Supplementary Material

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 $\widetilde{\mathcal{Y}} \subseteq \mathbb{R}^\ell$ and then solve ℓ independent standard learning problems (e.g. RLS, SVM, Boosting, etc. (Fergus et al., 2010)) or a single one with a joint loss (e.g. Ranking (Joachims et al., 2009)) 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 $L : \mathcal{Y}^T \to \widetilde{\mathcal{Y}}$ (Fergus et al., 2010). 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)

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

From the Representer theorem we know that the solution of (6) will have the form $f(x) = \sum_{i=1}^n k(x, x_i)\tilde{c}_i = \sum_{i=1}^n k(x, x_i) L c_i$, for some $c_i \in \mathbb{R}^T$ and $\tilde{c}_i = L c_i \in L(\mathbb{R}^T)$. Therefore, we can constrain (6) on matrices $\widetilde{C} = CL$ 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 (Joachims et al., 2009; Weston et al., 2011), see below), the "coded" Problem (6) on $\widetilde{\mathcal{Y}}$ 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 $\mathcal{L}(\tilde{y}, f(x)) = \mathcal{L}(\langle \tilde{y}, f(x) \rangle_{\widetilde{\mathcal{Y}}})$ for all $\tilde{y} \in \widetilde{\mathcal{Y}}$ and $x \in \mathcal{X}$, we have

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

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

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

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

$$tr(\widetilde{C}^\top K \widetilde{C}) = tr(LC^\top K C L^\top) = tr(L^\top L C^\top K C)$$

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

Semantic Label Sharing In (Fergus et al., 2010) the authors proposed a strategy to solve a large multi-class visual learning problem that exploited the semantic information provided by the WordNet (Fellbaum, 1998) 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. 7 (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

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

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

metric learning In (Lozano & Sindhwani, 2011) 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 ($\mathcal{Q}$), 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 (Lozano & Sindhwani, 2011) the authors do not impose the regularization term $tr(AC^\top KC)$. Notice however that such term allows us to apply Theorem 3.1 and thus obtain the equivalence between ($\mathcal{Q}$) and ($\mathcal{R}$). This is extremely useful from the optimization perspective since, for instance, for the least squares loss and log-determinant penalty mentioned above, Problem ($\mathcal{R}$) is actually convex jointly, which is not the case for the framework in (Lozano & Sindhwani, 2011).

Learning the tasks and their structure

Equivalence with the convex problem

We will make use of the following observation

Lemma 7.1. *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 $x \in \text{Ker}(C^\top\sqrt{K})$. □

Proof. (Theorem 3.1)

We need to prove that $\mathcal{C}$ is a convex set and that $tr(A^\dagger C^\top KC)$ 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 7.1, 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 $tr(A^\dagger C^\top KC)$ has been already pointed out elsewhere (see for instance (Argyriou et al., 2008c)). 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 (Dinuzzo et al., 2011) 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 $tr(A^\dagger C^\top KC) = vec(\sqrt{K}C)^\top(A^\dagger \otimes I_T)vec(\sqrt{K}C)$, where $\otimes$ identifies the Kronecker product and by $vec(\cdot)$ we denote the vectorization operator mapping a matrix $M \in \mathbb{R}^{n\times m}$ to the concatenation of all its columns $vec(M) \in \mathbb{R}^{nm}$. Since $\text{Ran}(A) \supseteq \text{Ran}(C^\top KC) = \text{Ran}(C\sqrt{K})$ we can apply the generalized Schur's complement to write the epigraph of $f(A, C) = tr(A^\dagger C^\top KC)$ as

$$epi\ f = \Big\{(t, A, C) \ \Big|\ t \geq tr(A^\dagger C^\top KC) = vec(C\sqrt{K})^\top(A^\dagger \otimes I_T)vec(C\sqrt{K}), (A, C) \in \mathcal{C}\Big\} =$$

$$= \left\{(t, A, C) \ \middle|\ \begin{pmatrix} A \otimes I_T & vec(C\sqrt{K}) \\ vec(C\sqrt{K})^\top & t \end{pmatrix} \succeq 0, \ (A, C) \in \mathcal{C}\right\}$$

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 $epi\ f$ follows by directly observing that for any couple $(t_1, A_1, C_1), (t_2, A_2, C_2) \in epi\ f$, 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 $(CA, A) \in domR$ 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. □

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 $domS^\delta = \mathbb{R}^{n\times T} \times S_{++}^T$. To see this, consider a generic minimizing sequence, i.e. a sequence $\{(C_n, A_n)\}_{n\in\mathbb{N}} \subset domS^\delta$ such that $S^\delta(C_n, A_n) \to inf_{C,A} S^\delta(C, A)$. 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 $tr(A_n^{-1}\widehat{C}_n^\top K\widehat{C}_n)$ to go to infinity as n grows. But this is not possible since $S^\delta(\widehat{C}_n, A_n) \to inf_{C,A} S^\delta(C, A) < +\infty$. Therefore $(\widehat{C}_n, A_n)$ admits a converging subsequence. Suppose without loss of generality that (C_n, A_n) converges to a point $(C^*, A^*) \in \overline{domS^\delta} = \mathbb{R}^{n\times T} \times S_+^T$. We want to show that (C^*, A^*) is actually in the $domS^\delta = \mathbb{R}^{n\times T} \times S_{++}^T$, 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 $S_+^T \backslash S_{++}^T$, we would have that $\delta^2\ tr(A_n^{-1}) \to +\infty$ and therefore $S^\delta(\widehat{C_n}, A_n) \to +\infty$ as $n \to +\infty$. Finally, by the continuity of S^δ, we have $S^\delta(\widehat{C}_n, A_n) \to S^\delta(C^*, A^*)$, 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 7.2, 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 $\delta^2 tr(A^{-1})$ is jointly convex which follows the same reasoning as for the convexity of $tr(A^{-1}C^\top KC)$ in Theorem (3.1).

Let us consider two sequences $\delta_n > 0$ and $\{(C_n, A_n)\}_{n\in\mathbb{N}} \subset domS^\delta = \mathbb{R}^{n\times T} \times S_{++}^T$ 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 $tr(A_n^{-1}C_n^\top KC_n)$ to go to infinity as n grows. But this is not possible since $S^{\delta_n}(C_n, A_n) \to g(0) < +\infty$. 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 $\{(C_n^*, A_n^*)\}_{n\in\mathbb{N}} \subseteq argmin_{C,A} R(C, A)$ 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 $\|C_{n_k} - C^*\|_F + \|A_{n_k} - A^*\|_F > M$ 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 $C_n = \widehat{C}_n + C_n^\perp$ with $\widehat{C}_n \in \text{Ran}(K)$ the range of K and $C_n^\perp \in \text{Ker}(K)$ its nullspace. Clearly, $S^{\delta_n}(\widehat{C}_n, A_n) = S^{\delta_n}(C_n, A_n) \to g(0)$ and therefore, from the discussion above we have a sequence $\{(\widehat{C}_n^*, A_n^*)\}_{n\in\mathbb{N}} \subseteq \text{argmin}_{C,A} R(C, A)$ such that $\|\widehat{C}_n - \widehat{C}_n^*\|_F + \|A_n - A_n^*\|_F \to 0$ as $n \to +\infty$. We can now observe that the sequence $(C_n^*, A_n^*) = (\widehat{C}_n^* + C_n^\perp, A_n^*)$ satisfies the statement of the Theorem: indeed (i) the (C_n^*, A_n^*) are minimizers for R since $R(C_n^*, A_n^*) = R(\widehat{C}_n^*, A_n^*)$ and (ii) $\|C_n - C_n^*\|_F = \|\widehat{C}_n - \widehat{C}_n^*\|_F \to 0$ for $n \to +\infty$. □

Lemma 7.2. *$min_{A,C} R(C, A) = inf_{A,C} S^0(C, A)$ and they have same infimizers:*

Proof. This fact follows from the observation that for all $\delta > 0$, $domS^\delta = domS^0$ is equal to the interior of $domR$ and that all minimizers for R belong to $domR$. To show this second statement we will prove that for any sequence $\{(C_n, A_n)\}_{n\in\mathbb{N}} \subset domR$ and converging to some point $(\bar{C}, \bar{A}) \in \mathbf{R}^{n\times T} \times S_+^T \setminus domR$, we have that $R(C_n, A_n) \to +\infty$ as n goes to infinity. For simplicity of notation let us denote $\bar{B} = \bar{C}^\top K\bar{C}$ and analogously $B_n = C_n^\top KC_n$. Since from hypothesis $\text{Ran}(\bar{A}) \not\supseteq \text{Ran}(\bar{C}^\top K\bar{C})$ 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 $v_n \to \bar{v}$ and their associated eigenvalue $\lambda_n \to 0$ 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 $\|B_n v_n\|_2 \to \|\bar{B}\bar{v}\| > 0$. Therefore we have

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

as n goes to infinity. □

Spectral Regularization

Proposition 3.6 follows directly from the following result

Proposition 7.3. *Let $A, M \in S_+^n$ with $\text{Ran}(A) \supseteq \text{Ran}(M)$, $rank(M) = r$. 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

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

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 eigendecompositionn $\Theta = S\Lambda S^\top$ 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

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

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 $tr(\Theta M) = tr(\Theta' M)$. 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 $k = rank(A)$. Therefore, since the R_{ij}^2s add up to 1 and the scalar function $(1/x)^p$ is convex in $x \in \mathbb{R}_+ +$, we have

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

where we have made use of the fact that for all $j = 1, \ldots, n$ 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. □

Applied to the minimization in problem ($\mathcal{R}$) with $C \in \mathbb{R}^{n\times T}$ fixed and p-Schatten penalty, Proposition 7.3 states that a minimizer $A_C \in S_+^T$ has the same system of eigenvalues as $C^\top KC$ and their spectrum have same sparsity pattern (i.e. $\text{Ran}(C^\top KC) = \text{Ran}(A)$). This observation leads directly to the closed formula to find a A_* stated in Proposition 3.6.

Proof. (Proposition 3.6) Consider the eigendecomposition $C^\top KC = M = U\Sigma U^\top$ with $U \in O^T$ and $\Sigma \in S_+^T$ diagonal with the eigenvalues arranged in descending order. We apply Proposition 7.3 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 $r = rank(M)$. Therefore, the optimization problem consists in minimizing the target function above with respect to the γ_ts. 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 $t = 1 \ldots r$, the minimizer is of the form $\gamma_t = \sqrt[p+1]{\sigma_t/\lambda}$, leading to the desired solution. □

Linear Multi-task Learning

Several works in multi-task learning have focused on linear models where the multi-output predictor $f : \mathbb{R}^d \to \mathbb{R}^T$ 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 $\Omega : \mathbb{R}^{d\times T} \to \mathbb{R}$ and regularization schemes of form

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

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 $\Omega(\cdot) = \|\cdot\|_{2,1}$ (Argyriou et al., 2008a), which encourages whole rows of W to be simultaneously sparse, see also (Jayaraman et al., 2014; Zhong & Kwok, 2012). 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 (Argyriou et al., 2008a) on the tasks, or search after tasks-cluster structures (Jacob et al., 2008). Ideas related to a combination of the above methods can also be considered (Chen et al., 2012).

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 shown that for several spectral norms, such as the p-

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

schatten norms, the penalty Ω can be written as

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

Here we report the example of the nuclear norm $\|\cdot\|_*$, that has already been observed in similar form in (Argyriou et al., 2008a; Grave et al., 2011) and that can be easily derived from Prop. 3.6 for the case $p = 1$.

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

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

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 (Argyriou et al., 2008a), the authors generalized a feature selection framework to the multi-task setting by formulating the linear problem

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

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 (Argyriou et al., 2008a). Although (11) is not convex, the authors prove that there exists an equivalent convex formulation of the form

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

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 $\widetilde{K} \in \mathbb{R}^{n \times \ell}$ such that $\widetilde{K}_{ij} = \psi_j(x_i)$. A possible way to do this is by considering a eigenvalue decomposition $U \Sigma U^\top$ of K and taking $\widetilde{K} = U \Sigma^{1/2}$ (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

$$\underset{\substack{B \in \mathbb{R}^{\ell \times T}, D \in S^\ell_+, \\ \text{Ran}(D) \supseteq \text{Ran}(B), tr(D) \leq 1}}{\text{minimize}} \quad V(Y, \widetilde{K}B) + \gamma \, tr(B^\top D^\dagger B) \qquad (13)$$

As can be noticed, the structure of problem (13) is very similar to the one of problem ($\mathcal{R}$) and indeed, as stated in Corollary 7.5 the two are equivalent when trace regularization is imposed on ($\mathcal{R}$). However, as shown in Theorem 7.4, a more general equivalence holds.

Theorem 7.4. *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 span$\{k(x_i, \cdot\}_{i=1}^n$ and $\widetilde{K} \in \mathbb{R}^{n \times \ell}$ with $\widetilde{K}_{ij} = \psi_j(x_i)$. Then*

$$\underset{\substack{B \in \mathbb{R}^{\ell \times T}, D \in S^\ell_+, \\ \text{Ran}(D) \supseteq \text{Ran}(B)}}{\textit{minimize}} \quad S(B, D) = V(Y, \widetilde{K}B) + tr(B^* D^\dagger B) + \lambda \, \|D\|_p \qquad (\mathcal{T})$$

is a convex optimization problem equivalent to ($\mathcal{R}$) 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 7.4 (which we prove below) consists in identifying the two mappings that allow to obtain a minimizer for problem ($\mathcal{R}$) from a solution of ($\mathcal{T}$) and vice-versa.

As a corollary of Theorem (7.4) we get the exact equivalence to the problem proposed in (Argyriou et al., 2008a).

Corollary 7.5. *Problem (13) is equivalent to ($\mathcal{T}$) for $p = 1$. In particular the two problems achieve the same minimum for $\lambda = \gamma^2/4$. As a consequence of Theorem 7.4 this implies also that (13) is also equivalent to ($\mathcal{R}$) when $F(\cdot) = \|\cdot\|_1 = tr(\cdot)$.*

This result follows from the direct comparison of the minimizers for the problems ($\mathcal{T}$) (from Proposition 3.6) and (13) (from (Argyriou et al., 2008a)). Notice, that although equivalent as convex optimizations, it is in general

more convenient to solve problems in the form ($\mathcal{R}$) rather than ($\mathcal{T}$) since in most cases $T << \ell$.

Proof. Theorem 7.4.

From the discussion in (Argyriou et al., 2008a) we can rewrite problem ($\mathcal{R}$) in the equivalent formulation

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

Therefore, to prove Theorem 7.4 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 Propositon 7.3, 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 $S(B,D_B) = V(Y,\widetilde{K}B) + tr(\Gamma_D^\dagger\Sigma^2) + \lambda\|\Gamma_D\|_p$ and the minimum achieved by $T(B,\cdot)$ as $T(B,A_B) = V(Y,\widetilde{K}B)+tr(\Gamma_A^\dagger\Sigma^2)+\lambda\|\Gamma_A\|_p$. In the light of these equations, it can be easily cheked 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. □