Learning with convex constraints

Marco Gori and Stefano Melacci

DII - University of Siena,
53100 - Siena, Italy
{marco,mela}@dii.unisi.it

Abstract. In this paper, we focus on multitask learning and discuss the notion of learning from constraints, in which they limit the space of admissible real values of the task functions. We formulate learning as a variational problem and analyze convex constraints, with special attention to the case of linear bilateral and unilateral constraints. Interestingly, we show that the solution is not always an analytic function and that it cannot be expressed by the classic kernel expansion on the training examples. We provide exact and approximate solutions and report experimental evidence of the improvement with respect to classic kernel machines.

Key words: kernel machines; constrained optimization; regularization.

1 Introduction

The powerful framework of regularization has been playing an enormous impact in machine learning, also in the case in which more tasks are jointly involved (see e.g. [1]). Unfortunately, the curse of dimensionality, especially in presence of many tasks, makes many complex real-world problems still hard to face. A possible direction to attach those problems is to be able to express constraints on the functional space so as to restrict the hypothesis space. Following the variational framework proposed in [2], in this paper we discuss the notion of *learning from constraints*, which limits the admissible real values of the task functions. The basic idea was proposed in [6], in which the principle of stage-based learning was also advocated. We focus on convex constraints, and, in particular, we prove that the solution is still representable as kernel expansion in the special case of linear bilateral constraints, which makes it possible to the re-use kernel-like apparatus to solve the problem. Differently, in the case of unilateral constraints, even when we simply force non-negativeness of a single function, there is no classic kernel expansion to solve the problem exactly. However, we propose a technique to approximate the solution that is based on the idea, sketched in [6], of sampling the penalty term which enforces the constraint. In addition, we suggest the adoption of a *linear soft-constraining scheme* and prove that, under an appropriate choice of the regularization parameters, we can enforce the perfect satisfaction of the non-negativeness constraint, a property that holds in general for any polytope. Finally, we present an experimental report to assess the improvement of the learning from constraints scheme with respect to classic kernel machines.

2 Learning from constraints

Given an input space X, a set of labeled samples $\mathcal{E} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) | \boldsymbol{x}_i \in X, \boldsymbol{y}_i \in \mathbb{R}^p, i = 1, \ldots, \ell\}$, and a functional space $\mathcal{F}$, we can generalize the variational formulation given in [2] to the case of multi-task learning by choosing $\boldsymbol{f} = [f_1, \ldots, f_p]$, $f_j \in \mathcal{F}$, $j = 1, \ldots, p$, such that

$$\begin{array}{l} \boldsymbol{f}(\boldsymbol{x}) = \arg\min_{\boldsymbol{f}} \left(\sum_{k=1}^{\ell} \mathcal{L}_e(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{f}(\boldsymbol{x}_k)) d\boldsymbol{x} + \frac{\lambda}{2} \int_X \parallel P\boldsymbol{f}(\boldsymbol{x}) \parallel^2 d\boldsymbol{x} \right) \\ \qquad s.t. \;\; \phi_h(\boldsymbol{x}) = \phi_h^f(\boldsymbol{x}, f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x})) \geq 0, \;\; h = 1, \ldots, q \end{array} \tag{1}$$

where $\mathcal{L}_e$ is a loss function, P is a pseudo-differential operator that is closely related to kernels [3], $\lambda > 0$ is a scalar weight, and ϕ_h, $h = 1, \ldots, q$, are constraints that model the relationships among f_j, $j = 1, \ldots, p$. An interesting case is the one of the operator $\odot_m$, selected such that $\parallel \odot_m f \parallel^2 = \sum_{r=0}^{m} \alpha_r \left(d^r f(\boldsymbol{x})\right)^2$, where the $\alpha_h \geq 0$ are constant values and the derivative operator d is a scalar operator if r is even and a vector operator is r is odd. More specifically, $d^{2r} = \Delta^r = \nabla^{2r}$ and $d^{2r+1} = \nabla\nabla^{2r}$, where Δ is the Laplacian operator and ∇ is the gradient, with the additional convention $d^o f = f$. Let us denote by $P^\star$ the adjoint of P and consider $L := P^\star P$. When $p > 1$ we can construct more general differential operators with the associated L that can give rise to cross-dependencies in multi-task learning [1], but in this paper we rely on the decoupling assumption, that is the pseudo-differential operators do not produce any cross-task effect.

The solution can be given within the Lagrangian formalism [4] that requires the satisfaction of the Euler-Lagrange (EL) equations for Eq. 1. If we indicate with $\mathcal{L}'_{e,f_j}$ the derivative of $\mathcal{L}_e$ w.r.t. f_j, and with $\delta(\cdot)$ the Dirac delta, we have

$$\lambda L f_j + \sum_{h=1}^{q} \hat{\rho}_h(\boldsymbol{x}) \frac{\partial \phi_h^f}{\partial f_j} = -\sum_{k=1}^{\ell} \delta(\boldsymbol{x} - \boldsymbol{x}_k) \mathcal{L}'_{e,f_j}(\boldsymbol{x}_k, y_{k,j}, f_j(\boldsymbol{x}_k)) \tag{2}$$

with $j = 1, \ldots, p$, that must be paired with the set of constraints $\phi_h(\boldsymbol{x}), h = 1, \ldots, q$ and with the boundary conditions on ∂X to determine the solution. Notice that, unlike for classic function optimization, the Lagrange multipliers $\hat{\rho}_h$ for variational problems with the given subsidiary conditions are functions on X (see e.g. [4], p. 46). In general, we end up in a non-linear equation for which the classic representer theorem on which kernel machines are based does not hold. If, following the spirit of statistics and machine learning, we accept a soft-fulfillment of the constraints then the Lagrangian formulation is replaced with the optimization of an index in which we add a penalty term

$$E = \sum_{i=1}^{\ell} \sum_{j=1}^{p} \mathcal{L}_e(\boldsymbol{x}_i, y_{i,j}, f_j(\boldsymbol{x}_i)) + \frac{\lambda}{2} \sum_{j=1}^{p} \int_X \parallel Pf_j \parallel^2 d\boldsymbol{x} + \sum_{h=1}^{q} \int_X \rho_h(\boldsymbol{x}) \mathcal{L}_c(\phi_h) d\boldsymbol{x} \tag{3}$$

where $\mathcal{L}_e$ and $\mathcal{L}_c$ are the loss functions related to the fitting of the examples and to the soft-fulfillment of the constraints, respectively, and $\rho_h(\boldsymbol{x})$ is an approximation of $\hat{\rho}_h(\boldsymbol{x})$. It can be shown that, in presence of convex loss functions and convex constraints also $\mathcal{F}$ becomes convex, thus simplifying dramatically the problem at hand [5].

3 Learning under linear bilateral constraints

Let us start considering the case of linear constraints in which $A\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{b}$, where $A \in I\!R^{q,p}$, $\boldsymbol{b} \in I\!R^q$, and $p > q$.

Lemma 1. *Let $L_p = diag[P^\star P]$ (the subscript on L indicates its order) and $A \in I\!R^{q,p}$ be. Then $AL_p = L_q A$.*

Proof. Straightforward consequence of linearity of L_p (see [5]).

Proposition 1. *Let $det(A \cdot A') \neq 0$ be. Then the Lagrange multipliers $\hat{\boldsymbol{\rho}}(\cdot)$ that yield the solution of 1 are:*

$$\hat{\boldsymbol{\rho}}(\boldsymbol{x}) = -[AA']^{-1} \cdot \left(\lambda \cdot \alpha_o \cdot \boldsymbol{b} + \sum_{k=1}^{\ell} A\mathcal{L}'_{e,\boldsymbol{f}}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{f}(\boldsymbol{x}_k))\delta(\boldsymbol{x} - \boldsymbol{x}_k) \right). \tag{4}$$

Proof. We start from the EL equations 2 in which we replace the general constraints ϕ_h with the linear constraint, and we get the proof (see [5]).

Theorem 1. *Let $Q := I_p - A'[AA']^{-1}A$ be, where $I_p \in I\!R^{p,p}$ is the identity matrix. For the solution of 1, under the constraints $A\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{b}$, the following kernel representation holds*

$$\boldsymbol{f}(\boldsymbol{x}) = \hat{\boldsymbol{\psi}}(\boldsymbol{x}) + \sum_{k=1}^{\ell} \boldsymbol{w}_k g(\boldsymbol{x} - \boldsymbol{x}_k), \qquad \boldsymbol{w}_k = -\frac{Q\mathcal{L}'_{e,\boldsymbol{f}}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{f}(\boldsymbol{x}_k))}{\lambda} \tag{5}$$

where $g(\cdot)$ is the Green's function of L, $\hat{\boldsymbol{\psi}}(\boldsymbol{x}) = \boldsymbol{\gamma}_c + \boldsymbol{\psi}(\boldsymbol{x})$, $\boldsymbol{\gamma}_c := A'[AA']^{-1}\boldsymbol{b}$, $\boldsymbol{\psi}(\cdot) \in Ker(L)$. Moreover, let $W = [\boldsymbol{w}_1|\ldots|\boldsymbol{w}_\ell] \in I\!R^{p,\ell}$ and $Y = [\boldsymbol{y}_1|\ldots|\boldsymbol{y}_\ell] \in I\!R^{p,\ell}$ be. In the case of the quadratic loss function[1] the solution can be obtained by solving $W(\lambda I_\ell + G) = Q(Y - \Psi - \boldsymbol{\gamma}_c \cdot \mathbf{1}_\ell{}')$, where $G \in I\!R^{\ell,\ell}$ is the Gram matrix, $\mathbf{1}_\ell$ is a vector of ℓ elements equal to 1, $\Psi = [\boldsymbol{\psi}(\boldsymbol{x}_1)|\ldots,|\boldsymbol{\psi}(\boldsymbol{x}_\ell)] \in I\!R^{p,\ell}$, and $A\boldsymbol{\psi}(\boldsymbol{x}_i) = 0$. The product by Q is not required in the case in which the examples are coherent with the constraints.

Proof. Straightforward consequence of replacing the Lagrange multipliers given by Proposition 1 into the the EL-equations (see [5]).

4 Learning under unilateral constraints

Let us consider the case of a single function ($p = 1$) along with the single inequality constraint $f(\boldsymbol{x}) \geq 0$. Assuming that $\mathcal{L}_e$ is the quadratic loss, we start by pointing out a significant difference with respect to the problem of equality linear constraints previously discussed by a simple example.

[1] We consider $\mathcal{L}_e$ scaled by $\frac{1}{2}$ when it is the quadratic loss function.

Example 1. Let us consider a learning task in which $X = [-2, +2] \subset \mathbb{R}$ *and* $\mathcal{E} = \{(-1, -1), (+1, +1)\}$. *We discuss the solution in the case* $P = \Delta$ *(see [5] for* $P = \nabla$*). We notice that the minimum for the loss function on the mismatch with respect to the training set requires that* $f(-1) = 0$ *and* $f(+1) = +1$. *Moreover, apart from eventual discontinuities, for any linear piece-wise function* $\Delta f(x) = \partial^2 f(x)/\partial x^2 = 0$ *and, therefore,* $\int_X \parallel Pf(x) \parallel^2 dx = 0$. *Finally, any non-negative linear piece-wise function such that* $f(-1) = 0$ *and* $f(+1) = +1$ *is a solution.*

The discussion of this example indicates that the solution may not be an analytic function and that the classic representation theorem of kernel machines does not hold. We can *approximate* f with a kernel expansion restricted to the a finite set, such as $\{\boldsymbol{x}_k\}_{k=1}^{\ell}$. Then the problem of Eq. 1 can be solved using Lagrange multipliers with the further assumption that they are limited to training points only, i.e. $\sum_{k=1}^{\ell} \rho_k \delta(\boldsymbol{x} - \boldsymbol{x}_k)$ (see [5]).

PENALTY-BASED SOLUTIONS

Alternatively, we can embed the unilateral constraint in the learning process with a penalty term $\rho \int_X \mathcal{L}_c(f(\boldsymbol{x}))d\boldsymbol{x}$, where $\rho < 0$. We start considering the case of a hinge-like penalty function, that is $\mathcal{L}_c(u) = 0$ if $u \geq 0$ and $\mathcal{L}_c(u) = u$ if $u < 0$. When $f(\boldsymbol{x}) < 0$, the Euler-Lagrange equations become

$$\lambda L f(\boldsymbol{x}) + \sum_{k=1}^{\ell} (f(\boldsymbol{x}_k) - y_k)\delta(\boldsymbol{x} - \boldsymbol{x}_k) + \rho = 0 \tag{6}$$

whereas if $f(\boldsymbol{x}) \geq 0$ the term ρ must be removed. Differently, in the case of a linear penalty $\mathcal{L}_c(u) = u$, which penalizes non-negative values of $f(\boldsymbol{x})$ but at the same time favors positive values, the Euler-Lagrange equations are the ones of Eq. 6, the representer theorem holds and we get

$$f(\boldsymbol{x}) = -\frac{\rho}{\alpha_o \lambda} + \sum_{k=1}^{\ell} w_k g(\boldsymbol{x} - \boldsymbol{x}_k), \qquad \boldsymbol{w} = (\lambda I_\ell + G)^{-1} \left(\boldsymbol{y} + \frac{\rho}{\alpha_o \lambda} \mathbf{1}_\ell \right). \tag{7}$$

Lemma 2. *If* $KerL = \emptyset$ *then equation 6 admits a unique solution.*

Proof. The proof can easily be given by contradiction [5].

In general the constraint $f(\boldsymbol{x}) \geq 0$ is only partially met.
Let $B := diag[\beta_1, \ldots, \beta_\ell]$ be the diagonal matrix similar to G such that $G = TBT'$, where T is orthogonal. We define $y_M := \max_k \{y_k\}$, $\beta_m = \min_k \{\beta_k\}$, and $G_M := \max_{\boldsymbol{x} \in X} \left\{ \sum_{k=1}^{\ell} g(\boldsymbol{x} - \boldsymbol{x}_k) \right\}$.

Theorem 2. *Let us assume that the following conditions hold*

$$i) \quad \zeta := \lambda - G_M \left\{ 1 + \parallel G(\lambda I + G)^{-1} \mathbf{1}_\ell \parallel_2 \right\} > 0$$

$$ii) \quad |\rho| \geq \frac{\alpha_o G_M y_M^2 \lambda^3 \ell}{(\lambda + \beta_m)^2 \zeta}$$

then for all coordinate w_k of 7 we have

$$w_k \geq \frac{\rho}{\alpha_o \lambda G_M}, \qquad \forall \boldsymbol{x} \in X : f(\boldsymbol{x}) \geq 0.$$

Proof. See [5].

5 Experimental Results

Given $\boldsymbol{f}(x) = [f_1(x), f_2(x), f_3(x)]'$, consider the bilateral constraints based on

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -3 & 5 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 2 \\ 7 \end{bmatrix}. \tag{8}$$

We artificially generated three mono-dimensional clusters of 50 labeled points each, randomly sampling three Gaussian distributions with different means and variances. Data from the same cluster share the same label. Labels are given by a supervisor and they are supposed to be perfectly coherent with the bilateral constraints (Fig. 1 top row), or noisy (Fig. 1 bottom row). We selected a Gaussian kernel of width $\sigma = 1$, and we set $\lambda = 10^{-2}$. The functions that are learned with an without enforcing the convex constraints are reported in the first three graphs of each row of Fig. 1. We also show the L_1 norm of the residual, $\|\boldsymbol{b} - A\boldsymbol{f}(x)\|_1$.

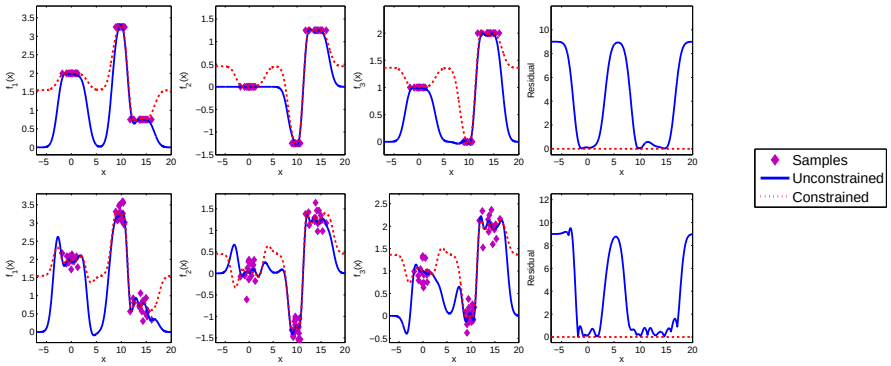

Fig. 1. $f_1(x), f_2(x), f_3(x)$ on a dataset of 150 labeled samples with and without enforcing the constraints of Eq. 8 (the L_1 norm of the residual is also shown). In the top row the labels are coherent with the constraints. In the bottom row labels are noisy.

When relying on labeled data only, the relationships of Eq. 8 are modeled only on the space regions where labeled points are distributed, and the label fitting may not be perfect due to the smoothness requirement. Differently, when constraints are introduced, the residual is zero on the entire input space.

In order to investigate the quality of the solutions proposed to enforce non-negativeness of $f(\boldsymbol{x})$, we selected a 2-class dataset with 1000 points (Fig. 2).

Classes are represented by blue dots and white crosses, and the corresponding targets are set to 0 and 1, respectively. Even if targets are non negative, $f(\boldsymbol{x})$ is not guaranteed to be strictly positive on the whole space (Fig. 2(a)). In Fig. 2(b-d), $f(\boldsymbol{x}) \geq 0$ is enforced by the procedures of Section 4. In Fig. 2(b) the function is constrained by the scheme based on Lagrange multipliers restricted to the training points, that assures $f(\boldsymbol{x}) \geq 0$ only when $\boldsymbol{x} \in \mathcal{E}$. Differently, Fig. 2(c) shows the result of the approach that linearly penalizes small values of the function (we set $\rho = -1.1$, $\lambda = 5$). Even if the positiveness of the function is fulfilled on the whole space, $f(\boldsymbol{x})$ is encouraged to assume larger values also out of the distribution of the training points. Finally, Fig. 2(d) shows a hinge loss based constraining ($\rho = 10$), that avoids penalizations of $f(\boldsymbol{x})$ where it is not needed.

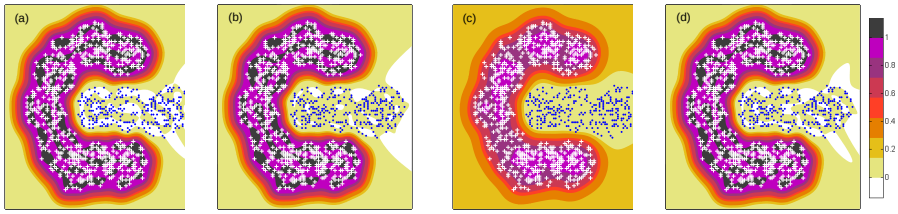

Fig. 2. A 2-class dataset (1000 samples). (a) $f(\boldsymbol{x})$ trained without any constraints - (b) when $f(\boldsymbol{x}) \geq 0$ is enforced by the Lagrange multiplier based scheme - (c) by a linear penalty - (d) by a hinge-loss penalty. $f(\boldsymbol{x})$ is negative on the white regions.

6 Conclusions

This paper contributes to the idea of extending the framework of learning from examples promoted by kernel machines to *learning from constraints*. Exact and approximate solutions in the case of convex functional spaces are proposed.

References

1. Evgeniou, T., Micchelli, C., Pontil, M.: Learning multiple tasks with kernel methods. Journal of Machine Learning Research **6** (2005) 615–637
2. Poggio, T., Girosi, F.: A theory of networks for approximation and learning. Technical report, MIT (1989)
3. Smola, A., Schoelkopf, B., Mueller, K.: The connection between regularization operators and support vector kernels. Neural Networks **11** (1998) 637– 649
4. Gelfand, I., Fomin, S.: Calculus of Variations. Dover publications, Inc (1963)
5. Gori, M., Melacci, S.: Learning with convex constraints. Technical report, Department of Information Engineering - University of Siena (2010)
6. Gori, M.: Semantic-based regularization and Piaget's cognitive stages. Neural Networks, vo. 22, no. 7, 1035-1036 (2009)