi}\sigma)^{h+\kappa-2}\sqrt{h\kappa}}$. Then, we get

$$
\begin{aligned}
G_{i,\kappa}^{j} &= \sum_{h=0}^{\infty} \sum_{\kappa=0}^{\infty} (-1)^{h+\kappa} \breve{q}_j^{h+\kappa} \langle u^h, u^\kappa \rangle \\
&= \sum_{h=0}^{\infty} \sum_{\kappa=0}^{\infty} (-1)^{h+\kappa} \breve{q}_j^{h+\kappa} \int_{\mathbb{R}^d} \frac{e^{-\frac{x^2}{2\sigma_u^2(h,\kappa)}}}{(\sqrt{2\pi}\sigma)^{h+\kappa-2}\sqrt{h\kappa}} dx \\
&= \sum_{h=0}^{\infty} \sum_{\kappa=0}^{\infty} (-1)^{h+\kappa} \breve{q}_j^{h+\kappa} \frac{\sqrt{2\pi}\sigma_u(h,\kappa)}{(\sqrt{2\pi}\sigma)^{h+\kappa-2}\sqrt{h\kappa}} \int_{\mathbb{R}^d} \frac{e^{-\frac{x^2}{2\sigma_u^2(h,\kappa)}}}{\sqrt{2\pi}\sigma_u(h,\kappa)} dx \\
&= (2\pi)^{\frac{3}{2}} \sigma^4 \sum_{h=0}^{\infty} \sum_{\kappa=0}^{\infty} (-1)^{h+\kappa} \breve{q}_j^{h+\kappa} \frac{\sqrt{h\kappa}}{(h+\kappa)(\sqrt{2\pi}\sigma)^{h+\kappa}} \\
&= (2\pi)^{\frac{3}{2}} \sigma^4 \sum_{h=0}^{\infty} \sum_{\kappa=0}^{\infty} (-1)^{h+\kappa} \frac{\sqrt{h\kappa}}{h+\kappa} \left(\frac{\breve{q}_j}{\sqrt{2\pi}\sigma} \right)^{h+\kappa}.
\end{aligned}
$$

Now, we can promptly see that, if there exists $\beta \in (0,1)$ such that, $\forall j \in \mathbb{N}_n : \; |\breve{q}_j| < \beta\sqrt{2\pi}\sigma$, then the series above converges. Indeed,

$$
\begin{aligned}
& (2\pi)^{\frac{3}{2}}\sigma^4 \left| \sum_{h=0}^{\infty}\sum_{\kappa=0}^{\infty}(-1)^{h+\kappa}\frac{\sqrt{h\kappa}}{h+\kappa}\left(\frac{\breve{q}_j}{\sqrt{2\pi}\sigma}\right)^{h+\kappa} \right| \leq (2\pi)^{\frac{3}{2}}\sigma^4 \left| \sum_{h=0}^{\infty}\sum_{\kappa=0}^{\infty}(-1)^{h+\kappa}\left(\frac{\breve{q}_j}{\sqrt{2\pi}\sigma}\right)^{h+\kappa} \right| \\
= \; & (2\pi)^{\frac{3}{2}}\sigma^4 \left| \sum_{h=0}^{\infty}(-1)^{h}\left(\frac{\breve{q}_j}{\sqrt{2\pi}\sigma}\right)^{h} \sum_{\kappa=0}^{\infty}(-1)^{\kappa}\left(\frac{\breve{q}_j}{\sqrt{2\pi}\sigma}\right)^{\kappa} \right| \leq (2\pi)^{\frac{3}{2}}\sigma^4 \frac{1}{(1-\beta)^2}.
\end{aligned}
$$

Finally, from (68) we can compute $\| f \|^2_{P,\gamma}$ since $G^j_{i,\kappa}$, likewise g^{ω}_j, is the limit of a convergent series. Then, if we plug equation (61) into the overall penalty loss we end up with a finite-dimensional optimization problem, which fits the classical approach used for kernel machines.

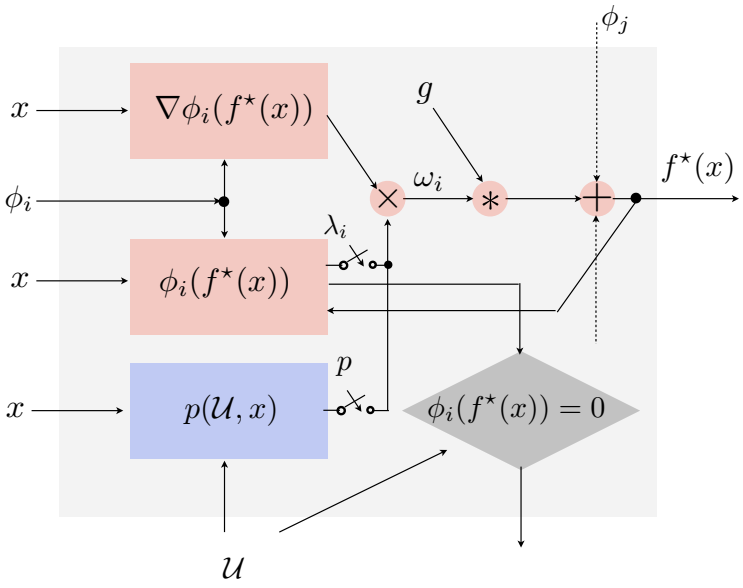


Figure 7: The structure of a support constraint machine. Depending on hard or soft constraints, the reaction of each constraint is computed by using the Lagrange multiplier or the (possibly generalized) probability density, which is often expressed in terms of the unsupervised data $\mathcal{U}$. The constraint reactions are then used to express the optimal solution $f^\star$ to the constrained learning problem. Finally, for any point x, the machine determines $f^\star(x)$ and checks the constraints.

5.2 Fixed-point algorithms

The constraint-induced kernels allow us to better represent the optimal solutions to certain problems of learning from constraints and, in some cases, the methodology perfectly fits the mathematical and algorithmic framework of kernel machines. For example, in case of simple geometry, constraint-induced kernels can be determined directly from the associated plain kernel. This is the case, e.g., of box kernels (see Fig. 2). For soft quadratic constraints, the computation of the constraint-induced kernel can be based on the Neumann-series recurrent scheme. In the general case, when looking at the representer theorems obtained for both hard and soft constraints, one can see that the computation of an optimal solution requires the associated constraint reaction ω, which depends on the optimal solution itself. This circular dependence is depicted in Fig. 7 (where we refer, without loss of generality, to a global optimal solution $f^\star$). Indeed we have to expect its emergence, which seems to be a sign of the inherent complexity of the problem at hand. For example, the representation given by Theorem 18 (ii) (see equation (20)) is a non-linear

version of the classical functional equation known as the *Fredholm equation of the II kind*[39]. The fact that classical supervised learning, learning from propositional descriptions, and other constraints yield to finite-dimensional optimization problems based on constraint-induced kernels that can be directly or efficiently determined by the plain kernel, reveals a specific property of those simple constraints. However, in general, the Lagrange multipliers that are not merely constants but functions or distributions: this makes the recurrent computation hard.

Based on these premises, whereas the reduction to classical kernel machines is a desirable link with existing approaches to machine learning, the given framework of Support Constraint Machines (SCMs) seems to be well-suited also for a *truly new way* of computing an optimal solution to the learning problem, which is not based on the discovering of the weights of a certain kernel expansion. More precisely, in the case of soft constraints we propose to attack directly the functional equations that have to be satisfied by an optimal solution by discovering a fixed-point of the corresponding operator. Indeed, in the case of soft holonomic constraints an optimal solution f^o can be written as $f^o = \mathcal{V}(f^o)$, where

$$\mathcal{V} : \mathcal{F} \to \mathcal{F} : f \to -\sum_{i=1}^{m} \gamma^{-1} g(\cdot) * p(\cdot) 1_{\mathcal{X}_i} \nabla_f \phi_i^{\geq}(\cdot, \cdot).$$

Related operators can be introduced for other types of soft constraints. A possible computational scheme to find f^o is the classical iteration $u^{(\kappa+1)} = \mathcal{V}(u^{(\kappa)})$, initialized by a given $u^{(0)} \in \mathcal{F}$.

An in-depth analysis of algorithms based on this scheme is outside the scope of this paper, but the following example of soft quadratic constraints, already addressed in Section 4.4 with the purpose of reduction to kernel machines, clearly shows the power of the direct computation of a fixed-point of $\mathcal{V}$. For instance, it follows from the proof of Proposition 28 that the optimal solution f^o is a fixed point of the operator $\mathcal{V}$ defined by

$$\mathcal{V}(u)_j \quad := \quad -p_{\gamma_j} \sum_{h=1}^{n} g * Q_{j,h} u_h + \frac{1}{m} \sum_{\kappa=1}^{m} \frac{y_{i,j} - u_j(x_i)}{\gamma_j} g(x - x_i)\,.$$

Then, in order to find f^o, one may apply, for such an operator, the iterative scheme $u^{(\kappa+1)} = \mathcal{V}(u^{(\kappa)})$ with the initialization $u_j^{(0)} := \sum_{i=1}^{m} \alpha_{i,j}^{(0)} g(x - x_i)$ for some coefficients $\alpha_{i,j}^{(0)}$. We can easily see that we can export to this case the proof technique applied to derive Proposition 30, and conclude that such an iterative scheme leads to the (unique) fixed point of $\mathcal{V}$, provided that $\mathcal{V}$ is a contraction operator. Moreover, the rate of convergence is linear, as it typically happens for iterative schemes based on contraction operators. As already noticed, the possible intractability of applying the operator $\mathcal{T} := g*$ to a generic function can be faced by appropriate plain kernels, like the Gaussian, for which one obtains the structural properties of convolution stated by Proposition 31. In that case, it easy to see that, if the iterative scheme above is initialized with Gaussian functions $u_j^{(0)}$, then, at each iteration κ, the functions $u_j^{(\kappa)}$ have Gaussian expansions, centered at $\{x_i,\ i = 1, \ldots, m\}$.

5.3 Applications

The theory of Support Constraint Machines (SCMs) presented in this paper provides foundations for applications to, amongst others, text categorization, face recognition, computer vision, medical diagnosis, and bioinformatics (see Table 4, second column). In the last few years, different types of constraints have been the subject of investigation, which motivates the overall view of learning from constraints given in this paper. Most emphasis has been given to holonomic constraints and propositional descriptions, but the applications based on iso-perimetric constraints (see Table 2) are clearly of interest. The applicative studies with holonomic constraints have primarily made use of the reduction to kernel-machines (Section 5.1), which arises when sampling the constraints on unsupervised data. Relevant applications have been published using propositional descriptions, which have been analyzed in Section 4.2. In this case, the corresponding representation of the learning tasks, given in (52), indicates that one goes beyond solutions based on plain kernels, thus showing the relevance of the case studies treated in Section 4.2. Only the case of box kernels has been concretely applied, but the remaining analysis of the section indicates other important contexts in which the proposed theory dictates the adoption of proper kernels.

39. There are a number of theoretical and numerical studies on this equation (see e.g., (Lishan (1996); Ivaz and Mostahkam (2006))), and particular attention has been devoted to the linear case (Kreyszig (1989)).

Learning from constraints: applications					
	application context	*type of constraints*	*results without / with additional constraints on the learning environment*	*source*	*instance of*
i.	Text classification	First-Order Logic (Class-relationships)	F1 Score 0.569 / 0.672	Frandina et al. (2012)	Sec. 3.2 Sec. 4.1.2 Sec. 5.1.2
ii.	Text classification	Coherence over the data manifold	Error Rate 20.04% / 9.34%	Melacci and Belkin (2011)	Sec. 3.2 Sec. 4.1.1 Sec. 5.1.2
iii.	Tagging bibliographical entries	First-Order Logic (Semantic relationships among tags)	F1 Score 0.140 / 0.155	Diligenti et al. (2012)	Sec. 3.2 Sec. 4.1.2 Sec. 5.1.2
iv.	Image Tagging	First-Order Logic descriptions	–	Saccà et al. (2011b)	Sec. 3.2 Sec. 4.1.2 Sec. 4.2
v.	Handwritten digit recognition	Propositional descriptions (Rules defined by observing digits)	Accuracy 89.78% / 92.55%	Melacci and Gori (2013)	Sec. 4.1.2 Sec. 4.2
vi.	Computer vision	Coherence on four views of the same object	Accuracy 92.53% / 94.67%	Melacci et al. (2009)	Sec. 3.2 Sec. 4.1.1 Sec. 5.1.2
vii.	Face recognition	Probabilistic constraints	Error Rate 41.36% / 39.32%	Melacci and Gori (2011)	Sec. 3.2 Sec. 4.1.1 Sec. 5.1.2
viii.	Computer vision	Mutual Information penalty	–	Gori et al. (2012)	Sec. 3.2 Sec. 4.1.1 Sec. 5.1.2
ix.	Breast cancer prognosis	Propositional descriptions (Rules provided by a physician)	Accuracy 82.58% / 90.97%	Melacci and Gori (2013)	Sec. 4.1.2 Sec. 4.2
x.	Bioinformatics	First-Order Logic constraints	AUC p: 0.808 / 0.820 d: 0.605 / 0.937 r: 0.591 / 0.676	Saccà et al. (2014)	Sec. 3.2 Sec. 4.1.1 Sec. 5.1.2

Table 4: Examples of applications that are instances of the theory proposed in this paper. Most of the experimental results are related to soft pointwise constraints. The cases *i,ii,iii,iv,vi,vii,viii,x* can be regarded as instances of *soft holonomic constraints*. Such constraints are then sampled onto a set of unsupervised points, obtaining *soft pointwise constraints*. The optimal solutions to the corresponding learning problems are described by the *sampling-induced kernels* presented in Section 5.1.2. Differently, *v* and *ix* are instances of *learning from propositional descriptions*, which is described in Section 4.2. The experimental results provide clear evidence on the beneficial effect of adding constraints to supervised examples. In the table, we use "–" to indicate that no comparison was carried out with the case of supervised learning with plain kernels.

The adoption of the methodology investigated in this paper requires one to go beyond learning from examples only, by modelling learning tasks in term of constraints. This guideline has already been followed in a number of remarkable different applications. Some of them are summarized in Table 4.

In *i* (Frandina et al. (2012)), a large set of scientific papers is classified by exploiting relationships between the classes, which are expressed in terms of First-Order Logic (FOL) formulas. They are translated into real-valued constraints by means of T-norms (Klement et al. (2000)), so as to gain a uniform real-valued representation of the constraints, which is the basic representational assumption of this paper. A subset of the CORA dataset[40] is then used to evaluate the impact of those additional constraints. The application faced in *ii* (Melacci and Belkin (2011)) exploits the popular manifold assumption applied to classification. It refers to a text classification task on newsgroups data[41] (see Melacci and Belkin (2011) for other examples and experiments). A constraint on the classifier output between each pair of data points is defined, which is weighted by a measure related to their distance. This constraint enforces smooth changes on the classifier output over the (estimated) manifold of the data distribution. Interestingly, this category of constraints can be simply regarded as an instance of the learning framework proposed in this paper. In *iii* (Diligenti et al. (2012)), a set of bibliographical entries (from the Bibtex dataset[42]) is paired with semantic relationships between their categories, for the purpose of tagging. The approach followed therein is similar to the one adopted in *i* but, instead of a pre-coordinate classification scheme, the intelligent agent deals with a post-coordinate scheme (tagging), which allows the users to attach keywords to documents without the constraint of placing them into a unique location. In *iv* (Saccà et al. (2011b)), related studies are carried out, where prior logic knowledge and graph regularization are integrated for an interesting application to image tagging. In *v* (Melacci and Gori (2013)), a classical pattern recognition problem is attacked using propositional descriptions. It is shown that handwritten digits (USPS dataset (Hull (1994))) are better discriminated when adding constraints on those digit portions that are well-known to be critical in the recognition process, thus dramatically reducing the need for large databases of supervised point-wise examples. The problem of object recognition is the topic of the experiments carried out in *vi* (Melacci et al. (2009)). One is given therein 100 distinct objects, and considers four classifiers operating on multiple views of the same object, which are constrained to produce a coherent decision (COIL-100 dataset, Columbia University[43]). The experiments cited in *vii* (Melacci and Gori (2011)) are based on constraining 295 classifiers to fulfill a probabilistic normalization, which is expressed by a linear constraint. This yields remarkable improvements in the face recognition benchmark XM2VTS database, University of Surrey[44]. In *viii* (Gori et al. (2012)), an extraction of computer vision features is proposed, which is based on the maximization of the mutual information between the video stream and a set of codes. The solution can be regarded as learning under a soft constraint related to mutual information. In *ix* (Melacci and Gori (2013)), the task consists in predicting whether a patient will remain cancer-free for at least 24 months, given the results of some medical tests (Wisconsin Breast Cancer Prognosis (Bache and Lichman (2013))). Two basic rules provided by a physician are considered, in the form of propositional descriptions, and two constraints on open sets are defined and exploited to improve the prediction accuracy. In *x* (Saccà et al. (2014)), it is shown that the adoption of the theory of learning from FOL constraints, properly translated into real-valued constraints as mentioned in i and iii, leads to a substantial improvement of performance over competitors in several experimental settings, for problems of protein-protein interactions. In particular, in the fourth column of row x, the improvement with respect to learning from examples with plain kernels is shown for three different levels of interactions (p: protein, d: domain, r: residuals).

Overall, inspection of the fourth column of Table 4 clearly shows the remarkable improvements of performances that have been reached by involving constraints. While these applications indicate that the interest in the field covered in this paper is growing, as discussed in Sections 5.1 and 5.2 there is also room for a significant follow-up of other ideas, along with their formalizations and corresponding algorithmic frameworks, possibly inspired by the present work.

6. Conclusions and open issues

In this paper, we have developed the mathematical foundations of a learning paradigm that is driven by a tight link of the parsimony principle with constraint-based representations of the environment[45]. These two ingredients lead

40. `http://people.cs.umass.edu/~mccallum/data.html`
41. `http://qwone.com/~jason/20Newsgroups/`
42. `http://mulan.sourceforge.net/datasets.html`
43. `http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`
44. `http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`
45. "Simplicity is the ultimate sophistication," Leonardo da Vinci.

to a variational framework that very much resembles the one used in other fields, such as Physics, Biology, and Economics. A major appeal of the proposed approach consists of its capability of providing functional representations of the optimal solutions, which are somehow prescribing the laws that govern learning in the given environment. While there are close connections with related studies especially in kernel machines, *a substantial novelty is the requirement of satisfying also constraints expressed by quantifiers acting on continuous infinite subsets of the perceptual space*. The given representation theorems, along with the emergence of their collapsed finite-dimensional versions, advocate the call for novel approaches to learning, which go beyond "sampling of concepts." Like support vectors are the only points in charge to support the decision in kernel machines, when shifting to more general constraints, we have shown that the same mechanism holds true for support constraints. Interestingly, the focus on constraints leads to dismiss the classical distinction between supervised and unsupervised learning, since agents only interact with constraints that, in case of soft fulfillment need (unsupervised) data to perform the check, while a supervised pair is just one of the simplest instances of a constraint.

We have provided foundations on a number of experimental studies and we have reviewed most of them (see, e.g., those based on the software simulator

`https://sites.google.com/site/semanticbasedregularization/home/software`

) at the light of the general theory herein proposed. To emphasize this, we have reported a summary of the experimental improvements achieved in a number of machine learning tasks, like text classification and tagging, handwritten character recognition, computer vision, face recognition, and applications to medicine and bioinformatics. By the introduction of constraint-induced kernels, we have analyzed a number of remarkable cases and have shown how to construct the corresponding *semantic kernel*. While this seems to be a promising research direction (see e.g., (Melacci and Gori (2013) for box kernels), the perspective of *fixed-point learning algorithms* might lead to more remarkable advances. As pointed out in Section 5.2, the iterative approach appears a natural computational scheme to take into account the inherent circular dependence of representer theorems. Such a dependency cannot be overcome in most interesting real-world problems.

A possible extension of our theory concerns the application of tools from Statistical Learning Theory (SLT), such as Rademacher's complexity (see, e.g, Mendelson (2003); Gnecco and Sanguineti (2008a,b) and the references therein), to investigate how the presence of 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 the presence of the latter, so one expects a smaller upper bound on the corresponding Rademacher's complexity, hence better bounds from SLT. However, this kind of investigation is outside the scope of the present work. We only mention that tools from SLT were applied in (Gnecco et al. (2013b)) to investigate the case of learning from supervised examples in the presence of additional boundary conditions.

While the results given in this paper might open the doors to a number of quite straightforward extensions and applications, there are still significant open issues that have not been faced.

First, we have assumed that the constraints are given, but in many real-world problems this might be quite difficult. We expect an enormous impact from the removal of the distinction between functions to be learned and constraints. Basically, this is equivalent to state that the constraints can be learned exactly like the functions. We conjecture that while the distinction can be profitably removed, in order to face complexity issues, it is necessary to introduce stages of learning, in which some functions that are kept fixed at some stage, so as to play the role of constraints, evolve at a further stage. There are some intriguing connections of this idea with developmental psychology (Inhelder and Piaget (1958); Piaget (1961)) and recent studies in developmental AI (see e.g., (Guerin (2008); Sloman (2009))), which might be sources of inspiration for further research.

Second, in some cases, the optimization task related to the learning problem is non-convex. This triggers a natural question related to the existence of locally-optimal solutions, and also to the choice of a good starting point for the solver. Issues related to bad local minima deserve attention in future studies. In this respect, one might think about the use of a multi-start technique, and also about the development of annealing strategies for the vector parameter γ. Indeed, large values of its components γ_j make the optimization problem "nearly convex". By gradually reducing them, one can hope to end up with an easier way to good local minima of problems with smaller values of the components γ_j, compared to the case in which such an optimization problem is addressed directly, without using annealing. Strategies that start the optimization from the optimal solution to a simpler optimization problem involving only a subset of the constraints could be thought as well. In some sense, this would prevent the

agent from being overloaded of information (constraints) during the early stages of the learning process, and thus becoming trapped in bad local solutions.

Third, a possible direction of evolution of this theory might come from noticing that, unlike what happens in other field of science, the variational approach used in this paper to provide foundations of learning and inference neglects the major role of time. We foresee an extension based on the overcoming of the distinction between learning and testing, based on the introduction of *temporal regularization mechanisms*, leading to on-line learning. A couple of papers (Frandina et al. (2013a), Frandina et al. (2013b)) gives insights into this view, which follows the growing interest for models of life-long learning (see e.g., the AAAI 2013 Spring Symposium "Lifelong Machine Learning": `http://www.seas.upenn.edu/~eeaton/AAAI-SSS13-LML`).

Finally, a further direction of improvement is represented by the exploration of the logic structure induced by a set of constraints and the mechanisms of *inference* of new constraints, somewhat related to classical notions of logic. In particular, a notion of *parsimonious inference* might be investigated, which is *supported* only by a subset of a given set of premises. In terms of constraints, when parsimony is invoked, a parsimonious inferential mechanism should emerge, that is based only on relevant constraints, while disregarding the rest.

Acknowledgments

We are primarily indebted to Marco Maggini, Michelangelo Diligenti, Marco Lippi, Claudio Saccà, and Salvatore Frandina for their invaluable support during our discussions at the AI-Lab of the University of Siena. We also received many constructive criticisms and suggestions during the preparation of this paper that we accumulated during some related talks. Marco Gori and Stefano Melacci were partly supported by the project PRIN2009 "Learning Techniques in Relational Domains and Their Applications" granted by MIUR (Italian Ministry of Education University and Research). Giorgio Gnecco and Marcello Sanguineti were partially supported 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 Genova. Marcello Sanguineti was also partially support by the Progetto di Ricerca di Ateneo 2013 "Dealing with High-Dimensional Data with Applications to Life Sciences". Marco Gori would like to dedicate this paper to the memory of his father.

Appendix A. Technical lemmas

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

The following lemma summarizes elementary properties of weakly lower semicontinuous functionals (see, e.g., (Evans, 2000, Appendix D)). It is exploited in the proofs of Theorems 13 and 14.

Lemma 32 *The following hold.*

(i) Every closed and convex subset of a Hilbert space is weakly closed (Mazur's Theorem (Evans (2000), 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[46]*.*

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

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

(v) Let $\mathcal{H}$ be a Hilbert space, S its nonempty and weakly closed subset, and $F : S \to \mathbb{R}$ a functional. If F is weakly lower semicontinuous and there exists $M \in \mathbb{R}$ such that the set $\{f \in S \mid F(f) \leq M\}$ is nonempty and weakly compact, then the problem $\inf_{f \in S} F(f)$ has a global minimizer, which is unique when F is strictly convex.

The next lemma, which is a consequence of the Implicit Function Theorem, is exploited in the proof of Theorem 18. For a scalar-valued function u of various vector arguments, we denote by $\nabla_i u$ the column vector of partial

46. In 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).

derivatives of u with respect to all the components of the i-th vector argument. 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 33 *Let $\Omega \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}^{n_1}$, $\mathcal{Z} \subseteq \mathbb{R}^{n_2}$ be open subsets, and $\phi : \Omega \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^{n_2}$ a given function. Let $y : \Omega \to \mathcal{Y}$ and $z : \Omega \to \mathcal{Z}$ be given functions that satisfy the (vector-valued) holonomic and bilateral constraint*

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

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

$$\nabla_3 \phi(x, y(x), z(x)) := \begin{pmatrix} \frac{\partial \phi_1(x,y(x),z(x))}{\partial z_1} & \cdots & \frac{\partial \phi_1(x,y(x),z(x))}{\partial z_{n_2}} \\ \cdots & \cdots & \cdots \\ \frac{\partial \phi_{n_2}(x,y(x),z(x))}{\partial z_1} & \cdots & \frac{\partial \phi_{n_2}(x,y(x),z(x))}{\partial z_{n_2}} \end{pmatrix} \tag{69}$$

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). Let η_y be an arbitrary function in $\mathcal{C}^k_0(\Omega, \mathbb{R}^{n_1})$ 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 $\eta_z \in \mathcal{C}^k_0(\Omega, \mathbb{R}^{n_2})$ 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

$$\Delta z(x) = \varepsilon \eta_z(x) + \mathcal{O}(\varepsilon^2)\,, \tag{70}$$

where the "hidden constant" inside the "big $\mathcal{O}$" notation above does not depend [47] on x, and $\eta_z(x)$ has the expression

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

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,

$$\frac{\partial^h}{\partial x_{j_1} \ldots \partial x_{j_h}} \Delta z_i(x) = \varepsilon \frac{\partial^h}{\partial x_{j_1} \ldots \partial x_{j_h}} \eta_{z_i}(x) + \mathcal{O}(\varepsilon^2)\,, \tag{71}$$

where, again, the "hidden constant" inside the "big $\mathcal{O}$" notation above does not depend on x.

Proof. Fix $x = x_0 \in \Omega$. Since $\phi \in \mathcal{C}^{k+1}(\Omega \times \mathcal{Y} \times \mathcal{Z}, \mathbb{R}^{n_2})$ for $k \geq 1$ and the Jacobian matrix (69) is nonsingular, we can apply the Implicit Function Theorem, according to which, on a suitable open ball $\mathcal{B}$ centered in $(0, 0)$ and of sufficiently small radius $\epsilon > 0$, there exists a unique function $u \in \mathcal{C}^{k+1}(\mathcal{B}, \mathbb{R}^{n_2})$ such that $u(0, 0) = 0$ and

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

Moreover, since[48] $k + 1 \geq 2$, each component $u_i(\Delta x, \Delta y)$ of the function $u(\Delta x, \Delta y)$ has the multivariate Taylor expansion

$$u_i(\Delta x, \Delta y) = \sum_{|\alpha|=1} D^\alpha u_i(0, 0)(\Delta x, \Delta y)^\alpha + \mathcal{O}(\|(\Delta x, \Delta y)\|^2)\,, \tag{72}$$

47. Instead, in (70) and (71) there is a dependence of the "hidden constants" on the specific choice of η_y, which may be removed by further assuming $\|\eta_y\|_{\mathcal{C}^k_0(\Omega, \mathbb{R}^{n_1})} \leq M_y$ for some given positive constant M_y.

48. In the lemma, we have made the assumption $\phi \in \mathcal{C}^{k+1}(\Omega \times \mathcal{Y} \times \mathcal{Z}, \mathbb{R}^{n_2})$ instead of the weaker one $\phi \in \mathcal{C}^k(\Omega \times \mathcal{Y} \times \mathcal{Z}, \mathbb{R}^{n_2})$, in order to be able to express the remainder in Taylor polynomial (72) by the integral Lagrange form, instead, e.g., of the Peano form. To avoid cumbersome notations, in (72) 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 $\mathcal{C}^2$ depending on a scalar argument x, we recall that one has

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

where the last term is the remainder in the integral Lagrange form. This formula can be generalized to the multivariate case, and such an extension is used to obtain terms of order $\mathcal{O}(\varepsilon^2)$ in (71).

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 $\mathcal{O}(\|(\Delta x, \Delta y)\|^2)$ denotes a function of class $\mathcal{C}^{k+1}(\mathcal{B})$, infinitesimal at $(0,0)$ with order at least 2. The "hidden constant" inside the "big $\mathcal{O}$" notation above depends only on the local behavior of ϕ on a neighborhood of $(x, y(x), z(x))$, and is independent of x itself, provided that, after the initial choice x_0 for x, x varies inside a compact subset Ω_C of the projection of the set[49] $\mathcal{B} + (x_0, y(x_0))$ on Ω. Now, let $\eta_y \in \mathcal{C}_0^k(\Omega_C, \mathbb{R}^{n_1}) \subseteq \mathcal{C}_0^k(\Omega, \mathbb{R}^{n_1})$ 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

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

where the replacement of the term $\mathcal{O}(\|(0, \varepsilon\eta_y(x))\|^2)$ by $\mathcal{O}(\varepsilon^2)$ follows by the fact that $\eta_y(x)$ is fixed and uniformly bounded. Then, we get (70) 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 $\mathcal{C}_0^k(\Omega_C, \mathbb{R}) \subseteq \mathcal{C}_0^k(\Omega, \mathbb{R})$, likewise η_y is in $\mathcal{C}_0^k(\Omega_C, \mathbb{R}^{n_1}) \subseteq \mathcal{C}_0^k(\Omega, \mathbb{R}^{n_1})$. An 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, (71) is derived directly by (70), by computing its partial derivatives of order h (i.e., by exploiting the expression of the remainder of Taylor polynomial (72) in Lagrange 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 33 is the following: in order to still be 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. Equation (71) shows that also the partial derivatives of $\Delta z(x)$ up to the order k have similar expressions.

References

R. A. Adams and J. F. Fournier. *Sobolev Spaces*. Academic Press, 2nd edition, 2003.

A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. *Advances in Neural Information Processing Systems*, 20:25–32, 2007.

H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces. Applications to PDEs and Optimization*. SIAM, Philadelphia, 2006.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

J.-L. Basdevant. *Variational Principles in Physics*. Springer, 2006.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. of Machine Learning Research*, 7:2399–2434, 2006.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, London, 2004.

M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IOP, Bristol, 1998.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

49. Here, we denote by $\mathcal{B} + (x_0, y(x_0))$ the translation of the set $\mathcal{B}$ by $(x_0, y(x_0))$.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004.

P. Bromiley. Products and convolutions of Gaussian distributions. Technical Report 2003-003, TINA Vision, 2003. URL `http://www.bibsonomy.org/bibtex/277a7563536e69987a49428ff9a734fba/ytyoun`.

A. Caponnetto, C.A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *J. of Machine Learning Research*, 9:1615–1646, 2008.

R. Caruana. Multi-task learning. *Machine Learning*, 28:41–75, 1997.

G. J. Chaitin. On the length of programs for computing finite binary sequences. *J. of the ACM*, 13:547–569, 1966.

Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14:2791–2846, 2002.

L. De Raedt, P. Frasconi, K. Kersting, and S.H. Muggleton, editors. *Probabilistic Inductive Logic Programming*, volume 4911 of *Lecture Notes in Artificial Intelligence*. Springer, 2008.

M. Diligenti, M. Gori, M. Maggini, and L. Rigutini. Multitask kernel-based learning with logic constraints. In *Proc. 19th European Conf. on Artificial Intelligence*, pages 433–438, 2010.

M. Diligenti, M. Gori, and M. Maggini. Learning to tag text from rules and examples. In *Lecture Notes in Computer Science, vol. 6934 (Proc. 12th Int. Conf. of the Italian Association for Artificial Intelligence)*, pages 45–56. Springer, 2011.

M. Diligenti, M. Gori, M. Maggini, and L. Rigutini. Bridging logic and kernel machines. *Machine Learning*, 86:57–88, 2012.

F. Dinuzzo and B. Schölkopf. The representer theorem for Hilbert spaces: A necessary and sufficient condition. In *Proc. 26th Annual Conf. on Neural Information Processing Systems (NIPS)*, pages 189–196, 2012.

J. A. Dubinskij. *Sobolev Spaces of Infinite Order and Differential Equations*. D. Reidel Publishing Company, 1986.

E. L. Ernestovic. *Differential Equations and Calculus of Variations*. Mir Publishers, Moskow, 1970.

L. C. Evans. *Partial Differential Equations*. AMS, Providence, 2000.

T. Evgeniou, C.A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *J. of Machine Learning Research*, 6:615–637, 2005.

T. Evgenious, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 1999.

G. E. Fasshauer and Q. Ye. Reproducing kernels of generalized Sobolev spaces via a Green function approach with distributional operators. *Numerische Mathematik*, 119:585–611, 2011.

S. Frandina, C. Saccà, M. Diligenti, and M. Gori. Constraint-based learning for text categorization. In *Proc. CoCoMile Workshop, 20th European Conf. on Artificial Intelligence*, pages 23–28, 2012.

S. Frandina, M. Gori, M. Lippi, M. Maggini, and S. Melacci. Variational foundations of online backpropagation. In V. Mladenov et al., editor, *Lecture Notes in Computer Science, vol. 8131 (Proc. 23rd Int. Conf. on Artificial Neural Networks)*, pages 82–89. Springer, Berlin Heidelberg, 2013a.

S. Frandina, M. Gori, M. Lippi, M. Maggini, and S. Melacci. Inference, learning, and laws of nature. In *Proc. NeSy Workshop - IJCAI2013*. Springer, 2013b. 4 pages.

I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Dover, 1963.

M. Giaquinta and S. Hildebrand. *Calculus of Variations I*, volume 1. Springer, 1996.

F. Girosi and G. Anzellotti. Rates of convergence for radial basis functions and neural networks. In R. J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 97–113. Chapman & Hall, London, 1993.

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7: 219–269, 1995.

F. Girosi, M. Jones, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

J. Gleick. *Genius: The Life and Science of Richard Feynman*. Pantheon Books, New York, 1992.

G. Gnecco and M. Sanguineti. Approximation error bounds via Rademacher complexity. *Applied Mathematical Sciences*, 2:153–176, 2008a.

G. Gnecco and M. Sanguineti. Estimates of the approximation error via Rademacher complexity: Learning vector-valued functions. *J. of Inequalities and Applications*, article ID 640758, 16 pages, 2008b.

G. Gnecco, M. Gori, S. Melacci, and M. Sanguineti. Learning with hard constraints. In V. Mladenov et al., editor, *Lecture Notes in Computer Science, vol. 8131 (Proc. 23rd Int. Conf. on Artificial Neural Networks)*, pages 146–153. Springer, 2013a.

G. Gnecco, M. Gori, and M. Sanguineti. Learning with boundary conditions. *Neural Computation*, 25:1029–1106, 2013b.

G. Gnecco, M. Gori, S. Melacci, and M. Sanguineti. A theoretical framework for supervised learning from regions. *Neurocomputing*, 129:25–32, 2014.

M. Gori. Semantic-based regularization and Piaget's cognitive stages. *Neural Networks*, 22:1035–1036, 2009.

M. Gori and S. Melacci. Learning with convex constraints. In K. Diamantaras, W. Duch, and L.S. Iliadis, editors, *Lecture Notes in Computer Science, vol. 6354 (Proc. 20th Int. Conf. on Artificial Neural Networks)*, pages 315–320. Springer, 2010.

M. Gori and S. Melacci. Constraint verification with kernel machines. *IEEE Trans. on Neural Networks and Learning Systems*, 24:825–831, 2013.

M. Gori, S. Melacci, M. Lippi, and M. Maggini. Information theoretic learning for pixel-based visual agents. In A. Fitzgibbon et al., editor, *Lecture Notes in Computer Science, vol. 7577 (Proc. 12th European Conf. on Computer Vision)*, pages 864–875. Springer, 2012.

I. S. Gradshteyn and I. M. Ryzhik. *Tables of Integrals, Series and Products. Corrected and Enlarged Edition Prepared by A. Jeffrey.* Academic Press, 1980.

F. Guerin. Constructivism in AI: Prospects, progress and challenges. In *Computing and Phylosophy: AISB 2008 Proc., Vol. 12*, pages 20–28. The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2008.

J. Hadamard. Sur les problemes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

R. Herbrich and R.C. Williamson. Algorithmic luckiness. *J. of Machine Learning Research*, 3:175–212, 2002.

B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

J.J. Hull. A database for handwritten text recognition research. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16:550–554, 1994.

B. Inhelder and J. Piaget. *The Growth of Logical Thinking from Childhood to Adolescence*. Basic Books, New York, 1958.

K. Ivaz and B.S. Mostahkam. Newton-Tau numerical solution of a system of nonlinear Fredholm integral equations of second kind. *Applied and Computational Mathematics*, 5:201–208, 2006.

J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*. Springer, 1994.

E.P. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer, 2000.

A. Kolmogorov. Two approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley & Sons, 1989.

L. Lishan. Iterative method for solutions and coupled quasi-solutions of nonlinear Fredholm integral equations in ordered Banach spaces. *Indian J. of Pure and Applied Mathematics*, 10:959–972, 1996.

D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley & Sons, 1969.

S. Melacci and M. Belkin. Laplacian Support Vector Machines Trained in the Primal. *J. of Machine Learning Research*, 12:1149–1184, 2011.

S. Melacci and M. Gori. Semi-supervised multiclass kernel machines with probabilistic constraints. In R. Pirrone and F. Sorbello, editors, *Lecture Notes in Computer Science, vol. 6934 (Proc. 12th Int. Conf. of the Italian Association for Artificial Intelligence)*, pages 21–32. Springer, 2011.

S. Melacci and M. Gori. Learning with box kernels. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35: 2680–2692, 2013.

S. Melacci, M. Maggini, and M. Gori. Semi-supervised learning with constraints for multi-view object recognition. In *Lecture Notes in Computer Science, vol. 5769 (Proc. 19th Int. Conf. on Artificial Neural Networks)*, pages 653–662. Springer, 2009.

S. Mendelson. A few notes on Statistical Learning Theory. In S. Mendelson and A. J. Smola, editors, *Lecture Notes on Computer Science (Advanced Lectures on Machine Learning)*, volume 2600, pages 1–40. Springer, 2003.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

J. Piaget. *La Psychologie de l'Intelligence*. Armand Colin, Paris, 1961.

T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical report, MIT - A. I. Memo No. 1140, C.B.I.P. Paper No. 31, 1989.

C. Saccà, M. Diligenti, M. Gori, and M. Maggini. Learning to tag from logic constraints in hyperlinked environments. In *Proc. 10th Int. Conf. on Machine Leraning and Applications Workshops, vol.2*, pages 251–256, 2011a.

C. Saccà, M. Diligenti, M. Gori, and M. Maggini. Integrating logic knowledge into graph regularization: An application to image tagging. In *Proc. 9th Workshop on Mining and Learning with Graphs*, San Diego, 2011b.

C. Saccà, S. Teso, M. Diligenti, and A. Passerini. Improved multi-level protein–protein interaction prediction with semantic-based regularization. *BMC Bioinformatics*, 15:103, 2014.

B. Schölkopf and A.J. Smola. From regularization operators to support vector kernels. In Morgan Kaufmann, editor, *Advances in Neural Information Processing Systems*, 1998.

B. Schölkopf and A.J. Smola. *Learning with kernels*. MIT Press, 2002.

L. Schwartz. *Théorie des Distributions*. Hermann, Paris, 1978.

A. Sloman. Some requirements for human-like robots: Why the recent over-emphasis on embodiment has held up progress. In B. Sendhof et al., editor, *Creating Brain-Like Intelligence*, Lecture Notes in Computer Science, vol. 5436, pages 248–277. Springer, 2009.

A.J. Smola, B. Schölkopf, and K.R. Mueller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637– 649, 1998.

R. Solomonoff. A formal theory of inductive inference - Part 1 and Part 2. *Information and Control*, pages 1–22,224–254, 1964.

E. M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-Posed Problems*. W.H. Winston, Washington, D.C., 1977.

A. Tsakonasa, G. Douniasa, J. Jantzenb, H. Axerc, B. Bjerregaard, and D. G. von Keyserlingk. Evolving rule-based systems in two medical domains using genetic programming. *Artificial Intelligence in Medicine*, 32:195–216, 2004.

B. van Brunt. *The Calculus of Variations*. Springer, 2003.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

G. Wahba. Smoothing noisy data by spline functions. *Numerische Mathematik*, 23:183–194, 1975.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

A. L. Yuille and N. M. Grzywacz. The motion coherence theory. In *Proc. 2nd Int. Conf. on Computer Vision*, pages 344–353, 1988.

A. L. Yuille and N. M. Grzywacz. A mathematical analysis of the motion coherence theory. *Int. J. of Computer Vision*, 3:155–175, 1989.