

MIT Open Access Articles

Convex learning of multiple tasks and their structure

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Ciliberto, Carlo et al. "Convex learning of multiple tasks and their structure." Journal of Machine Learning Research, Proceedings of the 32nd International Conference on Machine Learning, July 7-9 2015, Lille, France, MIT Press, July 2015 Copyright © 2015 The Author(s)

As Published: <https://dl.acm.org/citation.cfm?id=3045283>

Publisher: MIT Press

Persistent URL: <http://hdl.handle.net/1721.1/112313>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



Convex Learning of Multiple Tasks and their Structure

Carlo Ciliberto ^{*†} Youssef Mroueh ^{*†} Tomaso Poggio ^{*†}
 Lorenzo Rosasco ^{*†‡}

Abstract

Reducing the amount of human supervision is a key problem in machine learning and a natural approach is that of exploiting the relations (structure) among different tasks. This is the idea at the core of multi-task learning. In this context a fundamental question is how to incorporate the tasks structure in the learning problem. We tackle this question by studying a general computational framework that allows to encode a-priori knowledge of the tasks structure in the form of a convex penalty; in this setting a variety of previously proposed methods can be recovered as special cases, including linear and non-linear approaches. Within this framework, we show that tasks and their structure can be efficiently learned considering a convex optimization problem that can be approached by means of block coordinate methods such as alternating minimization and for which we prove convergence to the global minimum.

1 Introduction

Current machine learning systems achieve remarkable results in several challenging tasks, but are limited by the amount of human supervision required. Leveraging similarity among different problems is widely acknowledged to be a key approach to reduce the need for supervised data. Indeed, this idea is at the basis of multi-task learning, where the joint solution of different problems (tasks) has the potential to exploit tasks relatedness (structure) to improve learning accuracy. This idea has motivated a variety of methods, including frequentist [25, 3, 4] and Bayesian methods (see e.g. [1] and references therein), with connections to structured learning [6, 34].

The focus of our study is the development of a general regularization framework to learn multiple tasks as well as their structure. Following [25, 15] we consider a setting where tasks are modeled as the components of a vector-valued function and their

^{*}Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia, Via Morego, 30, 16100, Genova, Italy, (cciliber@mit.edu)

[†]Poggio Lab, Massachusetts Institute of Technology, Massachusetts Ave. 77, 02139, Cambridge, MA, USA (tp@ai.mit.edu)

[‡]DIBRIS, Università di Genova, Via Dodecaneso, 35, 16146, Genova, Italy, (lrosasco@mit.edu)

structure corresponds to the choice of suitable functional spaces. Exploiting the theory of reproducing kernel Hilbert spaces for vector-valued functions (RKHSvv) [25], we consider and analyze a flexible regularization framework, within which a variety of previously proposed approaches can be recovered as special cases, see e.g. [19, 24, 26, 37, 14, 31]. Our main technical contribution is a unifying study of the minimization problem corresponding to such a regularization framework. More precisely, we devise an optimization approach that can efficiently compute a solution and for which we prove convergence under weak assumptions. Our approach is based on a barrier method that is combined with block coordinate descent techniques [33, 30]. In this sense our analysis generalizes the results in [3] for which a low-rank assumption was considered; however the extension is not straightforward, since we consider a much larger class of regularization schemes (any convex penalty). Up to our knowledge, this is the first result in multi-task learning proving the convergence of alternating minimization schemes for such a general family of problems.

The RKHSvv setting allows to naturally deal both with linear and non-linear models and the approach we propose provides a general computational framework for learning output kernels as formalized in [14].

The rest of the paper is organized as follows: in Sec 2 we review basic ideas of regularization in RKHSvv. In Sec. 2.3 we discuss the equivalence of different approaches to encode known structures among multiple tasks. In Sec. 3 we discuss a general framework for learning multiple tasks and their relations where we consider a wide family of structure-inducing penalties and study an optimization strategy to solve them. This setting allows us, in Sec. 4, to recover several previous methods as special cases. Finally in Sec. 5 we evaluate the performance of the optimization method proposed.

Notation. With $S_{++}^n \subset S_+^n \subset S^n \subset \mathbb{R}^{n \times n}$ we denote respectively the space of positive definite, positive semidefinite (PSD) and symmetric $n \times n$ real-valued matrices. O^n denotes the space of orthonormal $n \times n$ matrices. For any square matrix $M \in \mathbb{R}^{n \times n}$ and $p \geq 1$, we denote by $\|M\|_p = (\sum_{i=1}^n \sigma_i(M)^p)^{1/p}$ the p -Schatten norm of M , where $\sigma_i(M)$ is the i -th largest singular value of M . For any $M \in \mathbb{R}^{n \times m}$, M^\top denotes the transpose of M . For any PSD matrix $A \in S_+^n$, A^\dagger denotes the pseudoinverse of A . We denote by $I_n \in S_{++}^n$ the $n \times n$ identity matrix. The notation $\text{Ran}(M) \subseteq \mathbb{R}^m$ identifies the range of columns of a matrix $M \in \mathbb{R}^{m \times n}$.

2 Background

We study the problem of jointly learning multiple tasks by modeling individual task-predictors as the components of a vector-valued function. Let us assume to have T supervised scalar learning problems (or tasks), each with a “training” set of input-output observations $\mathcal{S}_t = \{(x_{it}, y_{it})\}_{i=1}^{n_t}$ with $x_{it} \in \mathcal{X}$ input space and $y_{it} \in \mathcal{Y}$ output space¹. Given a loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ that measures the per-task prediction

¹To avoid clutter in the notation, we have restricted ourselves to the typical situation where all tasks share same input and output spaces, i.e. $\mathcal{X}_t = \mathcal{X}$ and $\mathcal{Y}_t \subseteq \mathbb{R}$.

errors, we want to solve the following joint regularized learning problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \sum_{t=1}^T \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_i^{(t)}, f_t(x_i^{(t)})) + \lambda \|f\|_{\mathcal{H}}^2 \quad (1)$$

where \mathcal{H} is an Hilbert space of vector-valued functions $f : \mathcal{X} \rightarrow \mathcal{Y}^T$ with scalar components $f_t : \mathcal{X} \rightarrow \mathcal{Y}$. In order to define a suitable space of hypotheses \mathcal{H} , in this section we briefly recall concepts from the theory of reproducing kernel Hilbert spaces for vector-valued functions (RKHSvv) and corresponding regularization theory, which plays a key role in our work. In particular, we focus on a class of reproducing kernels (known as separable kernels) that can be designed to encode specific tasks structures (see [15, 2] and Sec. 2.3). Interestingly, separable kernels are related to ideas such as defining a metric on the output space or a label encoding in multi-label problems (see Sec. 2.3)

Remark 2.1 (Multi-task and multi-label learning). Multi-label learning is a class of supervised learning problems in which the goal is to associate input examples with a label or a set of labels chosen from a discrete set. In general, due to discrete nature of the output space, these problems cannot be solved directly; hence, a so-called *surrogate* problem is often introduced, which is computationally tractable and whose solution allows to recover the solution of the original problem [32, 7, 28].

Multi-label learning and multi-task learning are strongly related. Indeed, surrogate problems typically consist in a set of distinct supervised learning problems (or tasks) that are solved simultaneously and therefore have a natural formulation in the multi-task setting. For instance, in multi-class classification problems the “One vs All” strategy is often adopted, which consists in solving a set of multiple binary classification problems, one for each class.

2.1 Learning Multiple Tasks with RKHSvv

In the scalar setting, reproducing kernel Hilbert spaces have already been proved to be a powerful tool for machine learning applications. Interestingly, the theory of RKHSvv and corresponding Tikhonov regularization scheme follow closely the derivation in the scalar case.

Definition 2.2. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space of functions from \mathcal{X} to \mathbb{R}^T . A symmetric, positive definite, matrix-valued function $\Gamma : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{T \times T}$ is called a reproducing kernel for \mathcal{H} if for all $x \in \mathcal{X}, c \in \mathbb{R}^T$ and $f \in \mathcal{H}$ we have that $\Gamma(x, \cdot)c \in \mathcal{H}$ and the following reproducing property holds $\langle f(x), c \rangle_{\mathbb{R}^T} = \langle f, \Gamma(x, \cdot)c \rangle_{\mathcal{H}}$.

In analogy to the scalar setting, it can be proved (see [25]) that the Representer Theorem holds also for regularization in RKHSvv. In particular we have that any solution of the learning problem introduced in Eq. (1) can be written in the form

$$f(x) = \sum_{t=1}^T \sum_{i=1}^{n_t} \Gamma(x, x_i^{(t)}) c_i^{(t)} \quad (2)$$

with $c_i^{(t)} \in \mathbb{R}^T$ coefficient vectors.

The choice of kernel Γ induces a joint representation of the inputs as well as a structure among the output components [1]; In the rest of the paper we will focus on so-called separable kernels, where these two aspects are factorized. In Section 3, we will see how separable kernels provide a natural way to learn the tasks structure as well as the tasks.

2.2 Separable Kernels

Separable (reproducing) kernels are functions of the form $\Gamma(x, x') = k(x, x')A \forall x, x' \in \mathcal{X}$ where $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a scalar reproducing kernel and $A \in S_+^T$ is a positive semi-definite (PSD) matrix. In this case, the representer theorem allows to rewrite problem (1) in a more compact matrix notation as

$$\underset{C \in \mathbb{R}^{n \times T}}{\text{minimize}} \quad V(Y, KCA) + \lambda \text{tr}(AC^\top KC). \quad (\mathcal{P})$$

Here $Y \in \mathbb{R}^{n \times T}$ is a matrix with $n = \sum_{t=1}^T n_t$ rows containing the output points; $K \in S_+^n$ is the empirical kernel matrix associated to k and $V : \mathbb{R}^{n \times T} \times \mathbb{R}^{n \times T} \rightarrow \mathbb{R}_+$ generalizes the loss in (1) and consists in a linear combination of the entry-wise application of \mathcal{L} . Notice that this formulation accounts also the situation where not all training outputs $y^{(t)}$ are observed when a given input $x \in \mathcal{X}$ is provided: in this case the functional V weights 0 the loss values of those entries of Y (and the associated entries of KCA) that are not available in training.

Finally, the second term in (\mathcal{P}) follows by observing that, for all $f \in \mathcal{H}$ of the form $f(\cdot) = \sum_{i=1}^n k(x_i, \cdot)Ac_i$, the squared norm can be written as $\|f\|_{\mathcal{H}}^2 = \sum_{i,j} k(x_i, x_j)c_i^\top Ac_j = \text{tr}(AC^\top KC)$ where $C \in \mathbb{R}^{n \times T}$ is the matrix with i -th row corresponding to the coefficient vector $c_i \in \mathbb{R}^T$ of f . Notice that we have re-ordered the index i to be in $\{1, \dots, n\}$ to ease the notation.

2.3 Incorporating Known Tasks Structure

Separable kernels provide a natural way to incorporate the task structure when the latter is known a priori. This strategy is quite general and indeed in the following we comment on how the matrix A can be chosen to recover several multi-task methods previously proposed in contexts such as regularization, coding/embeddings or output metric learning, postponing a more detailed discussion in the supplementary material. These observations motivate the extension in Sec. 3 of the learning problem (\mathcal{P}) to a setting where it is possible to infer A from the data.

Regularizers. Tasks relations can be enforced by devising suitable regularizers [15]. Interestingly, for a large class of such methods it can be shown that this is equivalent to the choice of the matrix A (or rather its pseudoinverse) [25]. If we consider the squared norm of a function $f = \sum_{i=1}^n k(x_i, \cdot)Ac_i \in \mathcal{H}$ we have (see [15])

$$\|f\|_{\mathcal{H}}^2 = \sum_{t,s=1}^T A_{ts}^\dagger \langle f_t, f_s \rangle_{\mathcal{H}_k} \quad (3)$$

where A_t is the t -th column of A , \mathcal{H}_k is the RKHS associated to the scalar kernel k and $f_t = \sum_{i=1}^n k(x_i, \cdot) A_t^\top c_i \in \mathcal{H}_k$ is the t -th component of f . The above equation suggests to interpret A^\dagger as the matrix that models the structural relations between tasks by directly coupling different predictors. For instance, by setting $A^\dagger = I_T + \gamma(\mathbf{1}\mathbf{1}^\top)/T$, with $\mathbf{1} \in \mathbb{R}^T$ the vector of all 1s, we have that the parameter γ controls the variance $\sum_{t=1}^T \|\bar{f} - f_t\|_{\mathcal{H}_k}^2$ of the tasks with respect to their mean $\bar{f} = \frac{1}{T} \sum_{t=1}^T f_t$. If we have access to some notion of similarity among tasks in the form of a graph with adjacency matrix $W \in S^T$, we can consider the regularizer $\sum_{t,s=1}^T W_{t,s} \|f_t - f_s\|_{\mathcal{H}_k}^2 + \gamma \sum_{t=1}^T \|f_t\|_{\mathcal{H}_k}^2$ which corresponds to $A^\dagger = L + \gamma I_T$ with L the graph Laplacian induced by W .

Output Metric. A different approach to model tasks relatedness consists in choosing a suitable metric on the output space to reflect the tasks structure [24]. Clearly a change of metric on the output space with the standard inner product $\langle y, y' \rangle_{\mathbb{R}^T}$ between two output points $y, y' \in \mathcal{Y}^T$ corresponds to the choice of a different inner product $\langle y, y' \rangle_\Theta = \langle y, \theta y' \rangle_{\mathbb{R}^T}$ for some positive definite matrix $\Theta \in S_{++}^T$. Indeed this can be directly related to the choice of a suitable separable kernel. In particular, for the least squares loss function a direct equivalence holds between choosing a metric deformation associated to a $\Theta \in S_{++}^T$ and a separable kernel $k(\cdot, \cdot)_{I_T}$ or use the canonical metric (i.e. with $\Theta = I_T$ the identity) and kernel $k(\cdot, \cdot)_\Theta$. The details of this equivalence can be found in the supplementary material.

Output Representation. The tasks structure can also be modeled by designing an ad-hoc embedding for the output space. This approach is particularly useful for multi-label scenarios, where output embedding can be designed to encode complex structures such as (e.g. trees, strings, graphs, etc.) [17, 21, 11]. Interestingly in these cases, or more generally whenever the embedding map $L : \mathcal{Y}^T \rightarrow \tilde{\mathcal{Y}}$, from the original to the new output space, is linear, then it is possible to show that the learning problem with new code is equivalent to (1) for a suitable choice of separable kernel with $A = L^\top L$. We refer again to the supplementary material for the details of this equivalence.

3 Learning the Tasks and their Structure

Clearly, an interesting setting occurs when knowledge of the tasks structure is not available and therefore it is not possible to design a suitable separable kernel. In this case a favorable approach is to infer the tasks relations directly from the data. To this end we propose to consider the following extension of problem (P)

$$\underset{C \in \mathbb{R}^{n \times T}, A \in S_+^T}{\text{minimize}} \quad V(Y, KCA) + \lambda \text{tr}(AC^\top KC) + F(A), \quad (\mathcal{Q})$$

where the penalty $F : S_+^T \rightarrow \mathbb{R}_+$ is designed to learn specific tasks structures encoded in the matrix A . The above regularization is general enough to encompass a large number of previously proposed approaches by simply specifying a choice of the scalar kernel and the penalty F . A detailed discussion of these connections is postponed to Section 4. In this section, we focus on computational aspects. Throughout,

we restrict ourselves to convex loss functions V and convex (and coercive) penalties F . In this case, the objective function in (Q) is separately convex in C and A but not jointly convex. Hence, block coordinate methods, which are often used in practice, e.g. alternating minimization over C and A , are not guaranteed to converge to a global minimum. Our study provides a general framework to provably compute a solution to problem (Q) . First, In Section 3.1, we prove our main results providing a characterization of the solutions of Problem (Q) and studying a barrier method to cast their computation as a convex optimization problem. Second, in Section 3.2, we discuss how block coordinate methods can be naturally used to solve such a problem, analyze their convergence properties and discuss some general cases of interest.

3.1 Characterization of Minima and A Barrier Method

We begin, in Section 3.1.1, providing a characterization of the solutions to Problem (Q) by showing that it has an equivalent formulation in terms of the minimization of a convex objective function, namely Problem (R) . Depending on the behavior of the objective function on the boundary of the optimization domain, Problem (R) might not be solved using standard optimization techniques. This possible issue motivates the introduction, in Section 3.1.2, of a barrier method; a family of “perturbated” convex programs is introduced whose solutions are shown to converge to those of Problem (R) (and hence of the original (Q)).

3.1.1 An Equivalent formulation for (Q)

The objective functional in (Q) is not convex, therefore in principle it is hard to find a global minimizer. As it turns out however, it is possible to circumvent this issue and efficiently find a global solution to (Q) . The following result represents a first step in this direction.

Theorem 3.1. *Let $K \in S_+^n$ and consider the convex set*

$$\mathcal{C} = \{(C, A) \in \mathbb{R}^{n \times T} \times S_+^T \mid \text{Ran}(C^\top K C) \subseteq \text{Ran}(A)\}.$$

Then, for any $F : S_+^T \rightarrow \mathbb{R}_+$ convex and coercive, problem

$$\underset{(C,A) \in \mathcal{C}}{\text{minimize}} V(Y, KC) + \lambda \text{tr}(A^\dagger C^\top K C) + F(A) \quad (\mathcal{R})$$

has convex objective function and it is equivalent to (Q) . In particular, the two problems achieve the same minimum value and, given a solution (C_R, A_R) for (\mathcal{R}) , the couple $(C_R A_R^\dagger, A_R)$ is a minimizer for (Q) . Vice-versa, given a solution (C_Q, A_Q) for (Q) , the couple $(C_Q A_Q, A_Q)$ is a minimizer for (\mathcal{R}) .

The above result highlights a remarkable connection between the problems (Q) (non-convex) and (\mathcal{R}) (convex). In particular, we have the following Corollary, which provides us with a useful characterization of the local minimizers of problem (Q) .

Corollary 3.2. *Let $Q : \mathbb{R}^{n \times T} \times S_+^T \rightarrow \mathbb{R}$ be the objective function of problem (Q) . Then, every local minimizer for Q on the open set $\mathbb{R}^{n \times T} \times S_{++}^T$ is also a global minimizer.*

Corollary 3.2 follows from Theorem 3.1 and the fact that, on the restricted domain $\mathbb{R}^{n \times T} \times S_{++}^T$, the map Q is the combination of the objective functional of (\mathcal{R}) and the invertible function $(C, A) \mapsto (CA, A)$. Moreover, if Q is differentiable, i.e. V and the penalty F are differentiable, this is exactly the definition of a *convexifiable* function, which in particular implies *invexity* [12]. The latter property ensures that, in the differentiable case, all the *stationary* points (rather than only local minimizers) are global minimizers. This result was originally proved in [14] for the special case of V the least-squares loss and $F(\cdot) = \|\cdot\|_F^2$ the Frobenius norm; Here we have proved its generalization to all convex losses V and penalties F .

We end this section adding two comments. First, we note that, while the objective function in Problem (\mathcal{R}) is convex, the corresponding minimization problem might not be a convex program (in the sense that the feasible set \mathcal{C} is not identified by a set of linear equalities and non-linear convex inequalities [9]). Second, Corollary (3.2) holds only on the interior of the minimization domain $\mathbb{R}^{n \times T} \times S_+^T$ and does not characterize the behavior of the target functional on its boundary. In fact, one can see that both issues can be tackled defining a *perturbed* objective functional having a suitable behavior on the boundary of the minimization domain. This is the key motivation for the barrier method we discuss in the next section.

3.1.2 A Barrier Method to Optimize (\mathcal{R})

Here we propose a barrier approach inspired by the work in [3] by introducing a perturbation of problem (\mathcal{R}) that enforces the objective functions to be equal to $+\infty$ on the boundary of $\mathbb{R}^{n \times T} \times S_+^T$. As a consequence, each perturbed problem can be solved as a convex optimization constrained on a closed cone. The latter comment is made more precise in the following result that we prove in the supplementary material.

Theorem 3.3. *Consider the family of optimization problems*

$$\begin{aligned} & \underset{\substack{C \in \mathbb{R}^{n \times T}, \\ A \in S_+^T}}{\text{minimize}} V(Y, KC) + \lambda \text{tr}(A^{-1}(C^\top KC + \delta^2 I_T)) + F(A) \end{aligned} \quad (\mathcal{S}^\delta)$$

with $I_T \in S_{++}^T$ the identity matrix. Then, for each $\delta > 0$ the problem (\mathcal{S}^δ) admits a minimum. Furthermore, the set of minimizers for (\mathcal{S}^δ) converges to the set of minimizers for (\mathcal{R}) as δ tends to zero. More precisely, given any sequence $\delta_m > 0$ such that $\delta_m \rightarrow 0$ and a sequence of minimizers $(C_m, A_m) \in \mathbb{R}^{n \times T} \times S_+^T$ for (\mathcal{S}^δ) , there exists a sequence $(C_m^*, A_m^*) \in \mathbb{R}^{n \times T} \times S_+^T$ of minimizers for (\mathcal{R}) such that $\|C_m - C_m^*\|_F + \|A_m - A_m^*\|_F \rightarrow 0$ as $m \rightarrow +\infty$.

The barrier $\delta^2 \text{tr}(A^{-1})$ is fairly natural and can be seen as preconditioning of the problem leading to favorable computations. The proposed barrier method is similar in spirit to the approach developed in [3] and indeed Theorem 3.3 and next Corollary 3.4 are a generalization over the two main results in [3] to any convex penalty F on the cone of PSD matrices. However, notice that since we are considering a much wider family of penalties (than the trace norm as in [3]) our results cannot directly derived from those in [3]. In the next section we discuss how to compute the solution of Problem (\mathcal{S}^δ) considering a block coordinate approach.

Algorithm 1 CONVEX MULTI-TASK LEARNING

Input: K, Y, ϵ tolerance, δ perturbation parameter, S objective functional of (S^δ) , V loss, F structure penalty.

Initialize: $(C, A) = (C_0, A_0), t = 0$

repeat

$C_{t+1} \leftarrow \text{SUPERVISEDSTEP}(V, K, Y, C_t, A_t)$

$A_{t+1} \leftarrow \text{UNSUPERVISEDSTEP}(F, K, \delta, C_{t+1}, A_t)$

$t \leftarrow t + 1$

until $|S(C_{t+1}, A_{t+1}) - S(C_t, A_t)| < \epsilon$

3.2 Block Coordinate Descent Methods

The characteristic block variable structure of the objective function in problem (S^δ) , suggests that it might be beneficial to use block coordinate methods (BCM) (see [8]) to solve it. Here with BCM we identify a large class of methods that, in our setting, iterate steps of an optimization on C , with A fixed, followed by an optimization of A , for C fixed.

A *meta* block coordinate algorithm to solve (S^δ) is reported in in Algorithm 1. Here we interpret each optimization step over C as a supervised step, and each optimization step over A as an unsupervised step (in the sense that it involves the inputs but not the outputs). Indeed, when the structure matrix A is fixed, problem (\mathcal{R}) boils down to the standard supervised multi-task learning frameworks where a priori knowledge regarding the tasks structure is available. Instead, when the coefficient matrix C is fixed, the problem of learning A can be interpreted as an unsupervised setting in which the goal is to actually find the underlying task structure [23].

Several optimization methods can be used as procedures for both SUPERVISEDSTEP and UNSUPERVISEDSTEP in Algorithm 1. In particular, a first class of methods is called Block Coordinate Descent (BCD) and identifies a wide class of iterative methods that perform (typically inexact) minimization of the objective function one block of variables at the time. Different strategies to choose which direction minimize at each step have been proposed: pre-fixed cyclic order, greedy search [30] or randomly, according to a predetermined distribution [29]. For a review of several BCD algorithms we refer the reader to [30] and references therein.

A second class of methods is called alternating minimization and corresponds to the situation where at each step in Algorithm 1 and exact minimization is performed. This latter approach is favorable when a closed form solution exists for at least one block of variables (see Section 3.2.1) and has been studied extensively in [33] in the abstract setting where an oracle provides a block-wise minimizer at each iteration. The following Corollary describes the convergence properties of BCD and Alternate minimization sequences provided by applying Algorithm 1 to (S^δ) .

Corollary 3.4. *Let the Problem (S^δ) be defined as in Theorem 3.3 then:*

- (a) **Alternating Minimization:** *Let the two procedures in Algorithm 1 each provide a block-wise minimizer of the functional with the other block held fixed. Then every limiting point of a minimization sequence provided by Algorithm 1, is a*

global minimizer for (S^δ) .

- (b) **Block Coordinate Descent:** Let the two procedures in Algorithm 1 each consist in a single step of a first order optimization method (e.g. Projected Gradient Descent, Proximal methods, etc.). Then every limiting point of a minimizing sequence provided by Algorithm 1 is a global minimizer for (S^δ) .

Corollary (3.4) follows by applying previous results on BCD and Alternate minimization. In particular, for the proof of part (a) we refer to Theorem 4.1 in [33], while for part (b) we refer to Theorem 2 in [30].

In the following we discuss the actual implementation of both SUPERVISED and UNSUPERVISED procedures in the case where V is chosen to be least-squares loss and the penalty F to be a spectral p -Schatten norm. This should provide the reader with a practical example of how the meta-algorithm introduced in this section can be specialized to a specific multi-task learning setting.

Remark 3.5. (Convergence of Block Coordinate Methods) Several works in multi-task learning have proposed some form of BCM strategy to solve the learning problem. However, up to our knowledge, so far only the authors in [3] have considered the issue of convergence to a global optimum. Their results were proved for a specific choice of structure penalty in a framework similar to that of problem (\mathcal{R}) (see Section 4) but do not extend straightforwardly to other settings. Corollary 3.4 aims to fill this gap, providing convergence guarantees for block coordinate methods for a large class of multi-task learning problems.

3.2.1 Closed Form solutions for Alternating Minimization: Examples

Here we focus on the alternating minimization case and discuss some settings in which it is possible to obtain a closed form solution for the procedures SUPERVISEDSTEP and UNSUPERVISEDSTEP.

(SUPERVISEDSTEP) Least Square Loss. When the loss function V is chosen to be least squares (i.e. $V(Y, Z) = \|Y - Z\|_F^2$ for any two matrices $Y, Z \in \mathbb{R}^{n \times m}$) and the structure matrix A is fixed, a closed form solution for the coefficient matrix C returned by the SUPERVISEDSTEP procedure can be easily derived (see for instance [1]):

$$\text{vec}(C) = (I_T \otimes K + \lambda A^{-1} \otimes I_n)^{-1} \text{vec}(Y).$$

Here, the symbol \otimes denotes the Kronecker product, while the notation $\text{vec}(M) \in \mathbb{R}^{nm}$ for a matrix $M \in \mathbb{R}^{n \times m}$ identifies the concatenation of its columns in a single vector. In [26] the authors proposed a faster approach to solve this problem in closed form based on Sylvester's method.

(UNSUPERVISEDSTEP) p -Schatten penalties. We consider the case in which F is chosen to be a spectral penalty of the form $F(\cdot) = \|\cdot\|_p^p$ with $p \geq 1$. Also in this setting the optimization problem has a closed form solution, as shown in the following.

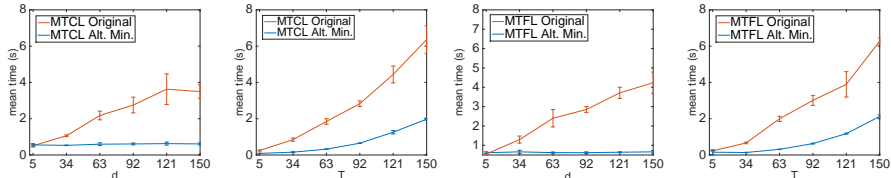


Figure 1: Comparison of the computational performance of the alternating minimization strategy studied in this paper with respect to the optimization methods proposed for MTCL in [19] and MTFM [3] in the original papers. Experiments are repeated for different number of tasks and input-space dimensions as described in Sec. 5.1.

Proposition 3.6. *Let the penalty of problem (S^δ) be $F = \|\cdot\|_p^p$ with $p \geq 1$. Then, for any $C \in \mathbb{R}^{n \times T}$ fixed, the optimization problem (S^δ) in the block variable A has a minimizer of the form*

$$A_C^\delta = \sqrt[p+1]{(C^\top K C + \delta^2 I_T)/\lambda}. \quad (4)$$

Proposition 3.6 generalizes a similar result originally proved in in [3] for the special case $p = 1$ and provides an explicit formula for the UNSUPERVISEDSTEP of Algorithm 1. We report the proof in the supplementary material.

4 Previous Work: Comparison and Discussion

The framework introduced in problem (Q) is quite general and accounts for several choices of loss function and task-structural priors. Section 3 has been mainly devoted to derive efficient and generic optimization procedures; in this section we focus our analysis on the modeling aspects, investigating the impact of different structure penalties on the multi-task learning problem. In particular, we will briefly review some multi-task learning method previously proposed, discussing how they can be formulated as special cases of problem (Q) (or, equivalently, (R)).

Spectral Penalties. The penalty $F = \|\cdot\|_F^2$ was considered in [14], together with a least squares loss function and the non convex problem (Q) is solved directly by alternating minimization. However, as pointed out in Sec. 3, solving the non convex problem (although invex, see the discussion on Corollary 3.2) directly could in principle become problematic when the alternating minimization sequence gets close to the boundary of $\mathbb{R}^{n \times T} \times S_{++}^T$. A related idea is that of considering $F(A) = \text{tr}(A)$ (i.e. the 1-Schatten norm). This latter approach can shown to be equivalent to the Multi-Task Feature Learning setting of [3] (see supplementary material).

Cluster Tasks Learning. In [19], the authors studied a multi-task setting where tasks are assumed to be organized in a fixed number r of unknown disjoint clusters. While

the original formulation was conceived for linear setting, it can be easily extended to non-linear kernels and cast in our framework. Let $E \in \{0, 1\}^{T \times r}$ be the binary matrix whose entry E_{st} has value 1 or 0 depending on whether task s is in cluster t or not. Set $M = I - E^\dagger E^\top$, and $U = \frac{1}{T} \mathbf{1}\mathbf{1}^\top$. In [19] the authors considered a regularization setting of the form of (\mathcal{R}) where the structure matrix A is parametrized by the matrix M in order to reflect the cluster structure of the tasks. More precisely:

$$A^{-1}(M) = \epsilon_M U + \epsilon_B (M - U) + \epsilon_W (I - M)$$

where the first term characterizes a global penalty on the average of all tasks predictors, the second term penalizes the between-clusters variance, and the third term controls the tasks variance within each cluster. Clearly, it would be ideal to identify an optimal matrix $A(M)$ minimizing problem (\mathcal{R}) . However, M belongs to a discrete non convex set, therefore authors propose a convex relaxation by constraining M to be in a convex set $\mathcal{S}_c = \{M \in S_+^T, 0 \preceq M \preceq I, \text{tr}(M) = r\}$. In our notations $F(A)$ is therefore the indicator function over the set of all matrices $A = A(M)$ such that $M \in \mathcal{S}_c$. The authors propose a pseudo gradient descent method to solve the problem jointly.

Convex Multi-task Relation Learning. Starting from a multi-task Gaussian Process setting, in [37], authors propose a model where the covariance among the coefficient vectors of the T individual tasks is controlled by a matrix $A \in S_{++}^T$ in the form of a prior. The initial maximum likelihood estimation problem is relaxed to a convex optimization with target functional of the form

$$\|Y - KC\|_F^2 + \lambda_1 \text{tr}(C^\top KC) + \lambda_2 \text{tr}(A^{-1}C^\top KC) \quad (5)$$

constrained to the set $\mathcal{A} = \{A \mid A \in S_{++}^T, \text{tr}(A) = 1\}$. This setting is equivalent to problem (\mathcal{R}) (by choosing F to be the indicator function of \mathcal{A}) with the addition of the term $\text{tr}(C^\top KC)$.

Non-Convex Penalties. Often times, interesting structural assumptions cannot be cast in a convex form and indeed several works have proposed non-convex penalties to recover interpretable relations among multiple tasks. For instance [2] requires A to be a graph Laplacian, or [13] imposes a low-rank factorization of A in two smaller matrices. In [27, 22] different sparsity models are proposed.

Interestingly, most of these methods can be naturally cast in the form of problem (\mathcal{Q}) or (\mathcal{R}) . Unfortunately our analysis of the barrier method does not necessarily hold also in these settings and therefore Alternating Minimization is not guaranteed to lead to a stationary point.

5 Experiments

We empirically evaluated the efficacy of the block coordinate optimization strategy proposed in this paper on both artificial and real datasets. Synthetic experiments were performed to assess the computational aspects of the approach, while we evaluated the quality of solutions found by the system on realistic settings.

	50 tr. samples per class		100 tr. samples per class		150 tr. samples per class		200 tr. samples per class	
	nMSE (\pm std)	nI	nMSE (\pm std)	nI	nMSE (\pm std)	nI	nMSE (\pm std)	nI
STL	0.2436 \pm 0.0268	0	0.1723 \pm 0.0116	0	0.1483 \pm 0.0077	0	0.1312 \pm 0.0021	0
MTFL	0.2333 \pm 0.0213	0.0416	0.1658 \pm 0.0107	0.0379	0.1428 \pm 0.0083	0.0281	0.1311 \pm 0.0055	0.0003
MTRL	0.2314 \pm 0.0217	0.0404	0.1653 \pm 0.0112	0.0401	0.1421 \pm 0.0081	0.0288	0.1303 \pm 0.0058	0.0071
OKL	0.2284 \pm 0.0232	0.0630	0.1604 \pm 0.0123	0.0641	0.1410 \pm 0.0087	0.0350	0.1301 \pm 0.0073	0.0087

Table 1: Comparison of Multi-task learning methods on the Sarcos dataset. The advantage of learning the tasks jointly decreases as more training examples became available.

5.1 Computational Times

As discussed in Sec. 4, several methods previously proposed in the literature, such as Multi-task Cluster Learning (MTCL) [19] and Multi-task Feature Learning (MTFL [3]), can be formulated as special cases of problem (\mathcal{Q}) or (\mathcal{R}). It is natural to compare the proposed alternating minimization strategy with the optimization solution originally proposed for each method. To assess the system’s performance with respect to varying dimensions of the feature space and an increasing number of tasks, we chose to perform this comparison in an artificial setting.

We considered a linear setting where the input data lie in \mathbb{R}^d and are distributed according to a normal distribution with zero mean and identity covariance matrix. T linear models $w_t \in \mathbb{R}^d$ for $t = 1, \dots, T$ were then generated according to a normal distribution in order to sample T distinct training sets, each comprising of 30 examples $(x_i^{(t)}, y_i^{(t)})$ such that $y_i^{(t)} = \langle w_t, x_i^{(t)} \rangle + \epsilon$ with ϵ Gaussian noise with zero mean and 0.1 standard deviation. On these learning problems we compared the computational performance of our alternating minimization strategy and the original optimization algorithms originally proposed for MTCL and MTFL and for which the code has been made available by the authors’. In our algorithm we used $A_0 = I$ identity matrix as initialization for the alternating minimization procedure. We used a least-squares loss for all experiments.

Figure 1 reports the comparison of computational times of alternating minimization and the original methods to converge to the same minima (of respectively the functional of MTCL and MTFL). We considered two settings: one where the number of tasks was fixed to $T = 100$ and d increased from 5 to 150 and a second one where d was fixed to 100 and T varied between 5 and 150. To account for statistical stability we repeated the experiments for each couple (T, d) and different choices of hyperparameters while generating a new random datasets at each time. We can make two observations from these results: 1) in the setting where T is kept fixed we observe a linear increase in the computational times for both original MTCL and MTFL methods, while alternating minimization is almost constant with respect to the input space dimension. 2) When d is fixed and the number of tasks increases, all optimization strategies require more time to converge. This shows that in general alternating minimization is a viable option to solve these problems and in particular, when $T \ll \min(d, n)$ – which is often the case in non-linear settings – this method is particularly efficient.

	Accuracy (%) per # tr. samples per class					
	50		100		150	
STL	72.23	±0.04	76.61	±0.02	79.23	±0.01
MTFL	73.23	±.08	77.24	±.05	80.11	±.03
MTRL	73.13	±0.08	77.53	±0.04	80.21	±0.05
OKL	72.25	±0.03	77.06	±0.01	80.03	±0.01

Table 2: Classification results on the 15-scene dataset. Four multi-task methods and the single-task baseline are compared.

5.2 Real dataset

We assessed the benefit of adopting multi-task learning approaches on two real dataset. In particular we considered the following algorithms: Single Task Learning (STL) as a baseline, Multi-task Feature Learning (MTFL) [3], Multi-task Relation Learning (MTRL) [37], Output Kernel Learning (OKL) [14]. We used least squares loss for all experiments.

Sarcos. Sarcos² is a regression dataset designed to evaluate machine learning solutions for inverse dynamics problems in robotics. It consists in a collection of 21-dimensional inputs, i.e. the joint positions, velocities and acceleration of a robotic arm with 7 degrees of freedom and 7 outputs (the tasks), which report the corresponding torques measured at each joint.

For each task, we randomly sampled 50, 100, 150 and 200 training examples while we kept a test set of 5000 examples in common for all tasks. We used a linear kernel and performed 5-fold crossvalidation to find the best regularization parameter according to the normalized mean squared error (nMSE) of predicted torques. We averaged the results over 10 repetitions of these experiments. The results, reported in Table 1, show clearly that to adopt a multi-task approach in this setting is favorable; however, in order to quantify more clearly such improvement, we report in Table 1 also the *normalized improvement (nI)* over single-task learning (STL). For each multi-task method MTL, the normalized improvement nI(MTL) is computed as the average

$$nI(\text{MTL}) = \frac{1}{n_{exp}} \sum_{i=1}^{n_{exp}} \frac{nMSE_i(\text{STL}) - nMSE_i(\text{MTL})}{\sqrt{nMSE_i(\text{STL}) \cdot nMSE_i(\text{MTL})}}$$

over all the $n_{exp} = 10$ experiments of the normalized differences between the nMSE achieved by respectively the STL approach and the given multi-task method MTL.

15-Scenes. 15-Scenes³ is a dataset designed for scene recognition, consisting in a 15-class classification problem. We represented images using LLC coding [35] and trained the system on a training set comprising 50, 100 and 150 examples per class. The test set consisted in 7500 images evenly divided with respect to the 15 scenes. Table 2 reports

²url<http://www.gaussianprocess.org/gpml/data/>

³http://www-cvr.ai.uiuc.edu/ponce_grp/data/

the mean classification accuracy on 20 repetitions of the experiments. It can be noticed that while all multi-task approach seem to achieve approximately similar performance, these are consistently outperforming the STL baseline.

6 Conclusions

We have studied a general multi-task learning framework where the tasks structure can be modeled compactly in a matrix. For a wide family of models, the problem of jointly learning the tasks and their relations can be cast as a convex program, generalizing previous results for special cases [3, 14]. Such an optimization can be naturally approached by block coordinate minimization, which can be seen as alternating between supervised and unsupervised learning steps optimizing respectively the tasks or their structure. We evaluated our method real data, confirming the benefit of multi-task learning when tasks share similar properties.

From an optimization perspective, future work will focus on studying the theoretical properties of block coordinate methods, in particular regarding convergence rates. Indeed, the empirical evidence we report suggests that similar strategies can be remarkably efficient in the multi-task setting. From a modeling perspective, future work will focus on studying wider families of matrix-valued kernels, overcoming the limitations of separable ones. Indeed, this would allow to account also for structures in the interaction space between the input and output domains jointly, which is not the case for separable models.

References

- [1] M. Álvarez, N. Lawrence, and L. Rosasco. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. see also <http://arxiv.org/abs/1106.6251>.
- [2] A. Argyriou, S. Cléménçon, and R. Zhang. Learning the Graph of Relations Among Multiple Tasks. Research report, Oct. 2013.
- [3] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73, 2008.
- [4] A. Argyriou, A. Maurer, and M. Pontil. An algorithm for transfer learning in a heterogeneous environment. In *ECML/PKDD (1)*, pages 71–85, 2008.
- [5] A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 25–32. MIT Press, Cambridge, MA, 2008.
- [6] G. H. Bakir, T. Hofmann, B. Scholkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan. Predicting structured data. *MIT Press*, 2007.
- [7] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- [8] A. Beck and L. Tsetuashvili. On the convergence of block coordinate descent type methods. *Technion, Israel Institute of Technology, Haifa, Israel, Tech. Rep*, 2011.
- [9] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

- [10] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):22, 2012.
- [11] K. Crammer and Y. Singer. On the learnability and design of output codes for multi-class problems. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 35–46, 2000.
- [12] B. Craven. Relations between invex properties. *WORLD SCIENTIFIC SERIES IN APPLICABLE ANALYSIS*, 5:25–34, 1995.
- [13] F. Dinuzzo. Learning output kernels for multi-task problems. *Neurocomputing*, 118:119–126, 2013.
- [14] F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. *International Conference on Machine Learning*, 2011.
- [15] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, pages 615–637, 2005.
- [16] C. Fellbaum. Wordnet: An electronic lexical database. 1998. *MIT Press*, 1998.
- [17] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. *European Conference on Computer Vision*, 2010.
- [18] E. Grave, G. R. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2187–2195. 2011.
- [19] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: a convex formulation. *Advances in Neural Information Processing Systems*, 2008.
- [20] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.
- [21] T. Joachims, T. Hofmann, Y. Yue, and C.-N. Yu. Predicting structured objects with support vector machines. *Commun. ACM*, 52(11):97–104, Nov. 2009.
- [22] A. Kumar and H. Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.
- [23] B. M. Lake and J. B. Tenenbaum. Discovering structure by learning sparse graphs. *Proceedings of the 32nd Cognitive Science Conference*, 2010.
- [24] A. Lozano and V. Sindhvani. Block variable selection in multivariate regression and high-dimensional causal inference. *Advances in Neural Information Processing Systems*, 2011.
- [25] C. A. Micchelli and M. Pontil. Kernels for multi-task learning. *Advances in Neural Information Processing Systems*, 2004.
- [26] H. Q. Minh and V. Sindhvani. Vector-valued manifold regularization. *International Conference on Machine Learning*, 2011.
- [27] Y. Mroueh, T. Poggio, and L. Rosasco. Multi-category and taxonomy learning: A regularization approach. In *NIPS Workshop on Challenges in Learning Hierarchical Models: Transfer Learning and Optimization*, 2011.
- [28] Y. Mroueh, T. Poggio, L. Rosasco, and J.-j. Slotine. Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2012.
- [29] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- [30] M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.
- [31] V. Sindhvani, A. C. Lozano, and H. Q. Minh. Scalable matrix-valued kernel learning and high-dimensional nonlinear causal inference. *CoRR*, abs/1210.4792, 2012.
- [32] I. Steinwart and A. Christmann. *Support vector machines*. Springer, 2008.

- [33] P. Tseng. Convergence of block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494, 2001.
- [34] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. *International Conference on Machine Learning*, 2004.
- [35] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.
- [36] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2764–2770. AAAI Press, 2011.
- [37] Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 733–742, Corvallis, Oregon, 2010. AUAI Press.
- [38] W. Zhong and J. Kwok. Convex multitask learning with flexible task clusters. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 49–56, New York, NY, USA, July 2012. Omnipress.

Appendix

Imposing Known Structure on the Tasks

Coding and Embedding

A common approach to encode knowledge of the tasks relations consists in mapping the output space \mathcal{Y}^T in a new $\tilde{\mathcal{Y}} \subseteq \mathbb{R}^\ell$ and then solve ℓ independent standard learning problems (e.g. RLS, SVM, Boosting, etc. [17]) or a single one with a joint loss (e.g. Ranking [21]) using the mapped outputs as training observation. The goal is to implicitly exploit the structure of the new space to enforce known (or desired) relations among tasks.

The most popular setting for these *embedding* (or *coding*) methods is multi-class classification since in several realistic learning problems, classes can be organized in informative structures such as hierarchies or trees. Interestingly, due to the symbolic nature of the classes representation as canonical basis of \mathbb{R}^T , nonlinear embeddings are not particularly meaningful in classification contexts. Indeed the literature on coding methods for multi-task learning has been mainly concerned with the design of linear operators $L : \mathcal{Y}^T \rightarrow \tilde{\mathcal{Y}}$ [17]. In the following we show that a tight connection exists between coding methods and our multi-task learning setting.

For a fixed linear operator $L \in \mathbb{R}^{\ell \times T}$, we can solve the “coded” problem using the notation of (\mathcal{P}) and a kernel of the form $\Gamma = kI_\ell$ with I_ℓ the $\ell \times \ell$ identity matrix (“independent tasks” kernel)

$$\underset{\tilde{C} \in \mathbb{R}^{n \times \ell}}{\text{minimize}} \quad V(\tilde{Y}, K\tilde{C}) + \lambda \text{tr}(\tilde{C}^\top K\tilde{C}) \quad (6)$$

From the Representer theorem we know that the solution of (6) will have the form $f(x) = \sum_{i=1}^n k(x, x_i) \tilde{c}_i = \sum_{i=1}^n k(x, x_i) Lc_i$, for some $c_i \in \mathbb{R}^T$ and $\tilde{c}_i = Lc_i \in L(\mathbb{R}^T)$. Therefore, we can constrain (6) on matrices $\tilde{C} = CL$ with $C \in \mathbb{R}^{n \times T}$, implying that the best solution for (6) belongs to the set of functions $f = L \circ g \in \mathcal{H}_{kI_\ell}$ with $g \in \mathcal{H}_{kI_T}$.

For those loss functions \mathcal{L} that depend only on the inner product between the vectors of prediction and the ground truth (e.g. logistic or hinge [21, 36], see below), the “coded” Problem (6) on $\tilde{\mathcal{Y}}$ with kernel kI_ℓ is equivalent to (\mathcal{P}) on \mathcal{Y} with kernel $kL^\top L$. More precisely, if the multi-output loss can be written so that $\mathcal{L}(\tilde{y}, f(x)) = \mathcal{L}(\langle \tilde{y}, f(x) \rangle_{\tilde{\mathcal{Y}}})$ for all $\tilde{y} \in \tilde{\mathcal{Y}}$ and $x \in \mathcal{X}$, we have

$$\langle \tilde{y}, f(x) \rangle_{\tilde{\mathcal{Y}}} = \langle Ly, Lg(x) \rangle_{\tilde{\mathcal{Y}}} = \langle y, L^\top Lg(x) \rangle_{\mathcal{Y}} \quad (7)$$

where $y \in \mathcal{Y}$ is such that $Ly = \tilde{y}$ and L^\top denotes the adjoint operator of L (in this case just the transpose matrix since L is a linear operator between vector spaces over the real field). Therefore, the two terms in the functional of (6) become

$$V(\tilde{Y}, K\tilde{C}) = V(YL^\top, KCL^\top) = V(Y, KCL^\top L)$$

where the last equality makes use of the property in eq. (7), and

$$\text{tr}(\tilde{C}^\top K\tilde{C}) = \text{tr}(LC^\top KCL^\top) = \text{tr}(L^\top LC^\top KC)$$

proving the aforementioned equivalence between Problems (6) and (P) by choosing $A = L^\top L$.

Semantic Label Sharing In [17] the authors proposed a strategy to solve a large multi-class visual learning problem that exploited the semantic information provided by the WordNet [16] to enforce specific relations among tasks. In particular, by designing a “semantic” distance between classes using the WordNet graph, the authors were able to generate a similarity matrix $L \in S_+^T$ encoding the most relevant class relations. They used this matrix to map the original outputs (i.e. the canonical basis of \mathbb{R}^T) into a new basis where euclidean distances between output codes would reflect the semantic ones induced by the WordNet priming. Then they applied a semi-supervised One-Vs-All approach on the new output space.

Output Metric

In multi-output settings, another approach to implicitly model the tasks relations consists in changing the metric on the output space \mathbb{R}^T . In particular, we can define a matrix $\Theta \in S_+^T$ and denote the induced inner product on \mathbb{R}^T as $\langle y, y' \rangle_\Theta = \langle y, \Theta y' \rangle_{\mathbb{R}^T}$ for all $y, y' \in \mathbb{R}^T$. For loss functions \mathcal{L} such as those mentioned in Sec. 6 (e.g. hinge, logistic, etc.) that depend only on the inner product between observations and predictions, we have that for a fixed Θ the new loss is defined as $\mathcal{L}_\Theta(y, f(x)) = \mathcal{L}(\langle y, f(x) \rangle_\Theta) = \mathcal{L}(\langle y, \Theta f(x) \rangle_{\mathbb{R}^T})$ and induces a learning problem of the form

$$\underset{C \in \mathbb{R}^{n \times T}}{\text{minimize}} \quad V(\tilde{Y}, KC\Theta) + \lambda \text{tr}(\Theta C^\top KC) \quad (8)$$

which is clearly equivalent to solving (P) choosing the kernel k_Θ . Notice that the second term in eq. (8) derives from the observation that with the new metric, the norm in the RKHS_{VV} becomes $\|f\|_{k_{I_T}}^2 = \langle f, f \rangle_{k_{I_T}} = \sum_{i,j}^n \sum_{t,s}^T k(x_i, x_j) \langle c_t, c_s \rangle_\Theta = \text{tr}(\Theta C^\top KC)$ as required.

metric learning In [24] the authors proposed a metric learning framework in which both the new metric A (or Θ) and the task predictors were estimated simultaneously. Adopting almost the same notation of Problem (Q), they used the least squares loss and imposed a penalty $F(A) = -\log(\det(A))$ on the metric/structure matrix. A further penalty was also imposed on A , in order to enforce specific sparsity patterns. The only difference with our framework is that in [24] the authors do not impose the regularization term $\text{tr}(AC^\top KC)$. Notice however that such term allows us to apply Theorem 3.1 and thus obtain the equivalence between (Q) and (R). This is extremely useful from the optimization perspective since, for instance, for the least squares loss and log-determinant penalty mentioned above, Problem (R) is actually convex jointly, which is not the case for the framework in [24].

Learning the tasks and their structure

Equivalence with the convex problem

We will make use of the following observation

Lemma 6.1. *Consider $K \in S_+^T$ and $C \in \mathbb{R}^{n \times T}$. Then $\text{Ran}(C^\top KC) = \text{Ran}(C^\top \sqrt{K}) = \text{Ran}(C^\top K)$.*

Proof. The second equivalence follows directly from the observation that $C^\top K = (C^\top \sqrt{K})\sqrt{K}$ and $C^\top \sqrt{K} = C^\top K(\sqrt{K})^\dagger$. Regarding the first equivalence, recall that for any $M \in \mathbb{R}^{T \times n}$, $\mathbb{R}^T = \text{Ran}(M) \oplus \text{Ker}(M)$, with $\text{Ker}(M)$ denoting the null space of M . Therefore we can alternatively prove that $\text{Ker}(C^\top KC) = \text{Ker}(C^\top \sqrt{K})$. Notice that clearly $\text{Ker}(C^\top \sqrt{K}) \subseteq \text{Ker}(C^\top KC)$. Now, let $x \in \text{Ker}(C^\top KC)$ so that $0 = x^\top C^\top KCx = x^\top (\sqrt{K}C)^\top (\sqrt{K}C)x$. This implies that x is a singular vector of $(\sqrt{K}C)$ with singular value equal to zero and therefore $x \in \text{Ker}(C^\top \sqrt{K})$. \square

Proof. (Theorem 3.1)

We need to prove that \mathcal{C} is a convex set and that $\text{tr}(A^\dagger C^\top KC)$ is jointly convex on \mathcal{C} . Regarding the first part, notice that for $A \in S_+^T$ and $C \in \mathbb{R}^{n \times T}$ the constraint $\text{Ran}(C^\top KC) \subseteq \text{Ran}(A)$ can be equivalently rewritten as $\text{Ker}(C^\top KC) \supseteq \text{Ker}(A)$. Therefore, using Lemma 6.1, we can check convexity of \mathcal{C} by showing that for any arbitrary couple $(A_1, C_1), (A_2, C_2) \in \mathcal{C}$ and any $\theta \in [0, 1]$ we have $\text{Ker}(\theta A_1 + (1 - \theta)A_2) \subseteq \text{Ker}(\theta C_1^\top K + (1 - \theta)C_2^\top K)$. Let us consider an arbitrary $x \in \text{Ker}(\theta A_1 + (1 - \theta)A_2)$. We have

$$0 = x^\top (\theta A_1 + (1 - \theta)A_2)x = \theta x^\top A_1 x + (1 - \theta)x^\top A_2 x.$$

Since both A_1 and A_2 are PSD, the terms $x^\top A_i x$ are necessarily non-negative for both $i = 1, 2$. Hence, from the equation above we have $x^\top A_i x = 0$, which is equivalent to $x \in \text{Ker}(A_1) \cap \text{Ker}(A_2) \subseteq \text{Ker}(C_1^\top K) \cap \text{Ker}(C_2^\top K)$. This means that x is in the nullspace of both $C_1^\top K$ and $C_2^\top K$ and therefore also in the nullspace of any linear combination of the two. In particular $x \in \text{Ker}(\theta C_1^\top K + (1 - \theta)C_2^\top K)$.

The proof for the convexity of $\text{tr}(A^\dagger C^\top KC)$ has been already pointed out elsewhere (see for instance [5]). For completeness, we provide an simpler derivation of this result which makes use of a Schur's complement argument and simple algebraic properties in line with [14] to show that the epigraph of the function is convex. Consider $A \in S_+^T$ and $C \in \mathbb{R}^{n \times T}$. From simple properties of the trace we have the equivalence $\text{tr}(A^\dagger C^\top KC) = \text{vec}(\sqrt{K}C)^\top (A^\dagger \otimes I_T) \text{vec}(\sqrt{K}C)$, where \otimes identifies the Kronecker product and by $\text{vec}(\cdot)$ we denote the vectorization operator mapping a matrix $M \in \mathbb{R}^{n \times m}$ to the concatenation of all its columns $\text{vec}(M) \in \mathbb{R}^{nm}$. Since $\text{Ran}(A) \supseteq \text{Ran}(C^\top KC) = \text{Ran}(C\sqrt{K})$ we can apply the generalized Schur's com-

plement to write the epigraph of $f(A, C) = \text{tr}(A^\dagger C^\top K C)$ as

$$\begin{aligned} \text{epi } f &= \{(t, A, C) \mid t \geq \text{tr}(A^\dagger C^\top K C)\} = \\ &= \text{vec}(C\sqrt{K})^\top (A^\dagger \otimes I_T) \text{vec}(C\sqrt{K}), (A, C) \in \mathcal{C} \} = \\ &= \left\{ (t, A, C) \mid \begin{pmatrix} A \otimes I_T & \text{vec}(C\sqrt{K}) \\ \text{vec}(C\sqrt{K})^\top & t \end{pmatrix} \succeq 0, \right. \\ &\quad \left. (A, C) \in \mathcal{C} \right\} \end{aligned}$$

where we write $X \succeq Y$ for any two symmetric matrices $X, Y \in S^m$ if and only if $X - Y \in S^m_+$. Notice that the block components of the matrix in the equation above are all linear with respect to A, C and t and therefore the convexity of $\text{epi } f$ follows by directly observing that for any couple $(t_1, A_1, C_1), (t_2, A_2, C_2) \in \text{epi } f$, the PSD constraint holds for any convex combination of the two.

We finally prove that the mapping between minimizers stated in Theorem (3.1). First notice that for any $(C, A) \in \mathbb{R}^{n \times T} \times S^T_+$ we have $Q(C, A) = R(CA, A)$, with $(CA, A) \in \text{dom} R$ since clearly $\text{Ran}(A) \supseteq \text{Ran}(AC^\top KCA)$. Therefore $\inf \{Q(C, A) \mid C \in \mathbb{R}^{n \times T}, A \in S^T_+\} \geq \inf \{R(C, A) \mid (C, A) \in \mathcal{C}\}$. Analogously, given a point $(C, A) \in \mathcal{C}$ we have that $R(C, A) = R(CA^\dagger A, A)$ since $\text{Ran}(C^\top K) \subseteq \text{Ran}(A)$ and thus $V(y, KCAA^\dagger) = V(y, KC)$. Therefore $R(C, A) = R(CA^\dagger A, A) = Q(CA^\dagger, A)$, implying that $\inf \{R(C, A) \mid (C, A) \in \mathcal{C}\} \geq \inf \{Q(C, A) \mid C \in \mathbb{R}^{n \times T}, A \in S^T_+\}$ and concluding the proof. \square

A Barrier Method to Optimize (\mathcal{R})

Proof. (Theorem 3.3) To prove the existence of finite minimizers we need to show that there exists a minimizing sequence for S^δ such that it converges to a point in $\text{dom} S^\delta = \mathbb{R}^{n \times T} \times S^T_{++}$. To see this, consider a generic minimizing sequence, i.e. a sequence $\{(C_n, A_n)\}_{n \in \mathbb{N}} \subset \text{dom} S^\delta$ such that $S^\delta(C_n, A_n) \rightarrow \inf_{C, A} S^\delta(C, A)$. Notice that we can separate C_n in $C_n = \widehat{C}_n + C_n^\perp$ with $\widehat{C}_n \in \text{Ran}(K)$ the range of the Gram matrix K and $C_n^\perp \in \text{Ker}(K)$ its nullspace and that therefore $S^\delta(\widehat{C}_n, A_n) = S^\delta(C_n, A_n)$. This implies that the sequence (\widehat{C}_n, A_n) is bounded, since, if it was not, we would have the coercive penalty F or the $\text{tr}(A_n^{-1} \widehat{C}_n^\top K \widehat{C}_n)$ to go to infinity as n grows. But this is not possible since $S^\delta(\widehat{C}_n, A_n) \rightarrow \inf_{C, A} S^\delta(C, A) < +\infty$. Therefore (\widehat{C}_n, A_n) admits a converging subsequence. Suppose without loss of generality that (C_n, A_n) converges to a point $(C^*, A^*) \in \text{dom} S^\delta = \mathbb{R}^{n \times T} \times S^T_+$. We want to show that (C^*, A^*) is actually in the $\text{dom} S^\delta = \mathbb{R}^{n \times T} \times S^T_{++}$, i.e. that A^* is positive definite. But this is obvious since $\delta > 0$ and therefore if the A_n were to converge to a point in $S^T_+ \setminus S^T_{++}$, we would have that $\delta^2 \text{tr}(A_n^{-1}) \rightarrow +\infty$ and therefore $S^\delta(\widehat{C}_n, A_n) \rightarrow +\infty$ as $n \rightarrow +\infty$. Finally, by the continuity of S^δ , we have $S^\delta(\widehat{C}_n, A_n) \rightarrow S^\delta(C^*, A^*)$, therefore proving that $(C^*, A^*) \in \text{argmin}_{C, A} S^\delta(C, A)$.

The second part of the proof requires the following preliminary steps:

1. $\min_{C, A} R(C, A) = \inf_{A, C} S^0(C, A)$ and they have same infimizers.
2. $g(\delta) = \inf_{A, C} S^\delta(C, A)$ is continuous (in fact convex) with minimum in 0.

We prove the first point in Lemma 6.2, while the second observation follows from the fact that the function g is the point-wise infimum of a jointly convex function over a convex set. This requires to show that $\delta^2 \text{tr}(A^{-1})$ is jointly convex which follows the same reasoning as for the convexity of $\text{tr}(A^{-1}C^\top KC)$ in Theorem (3.1).

Let us consider two sequences $\delta_n > 0$ and $\{(C_n, A_n)\}_{n \in \mathbb{N}} \subset \text{dom}S^\delta = \mathbb{R}^{n \times T} \times S_{++}^T$ satisfying the hypothesis of the Theorem, i.e. $S^{\delta_n}(C_n, A_n) = \min_{C,A} S^{\delta_n}(C, A)$. We will first prove the result for C_n in the range of the Gram matrix K . Notice that under this requirement, the (C_n, A_n) are bounded, since, analogously as for the proof above, if they were not we would have the coercive penalty F or the $\text{tr}(A_n^{-1}C_n^\top KC_n)$ to go to infinity as n grows. But this is not possible since $S^{\delta_n}(C_n, A_n) \rightarrow g(0) < +\infty$. Therefore, by points 1. and 2., $g(0) = \min_{C,A} R(C, A)$ and the limit points of (C_n, A_n) are minimizers for R . This finally implies that there exists a sequence $\{(C_n^*, A_n^*)\}_{n \in \mathbb{N}} \subseteq \text{argmin}_{C,A} R(C, A)$ such that $\|C_n - C_n^*\|_F + \|A_n - A_n^*\|_F$ tends to zero as n goes to infinity. To see this, suppose by contradiction that it is not true and that there exists a subsequence $\{(C_{n_k}, A_{n_k})\}_{k \in \mathbb{N}}$ and an $M > 0$ such that $\|C_{n_k} - C_n^*\|_F + \|A_{n_k} - A_n^*\|_F > M$ for all $k > 0$ and for all $(C^*, A^*) \in \text{argmin}_{C,A} R(C, A)$. Now, since (C_{n_k}, A_{n_k}) is a subsequence of (C_n, A_n) , we have that: (i) (C_{n_k}, A_{n_k}) is bounded (hence admits a converging subsequence) and (ii) every converging subsequence tends to a minimizer of R . This clearly contradicts the hypothesis.

Now, consider the general case in which C_n is not in the range of K : notice that similarly as before, C_n can be separated in $C_n = \hat{C}_n + C_n^\perp$ with $\hat{C}_n \in \text{Ran}(K)$ the range of K and $C_n^\perp \in \text{Ker}(K)$ its nullspace. Clearly, $S^{\delta_n}(\hat{C}_n, A_n) = S^{\delta_n}(C_n, A_n) \rightarrow g(0)$ and therefore, from the discussion above we have a sequence $\{(\hat{C}_n^*, A_n^*)\}_{n \in \mathbb{N}} \subseteq \text{argmin}_{C,A} R(C, A)$ such that $\|\hat{C}_n - \hat{C}_n^*\|_F + \|A_n - A_n^*\|_F \rightarrow 0$ as $n \rightarrow +\infty$. We can now observe that the sequence $(C_n^*, A_n^*) = (\hat{C}_n^* + C_n^\perp, A_n^*)$ satisfies the statement of the Theorem: indeed (i) the (C_n^*, A_n^*) are minimizers for R since $R(C_n^*, A_n^*) = R(\hat{C}_n^*, A_n^*)$ and (ii) $\|C_n - C_n^*\|_F = \|\hat{C}_n - \hat{C}_n^*\|_F \rightarrow 0$ for $n \rightarrow +\infty$. \square

Lemma 6.2. $\min_{A,C} R(C, A) = \inf_{A,C} S^0(C, A)$ and they have same infimizers:

Proof. This fact follows from the observation that for all $\delta > 0$, $\text{dom}S^\delta = \text{dom}S^0$ is equal to the interior of $\text{dom}R$ and that all minimizers for R belong to $\text{dom}R$. To show this second statement we will prove that for any sequence $\{(C_n, A_n)\}_{n \in \mathbb{N}} \subset \text{dom}R$ and converging to some point $(\bar{C}, \bar{A}) \in \mathbb{R}^{n \times T} \times S_+^T \setminus \text{dom}R$, we have that $R(C_n, A_n) \rightarrow +\infty$ as n goes to infinity. For simplicity of notation let us denote $\bar{B} = \bar{C}^\top K \bar{C}$ and analogously $B_n = C_n^\top K C_n$. Since from hypothesis $\text{Ran}(\bar{A}) \not\subseteq \text{Ran}(\bar{C}^\top K \bar{C})$ we have that $\text{Ker}(\bar{A}) \not\subseteq \text{Ker}(\bar{B})$, or, in other words, there exists an eigenvector \bar{v} for \bar{A} such that $v \in \text{Ker}(A)$ and $\|\bar{B}\bar{v}\|_2 > 0$.

Since the sequence A_n converges to \bar{A} , we can identify a sequence of eigenvectors v_n for A_n such that $v_n \rightarrow \bar{v}$ and their associated eigenvalue $\lambda_n \rightarrow 0$ as n goes to infinity. Notice that we can assume without loss of generality that $\lambda_n > 0$ for all n since $\lambda_n = 0$ would imply $v_n \in \text{Ker}(A_n) \subseteq \text{Ker}(B_n)$ but we have from hypothesis that $\|B_n v_n\|_2 \rightarrow \|\bar{B}\bar{v}\|_2 > 0$. Therefore we have

$$\text{tr}(A_n^\dagger B_n) \geq \lambda_n^{-1} v_n^\top B_n v_n = \lambda_n^{-1} \|B_n v_n\|_2^2 \rightarrow +\infty$$

as n goes to infinity. \square

Spectral Regularization

Proposition 3.6 follows directly from the following result

Proposition 6.3. *Let $A, M \in S_+^n$ with $\text{Ran}(A) \supseteq \text{Ran}(M)$, $\text{rank}(M) = r$. Let $M = U\Sigma U^\top$ be an eigendecomposition of M with $U \in O^n$ and $\Sigma \in S_+^n$ a diagonal matrix with eigenvalues in decreasing order. Then, there exists a matrix $A_* = U\Gamma U^\top \in S_+^n$ with $\Gamma \in S_+^n$ diagonal with $\Gamma_{i,i} = 0 \forall i < r$, such that*

$$\text{tr}(A_*^\dagger M) = \text{tr}(A^\dagger M) \quad \text{and} \quad \|A_*\|_p \leq \|A\|_p \quad \forall p \geq 1 \quad (9)$$

with the equality holding if and only if $A_* = A$.

Proof. To keep the notation uncluttered we prove the result for $\Theta = A^\dagger$. Consider an eigendecomposition $\Theta = \Lambda S \Lambda^\top$ with $S \in O^n$ and $\Lambda \in S_+^n$ diagonal with eigenvalues in decreasing order. Let us define $R = U^\top S \in O^n$. Then

$$\text{tr}(\Theta M) = \text{tr}(R \Lambda R^\top \Sigma) = \sum_{i=1}^r \sigma_i \sum_{j=1}^n R_{ij}^2 \lambda_j = \sum_{i=1}^r \sigma_i \gamma_i$$

where σ_i and λ_i are respectively the i -th eigenvalues of M and Θ and we have defined $\gamma_i = \sum_{j=1}^n R_{ij} \lambda_j$ for $i \leq r$ and $\gamma_i = 0$ otherwise. Hence, if we consider a diagonal matrix $\Gamma \in S_+^n$ such that $\Gamma_{ii} = \gamma_i$ and set $\Theta' = U\Gamma U^\top$ we obtain the left equivalence of eq. (9), namely $\text{tr}(\Theta M) = \text{tr}(\Theta' M)$. Now, consider the p -Schatten norm of Θ'

$$\|(\Theta')^\dagger\|_p = \left(\sum_{i=1}^r \frac{1}{\gamma_i^p} \right)^{1/p} = \left(\sum_{i=1}^r \frac{1}{\left(\sum_{j=1}^n R_{ij}^2 \lambda_j \right)^p} \right)^{1/p}.$$

Notice that $R_{ij} = U_i^\top \cdot S_j$ corresponds to the projection of the i -th eigenvector of M on the j -th eigenvector of Θ . Since $\text{Ran}(\Theta) = \text{Ran}(A) \supseteq \text{Ran}(M)$, for any eigenvector $s \in \mathbb{R}^n$ in the nullspace of Θ (i.e. with associated eigenvalue $\lambda = 0$), we have that $U_i^\top \cdot s = 0$ for all $i \leq r$. Hence, $\forall i \leq r$, $1 = R_i^\top \cdot R_i = \sum_{j=1}^n R_{ij}^2 = \sum_{j=1}^k R_{ij}^2$, where $k = \text{rank}(A)$. Therefore, since the R_{ij}^2 s add up to 1 and the scalar function $(1/x)^p$ is convex in $x \in \mathbb{R}_+$, we have

$$\begin{aligned} \sum_{i=1}^r \frac{1}{\left(\sum_{j=1}^n R_{ij}^2 \lambda_j \right)^p} &\leq \sum_{i=1}^r \sum_{j=1}^k R_{ij}^2 \frac{1}{\lambda_j^p} \leq \\ &\leq \sum_{j=1}^k \frac{1}{\lambda_j^p} \sum_{i=1}^n R_{ij}^2 = \sum_{j=1}^k \frac{1}{\lambda_j^p} = \|\Theta^\dagger\|_p^p \end{aligned}$$

where we have made use of the fact that for all $j = 1, \dots, n$ we have $\sum_{i=1}^n R_{ij} = R_j^\top \cdot R_j = 1$. Therefore, $\|(\Theta')^\dagger\|_p \leq \|\Theta^\dagger\|_p$. By taking $A' = (\Theta')^\dagger$ we have the desired result. \square

Applied to the minimization in problem (\mathcal{R}) with $C \in \mathbb{R}^{n \times T}$ fixed and p -Schatten penalty, Proposition 6.3 states that a minimizer $A_C \in S_+^T$ has the same system of eigenvalues as $C^\top K C$ and their spectrum have same sparsity pattern (i.e. $\text{Ran}(C^\top K C) = \text{Ran}(A)$). This observation leads directly to the closed formula to find a A_* stated in Proposition 3.6.

Proof. (Proposition 3.6) Consider the eigendecomposition $C^\top K C = M = U \Sigma U^\top$ with $U \in O^T$ and $\Sigma \in S_+^T$ diagonal with the eigenvalues arranged in descending order. We apply Proposition 6.3 and obtain the minimizer $A_* = U \Gamma U^\top$ for $\Gamma \in S_+^T$ diagonal with same sparsity pattern as Σ . We can rewrite the target function as

$$\sum_{t=1}^r \frac{\sigma_t}{\gamma_t} + \lambda \gamma_t.$$

where $r = \text{rank}(M)$. Therefore, the optimization problem consists in minimizing the target function above with respect to the γ_t s. This is an unconstrained convex optimization of a differentiable coercive function bounded below and therefore it is sufficient to set the gradient to zero and solve with respect to the γ_t . It is clear that for each $t = 1 \dots r$, the minimizer is of the form $\gamma_t = \sqrt[p+1]{\sigma_t/\lambda}$, leading to the desired solution. \square

Linear Multi-task Learning

Several works in multi-task learning have focused on linear models where the multi-output predictor $f : \mathbb{R}^d \rightarrow \mathbb{R}^T$ is parameterized by a matrix $W \in \mathbb{R}^{d \times T}$ whose columns $w_t \in \mathbb{R}^d$ are associated to the individual task-predictors $f_t(x) = \langle w_t, x \rangle_{\mathbb{R}^d}$ for any $x \in \mathbb{R}^d$. In this tasks structure can be imposed considering suitable matrix penalty $\Omega : \mathbb{R}^{d \times T} \rightarrow \mathbb{R}$ and regularization schemes of form

$$\min_{W \in \mathbb{R}^{d \times T}} V(Y, XW) + \Omega(W) \quad (10)$$

where $X \in \mathbb{R}^{n \times d}$ is the matrix whose rows correspond to the (transposed) input points in the training sets, ordered accordingly to the order in Y ⁴. We can recognize two main classes of penalty functions. A first class correspond to methods that impose structured sparsity on the input features across the multiple tasks, for instance considering the penalty $\Omega(\cdot) = \|\cdot\|_{2,1}$ [3], which encourages whole rows of W to be simultaneously sparse, see also [20, 38]. A second class corresponds to spectral regularization methods defined by penalties Ω acting on the singular values of W . Examples in this class include methods that impose low-rank assumptions [3] on the tasks, or search after tasks-cluster structures [19]. Ideas related to a combination of the above methods can also be considered [10].

Most Linear multi-task learning problems of the form (10) with Ω spectral penalty, can be formulated in terms of problem (\mathcal{R}) for a suitable choice of F . Indeed it can be

⁴Again V would weight with zeros the loss associated to entries for which examples are not available during training

shown that for several spectral norms, such as the p-schatten norms, the penalty Ω can be written as

$$\Omega(W) = \inf_{A \in S_{++}^T} \text{trace}(WA^{-1}W^\top) + F_\Omega(A) \quad \forall W \in \mathbb{R}^{n \times T}$$

Here we report the example of the nuclear norm $\|\cdot\|_*$, that has already been observed in similar form in [3, 18] and that can be easily derived from Prop. 3.6 for the case $p = 1$.

$$\|W\|_* = \frac{1}{2} \inf_{A \in S_{++}^T} \text{trace}(WA^{-1}W^\top) + \text{trace}(A).$$

Indeed, from Prop. (3.6) we have that the solution to the minimization problem is $A_* = \sqrt{(W^\top \text{op} W)}$ and therefore, the minimum of such functional will be exactly $\text{trace}(\sqrt{WW^\top}) = \|W\|_*$.

Impose Tasks Relationships by enforcing structure on the feature space

Relations among tasks can be also modeled by enforcing shared structures on the input space. For instance in [3], the authors generalized a feature selection framework to the multi-task setting by formulating the linear problem

$$\underset{U \in O^d, M \in \mathbb{R}^{d \times T}}{\text{minimize}} \quad V(Y, XUM) + \gamma \|M\|_{2,1} \quad (11)$$

where $X \in \mathbb{R}^{n \times d}$ is the matrix whose i -th row corresponds to the input vector $x_i \in \mathbb{R}^d$ and the $(2, 1)$ -norm $\|M\|_{2,1} = \sum_{k=1}^d \|M^k\|_2$ is introduced to enforce sparsity among the rows M^k of M . This penalty generalizes feature selection to the multi-task case by directly manipulating the covariance on the input space. However, since input and output distributions are connected by the training data, it is reasonable to expect this process to indirectly affect also the covariance on the output space. Indeed, in this Section we present an interesting result connecting multi-task problems that impose structure on the input covariance and problems that instead aim to control the output covariance (i.e. in the form of (\mathcal{R})).

To show this connection, we need to discuss in more detail the work in [3]. Although (11) is not convex, the authors prove that there exists an equivalent convex formulation of the form

$$\underset{\substack{W \in \mathbb{R}^{d \times T}, D \in S_{++}^d \\ \text{Ran}(D) \supseteq \text{Ran}(W), \text{tr}(D) \leq 1}}{\text{minimize}} \quad V(Y, XW) + \gamma \text{tr}(W^\top D^\dagger W). \quad (12)$$

The authors then proceed to generalize this framework to the nonlinear case using the advantages of the RKHS notation. In this setting, the original idea of identifying a low dimensional set of directions in the feature space translates naturally to the problem of finding a small set of orthogonal directions in the Hilbert space. To this end,

the authors perform a preprocessing step whose goal is to identify an orthonormal basis of functions $\psi_1, \dots, \psi_\ell \in \mathcal{H}_k$ for set spanned by the $k(x_i, \cdot)$ and define a matrix $\tilde{K} \in \mathbb{R}^{n \times \ell}$ such that $\tilde{K}_{ij} = \psi_j(x_i)$. A possible way to do this is by considering an eigenvalue decomposition $U\Sigma U^\top$ of K and taking $\tilde{K} = U\Sigma^{1/2}$ (taking out from $\Sigma^{1/2}$ the columns equal to zero). It is easy to show that the standard learning problem in RKHS settings can be cast equivalently in this new notation. However, this framework has the further advantage that it can be generalized to take into account the eventuality of a transformation in the feature space, leading to the extension of problem (12) for the non linear case

$$\begin{aligned} & \underset{\substack{B \in \mathbb{R}^{\ell \times T}, D \in S_+^\ell, \\ \text{Ran}(D) \supseteq \text{Ran}(B), \text{tr}(D) \leq 1}}}{\text{minimize}} & V(Y, \tilde{K}B) + \gamma \text{tr}(B^\top D^\dagger B) \end{aligned} \quad (13)$$

As can be noticed, the structure of problem (13) is very similar to the one of problem (R) and indeed, as stated in Corollary 6.5 the two are equivalent when trace regularization is imposed on (R). However, as shown in Theorem 6.4, a more general equivalence holds.

Theorem 6.4. *Let $\lambda > 0$, $p \geq 1$, $\mathbf{R}^{n \times T}$, $\{x_i, y_i\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^T$ a set of input-output pairs with $\mathbf{y} \in \mathbb{R}^{n \times T}$ the matrix whose i -th row corresponds to y_i . Let $\psi_1, \dots, \psi_\ell \in \mathcal{H}_k$ be an orthonormal basis for $\text{span}\{k(x_i, \cdot)\}_{i=1}^n$ and $\tilde{K} \in \mathbb{R}^{n \times \ell}$ with $\tilde{K}_{ij} = \psi_j(x_i)$. Then*

$$\begin{aligned} & \underset{\substack{B \in \mathbb{R}^{\ell \times T}, D \in S_+^\ell, \\ \text{Ran}(D) \supseteq \text{Ran}(B)}}}{\text{minimize}} & S(B, D) = V(Y, \tilde{K}B) + \text{tr}(B^* D^\dagger B) + \lambda \|D\|_p \end{aligned} \quad (\mathcal{T})$$

is a convex optimization problem equivalent to (R) with penalty function $F(A) = \|A\|_p$. In particular the two problems achieve the same minimum and, given a minimizer for one problem it is possible to obtain a solution for the other and vice-versa.

The crucial aspect of the proof of Theorem 6.4 (which we prove below) consists in identifying the two mappings that allow to obtain a minimizer for problem (R) from a solution of (T) and vice-versa.

As a corollary of Theorem (6.4) we get the exact equivalence to the problem proposed in [3].

Corollary 6.5. *Problem (13) is equivalent to (T) for $p = 1$. In particular the two problems achieve the same minimum for $\lambda = \gamma^2/4$. As a consequence of Theorem 6.4 this implies also that (13) is also equivalent to (R) when $F(\cdot) = \|\cdot\|_1 = \text{tr}(\cdot)$.*

This result follows from the direct comparison of the minimizers for the problems (T) (from Proposition 3.6) and (13) (from [3]). Notice, that although equivalent as convex optimizations, it is in general more convenient to solve problems in the form (R) rather than (T) since in most cases $T \ll \ell$.

Proof. Theorem 6.4.

From the discussion in [3] we can rewrite problem (R) in the equivalent formulation

$$\begin{aligned} & \underset{\substack{B \in \mathbb{R}^{\ell \times T}, A \in S_+^T, \\ \text{Ran}(A) \supseteq \text{Ran}(B^\top)}}}{\text{minimize}} & T(B, A) = V(Y, \tilde{K}B) + \text{tr}(A^\dagger B^\top B) + \lambda \|A\|_p \end{aligned} \quad (\mathcal{U})$$

Therefore, to prove Theorem 6.4 it is sufficient to show that problem (\mathcal{T}) and (\mathcal{U}) are equivalent. Assume without loss of generality $T \leq \ell$. Consider an arbitrary matrix $B \in \mathbb{R}^{\ell \times T}$ and a singular value decomposition $B = V \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} U^\top$ where $0 \in \mathbb{R}^{(\ell-T) \times T}$ identifies a matrix of all zeros, $V \in O^\ell, U \in O^T$ and $\Sigma \in S_+^T$ a diagonal matrix with eigenvalues in descending order. From Proposition 6.3, we obtain that the minimizers of the two functions $S(B, \cdot)$ and $T(B, \cdot)$ are unique and can be written respectively in the forms

$$D_B = V \begin{pmatrix} \Gamma_D & 0 \\ 0 & 0 \end{pmatrix} V^\top \in S_+^\ell \quad \text{and} \quad A_B = U \Gamma_A U^\top \in S_+^T$$

where $\Gamma_D, \Gamma_A \in S_+^T$ have same sparsity pattern as Σ and the zero matrices in the formulation of D_B are of appropriate dimension. We can therefore write the minimum value achieved by $S(B, \cdot)$ as $S(B, D_B) = V(Y, \tilde{K}B) + \text{tr}(\Gamma_D^\dagger \Sigma^2) + \lambda \|\Gamma_D\|_p$ and the minimum achieved by $T(B, \cdot)$ as $T(B, A_B) = V(Y, \tilde{K}B) + \text{tr}(\Gamma_A^\dagger \Sigma^2) + \lambda \|\Gamma_A\|_p$. In the light of these equations, it can be easily checked that by setting $A_B^{(D)} = U \Gamma_D U^\top \in S_+^T$ we have

$$S(B, D_B) = T(B, A_B^{(D)}) \geq T(B, A_B)$$

where the inequality follows from the fact that A_B is a minimizer for $T(B, \cdot)$. Analogously, we can design a matrix $D_B^{(A)} \in S_+^\ell$ such that $T(B, A_B) = S(B, D_B^{(A)}) \geq S(B, D_B)$. Since the minimizers A_B and D_B are unique, it follows that $\Gamma_D = \Gamma_A$. In the perspective of this result, we have that for any minimizer $(B_*, D_*) \in \mathbb{R}^{\ell \times T} \times S_+^\ell$ for (\mathcal{T}) , the couple $(B_*, A_{B_*}^{(D_*)}) \in \mathbb{R}^{\ell \times T} \times S_+^T$ is a minimizer for (\mathcal{U}) and furthermore, the two functions achieve the same minimum value. The same result holds in the opposite direction. \square