Learning with hard constraints

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

[1] DIBRIS
University of Genoa, Genova, Italy
giorgio.gnecco@unige.it, marcello.sanguineti@unige.it
[2] Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche
University of Siena, Siena, Italy
marco@dii.unisi.it, mela@dii.unisi.it

Abstract. A learning paradigm is proposed, in which one has both classical supervised examples and constraints that cannot be violated, called here "hard constraints", such as those enforcing the probabilistic normalization of a density function or imposing coherent decisions of the classifiers acting on different views of the same pattern. In contrast, supervised examples can be violated at the cost of some penalization (quantified by the choice of a suitable loss function) and so play the roles of "soft constraints". Constrained variational calculus is exploited to derive a representation theorem which provides a description of the "optimal body of the agent", i.e. the functional structure of the solution to the proposed learning problem. It is shown that the solution can be represented in terms of a set of "support constraints", thus extending the well-known notion of "support vectors".

Keywords: Learning from constraints, learning with prior knowledge, multi-task learning, support constraints, constrained variational calculus.

1 Introduction

Examples of constraints in machine learning come out naturally regardless of the context: for instance, constraints may represent prior knowledge provided by an expert (e.g., a physician in the case of a medical application: in such a case constraints may be expressed in the form of rules which help in the detection of a disease [8]). The expressive power of constraints becomes particularly significant when dealing with a specific problem, like vision, control, text classification, ranking in hyper-textual environment, and prediction of the stock market.

Table 1 provides some examples of constraints that are often encountered in practical problems arising in different domains. The first example (i) describes the simplest case in which we handle several pairs (x_i, y_i) provided for supervised learning in classification, where $y_i \in \{-1, 1\}$. If $f(\cdot)$ 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 classic "robust" sign agreement between the target and the function to be learned.

Table 1. Examples of constraints from different environments.

	linguistic description	*real-valued representation*
i.	*i*-th supervised pair for classification	$y_i \cdot f(x_i) - 1 \geq 0$
ii.	normalization of a density function	$\int_{\mathcal{X}} f(x)dx = 1$, and $\forall x \in \mathcal{X} :\ f(x) \geq 0$
iii.	coherence constraint (two classes)	$\forall x \in \mathcal{X} :\ f_1(S_1 x) \cdot f_2(S_2 x) > 0$
iv.	brightness invariance - optical flow	$\frac{\partial E}{\partial x} u + \frac{\partial E}{\partial y} v + \frac{\partial E}{\partial t} = 0$

Example *ii* is the probabilistic normalization of a density function, while example *iii* imposes the coherence between the decisions (in a binary classification problem) taken on $S_1 x$ and $S_2 x$, for the object x, where S_1 and S_2 are matrices used to select two different views of the same object x (see [9]). In the example *iv* we report a constraint from computer vision coming from the classic problem of determining the optical flow. It consists of finding the smoothest solution for the velocity field under the constraint that the brightness of any point in the movement pattern is constant. If $u(x, y, t)$ and $v(x, y, t)$ denote the components of the velocity field and $E(x, y, t)$ denotes the brightness of any pixel (x, y) at time t, then the velocity field satisfies the linear constraint indicated in Table 1 [6].

Unlike the classic framework of learning from examples, the beauty and the elegance of simplicity behind the parsimony principle - for which simple explanations are preferred to complex ones - has not been profitably used yet for the formulation of systematic theories of learning in a constrained-based environment. In those cases, 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 tasks and of providing the inductive behavior naturally required in some tasks. In the paper, instead, we propose the study of parsimonious agents interacting simultaneously with examples and constraints in a multi-task environment with the purpose of developing the simplest (smoothest) vectorial function in a set of feasible solutions.

More precisely, we think of an intelligent agent acting on a subset $\mathcal{X}$ 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 $\mathbb{R}^n$. Each function f_j is referred to as a *task of the agent*. As it is usual in supervised learning, we are also given a supervised learning set $\big\{(x_\kappa, y_\kappa),\ x_\kappa \in \mathbb{R}^d,\ y_\kappa \in \mathbb{R}^n,\ \kappa \in \mathbb{N}_{m_d}\big\}$.

In addition, we assume that additional prior knowledge is available, modeled by the exact fulfillment of constraints that are expressed either as

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

or as

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

where the sets $\mathcal{X}_i$ are open and $\phi_i, \check{\phi}_i$ are scalar-valued functions. We denote by $\mathcal{C}$ the collection of constraints (1) or (2). Following the terminology in variational calculus, we call (1) *hard bilateral holonomic constraints* and (2) *hard unilateral*

holonomic constraints. Such constraints are called *hard* since they cannot be violated; constraints that can be violated (at the cost of some penalization) play the role of *soft* constraints (e.g., usually, the ones associated with the supervised pairs of the learning set). Examples of learning problems with holonomic constraints are given, e.g., in [2,5], where the constraints arise by a suitable functional representation of prior knowledge expressed in terms of first-order-logic clauses. In the paper, we investigate theoretically the problem of learning in a constrained-based environment that takes into account at the same time both the examples and constraints of holonomic type.

The paper is organized as follows. In Section 2 we formalize the problem of learning from examples with hard constraints. A representer theorem for the solution is derived in Section 3. Section 4 is devoted to the concepts of reactions of the constraints and support constraint machines. Section 5 is a short discussion.

2 Formulation of the learning problem

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 [1]. In particular, we consider the case in which, $\forall j \in \mathbb{N}_n := \{1, \ldots, n\}$ and some positive integer k, the function $f_j : \mathcal{X} \to \mathbb{R}$ 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. So,

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

Finally, we assume $k > \frac{d}{2}$ since, by the Sobolev Embedding Theorem (see, e.g., [1, Chapter 4]), for $k > \frac{d}{2}$ each element of $\mathcal{W}^{k,2}(\mathcal{X})$ has a continuous representative, on which the constraints (1) and (2) can be evaluated unambiguously.

We can introduce a seminorm $\| f \|_{P,\gamma}$ on $\mathcal{F}$ by the pair (P, γ), where

$$P := [P_0, \ldots, P_{l-1}]'$$

is a suitable (vectorial) finite-order differential operator of order k with l components, and $\gamma \in \mathbb{R}^n$ is a fixed vector of positive components. Let

$$< Pf_j, Pf_j > = \| f_j \|_P^2 = \sum_{r=0}^{l-1} \int_{\mathcal{X}} (P_r f_j(x) P_r f_j(x)) dx \,,$$

$V(\cdot) := \frac{1}{2}(\cdot)^2$ denote the quadratic loss function, $\mu \geq 0$ be a fixed constant, and

$$\begin{aligned} \mathcal{L}_s(f) :=& \ \| f \|_{P,\gamma}^2 + \frac{\mu}{m_d} \sum_{l=1}^{m_d} \sum_{j=1}^{n} V(y_{i,j} - f_j(x)) \\ =& \sum_{j=1}^{n} \gamma_j < Pf_j, Pf_j > + \frac{\mu}{m_d} \sum_{l=1}^{m_d} \frac{1}{2} \sum_{j=1}^{n} (y_{i,j} - f_j(x))^2, \end{aligned} \tag{3}$$

the objective functional to be minimized. We state the following problem.

Problem LHC (Learning from examples with Hard Constraints). *Let $\mathcal{F}_{\mathcal{C}} \subseteq \mathcal{F}$ be the subset of functions that belong to the given functional space $\mathcal{F}$ and are compatible with a given collection $\mathcal{C}$ of hard holonomic constraints. The problem of determining a constrained (local or global) minimizer f^o of $\mathcal{L}_s$ over $\mathcal{F}_{\mathcal{C}}$ is referred to as learning from the soft constraints induced by the supervised examples and the square loss, and the hard holonomic constraint collection $\mathcal{C}$.*

So, Problem LHC is a problem of learning from examples with hard holonomic constraints. Of course, generalizations of this problem can be considered, in which other combinations of the constraints and other kinds of constraints are considered [3]. If we choose for P the form used in Tikhonov's stabilizing functionals [11], for $n = 1$ 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 nonnegative, $P_r := \sqrt{\rho_r(x)} D_r$, and D_r denotes a differential operator with constant coefficients and containing only partial derivatives of order r. An interesting case corresponds to the choice $\rho_r(x) \equiv \rho_r \geq 0$ and

$$D_{2r} = \Delta^r = \nabla^{2r} \,, \tag{4}$$

$$D_{2r+1} = \nabla \nabla^{2r} \,, \tag{5}$$

where Δ denotes the Laplacian operator and ∇ the gradient, with the additional condition $D_0 f = f$ (see [10, 12]). According to (3), when $n > 1$ the operator P acts separately on all the components of f, i.e.,

$$Pf := [Pf_1, Pf_2, \ldots, Pf_n]' \,.$$

Note that in this case we have overloaded the notation and used the symbol P for both the vector differential operator and the scalar ones that constitute its components. We focus on the case in which the operator P is invariant under spatial shift and has constant coefficients. We use the following notation. For a function u and a multiindex α with d nonnegative components α_j, we write $D^\alpha u$ to denote $\frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} u$, where $|\alpha| := \sum_{j=1}^{d} \alpha_j$. So, the generic component P_i of the operator P has the expression

$$P_i = \sum_{|\alpha| \leq k} b_{i,\alpha} D^\alpha \,,$$

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

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

Finally, we define the operator $L := (P^\star)'P$ and, using again an overloaded notation, the one $\gamma L := [\gamma_1 L, \ldots, \gamma_n L]'$.

3 The representer theorem for learning with constraints

Given a set of m holonomic constraints (defined, in general, on 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. By $\hat{\mathcal{X}}$ we denote any open subset of $\mathcal{X}$ where the same subset of constraints is defined in all its points, in such a way that $m(x)$ is constant on the same $\hat{\mathcal{X}}$. For every set $\mathcal{X}_i$, by "cl($\mathcal{X}_i$)" we denote the closure of $\mathcal{X}_i$ in the Euclidean topology. Finally, for two vector-valued functions h_1 and h_2 of the same dimension, $h_1 \otimes h_2$ denotes the vector-valued function v whose first component is the convolution of the first components of h_1 and h_2, the second component is the convolution of the second components of h_1 and h_2, and so on, i.e., $v_i = (h_1 \otimes h_2)_i = h_{1,i} \otimes h_{2,i}$, for each index i. 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 it is *inactive* in x_0 at local optimality. Recall that a free-space Green's function g associated an operator O is a solution to the distributional differential equation $Og = \delta$, where δ is the Dirac distribution, centered on the origin.

The next theorem prescribes the functional representation of a local solution to Problem LHC. It is stated for a constrained local minimizer f^o. Its proof can be obtained as a variation of the one of [3, Theorem 15].

Theorem 1 (Representer Theorem for Problem LHC). *Let us consider Problem LHC in the case of $m < n$ 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 functional space $\mathcal{F}$, where $\forall i \in \mathbb{N}_m : \phi_i \in \mathcal{C}^\infty(\mathrm{cl}(\mathcal{X}_i) \times \mathbb{R}^m)$. Let $f^o \in \mathcal{F}_C$ be any constrained local minimizer of the functional (3). 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, such that the Jacobian matrix

$$\frac{\partial(\phi_{\sigma_\phi(1)}, \ldots, \phi_{\sigma_\phi(m(x_0))})}{\partial(f^o_{\sigma_f(1)}, \ldots, f^o_{\sigma_f(m(x_0))})}, \tag{6}$$

evaluated in x_0, is not singular. Suppose also that (6) 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, in addition to the above constraints, f^o satisfies on $\hat{\mathcal{X}}$ the Euler-Lagrange equations

$$\gamma L f^o(x) + \sum_{i=1}^{m} \lambda_i(x) 1_{\mathcal{X}_i}(x) \cdot \nabla_f \phi_i(x, f^o(x)) + \frac{1}{m_d} \sum_{\kappa=1}^{m_d} (f^o(x) - y_\kappa)\delta(x - x_\kappa) = 0, \tag{7}$$

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

(ii) Let $\gamma^{-1}g := [\gamma_1^{-1}g, \ldots, \gamma_n^{-1}g]'$. *If for all* i *one has* $\mathcal{X}_i = \mathcal{X} = \mathbb{R}^d$ *and* g *is a free-space Green's function of* L, *then* f^o *has the representation*

$$f^o(\cdot) = \sum_{i=1}^{m} \gamma^{-1}g(\cdot) \vec{\otimes} \phi_i(\cdot, f^o(\cdot)) - \frac{1}{m_d} \sum_{\kappa=1}^{m_d} (f^o(x_\kappa) - y_\kappa)\gamma^{-1}g(\cdot - x_\kappa)\,, \quad (8)$$

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

(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 functional space* $\mathcal{F}$, *(i) and (ii) still hold (with every occurrence of* ϕ_i *replaced by* $\check{\phi}_i$*) if one requires the nonsingularity of the Jacobian matrix (see (6)) to hold when restricting the constraints defined in* x_0 *to the ones that are active in* x_0 *at local optimality. Moreover, each Lagrange multiplier* $\lambda_i(x)$ *is nonpositive and equal to* 0 *when the correspondent constraint is inactive in* x *at local optimality.*

4 Support constraint machines

4.1 Reactions of the constraints

The next definition formalizes a concept that plays a basic role in the following developments.

Definition 1. *The distribution* ω_i *in Theorem 1 is called the* reaction of the i-th constraint *and* $\omega := \sum_{i=1}^{m} \omega_i$ *is the overall reaction of the given constraints.*

We emphasize the fact that the reaction of a constraint is a concept associated with the constrained local minimizer f^o. In particular, two different constrained local minimizers f^o may be associated with different reactions. A similar remark holds for the overall reaction of the constraints. Loosely speaking, under the assumptions of Theorem 1, the reaction of the i-th constraint provides the way under which such constraint contributes to the expansion of f^o. For instance, under the assumptions of Theorem 1 (ii), one has the expansion

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

which emphasizes the roles of ω_i and ω in the expansion of f^o. So, solving Problem LHC is reduced to finding the reactions of the constraints.

4.2 Support constraints

Starting from the concept of the reaction of a constraint, we now introduce the following two concepts.

Definition 2. *A* support constraint *is a constraint associated with a reaction that is different from* 0 *at least in one point of the domain* $\mathcal{X}$.

Definition 3. *A* support constraint machine *is any machine capable of finding a (local or global) solution for which the representer theorem (see Theorem 1) holds.*

So, under the assumptions of the Theorem 1, a constrained local minimizer f^o for Problem LHC can be obtained by the knowledge of the reactions associated merely with the support constraints. This motivates the use of the terminology "support constraints" as an extension of the concept of "support vectors". Interestingly, support vectors are particular cases of support constraints [3]. The connection with kernel methods arise also because, under certain conditions, the free-space Green's function g associated with the operator L is a kernel of a reproducing kernel Hilbert space (see, e.g., [4] and the references therein).

The emergence of constraints whose reaction is identically 0 at local optimality (thus, of constraints that are not support constraints) is particularly evident for the case of hard unilateral constraints. For instance, under the assumptions of Theorem 1 (iii), a constraint that is inactive at local optimality for all $x \in \mathcal{X}$ is associated with a Lagrange multiplier distribution $\lambda_i(\cdot)$ 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 a problem of learning from hard constraints in which one of the constraints is redundant, in the sense that the fulfillment of the other constraints guarantees its fulfillment, too. Of course, without loss of generality, such a redundant constraint can be discarded from the problem formulation and, provided that the assumptions of Theorem 1 hold, one still has the representation (8), 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.

4.3 Computational issues

Although Problem LHC is reduced to finding the reactions of the constraints, a serious issue in the application of this recipe is that it is requires the knowledge of the Lagrange multipliers $\lambda_i(\cdot)$. Indeed, in addition to the fact that f^o has to satisfy (7), the constraints $\forall x \in \mathcal{X}, \forall i \in \mathbb{N}_m : \ \phi_i(x, f(x)) = 0$ must be verified, too. Such an implicit dependence makes it hard to solve the problem directly, unless special assumptions are satisfied (e.g., the linearity of the constraints, for which one obtains closed-form solutions [3]). The functional representation given by Theorem 1 (i) (see formula (8)) is a non-linear version of the classic functional equation known as the "Fredholm Equation of the II Kind". There are a number of theoretical and numerical studies on this equation, and particular attention has been devoted to the linear case [7]. In the general case, approximate reactions of the constraints can be obtained by discretizing the constraints, e.g., using unsupervised examples [3].

5 Discussion

In the paper, we have introduced a general theoretical framework of learning that involves agents acting in a constrained-based environment, for constraints of holonomic type. Our study focus 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. Extensions of this work to other kinds of constraints, a general theory of learning from constraints, and specific examples are discussed in [3].

Acknowledgement

This research was partially supported by the research grant PRIN2009 "Learning Techniques in Relational Domains and Their Applications" (2009LNP494) from the Italian MURST.

References

1. R. A. Adams and J. F. Fournier. *Sobolev spaces.* Academic Press, 2003.
2. M. Diligenti, M. Gori, M. Maggini, and L. Rigutini. Bridging logic and kernel machines. *Machine Learning*, 86:57–88, 2012. 10.1007/s10994-011-5243-x.
3. G. Gnecco, M. Gori, S. Melacci, and M. Sanguineti. Foundations of support constraint machines. Technical report, DII-UNISI, 2013.
4. G. Gnecco, M. Gori, and M. Sanguineti. Learning with boundary conditions. *Neural Computation*, 25:1029–1106, 2013.
5. M. Gori and S. Melacci. Constraint verification with kernel machines. *IEEE Trans. Neural Netw. Learning Syst.*, 24(5):825–831, 2013.
6. B. K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.
7. E. Kreyszig. *Introductory Functional Analysis with Applications.* Wiley & Sons, 1989.
8. G. Kunapuli, K. P. Bennett, A. Shabbeer, R. Maclin, and J. Shavlik. Online knowledge-based support vector machines. In *Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases*, pages 145–161, Berlin, Heidelberg, 2010. Springer.
9. S. Melacci, M. Maggini, and M. Gori. Semi–supervised learning with constraints for multi–view object recognition. In *Proc. 19th Int. Conf. on Artificial Neural Networks*, pages 653–662, Springer, 2009.
10. T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical report, MIT, 1989.
11. A.N. Tikhonov and V. Y. Arsenin. *Solution of ill-posed problems.* W.H. Winston, Washington, D.C., 1977.
12. A.L. Yuille and N.M. Grzywacz. A mathematical analysis of the motion coherence theory. *Int. J. of Computer Vision*, 3:155–175, 1989.