Conditional Meta-Learning of Linear Representations

Giulia Denevi[*]
Leonardo Labs (Italy)
giulia.denevi.ext@leonardo.com

Massimiliano Pontil
Istituto Italiano di Tecnologia (Italy) & University College of London (UK)
massimiliano.pontil@iit.it

Carlo Ciliberto
University College of London (UK) & Istituto Italiano di Tecnologia (Italy)
c.ciliberto@ucl.ac.uk

Abstract

Standard meta-learning for representation learning aims to find a common representation to be shared across multiple tasks. The effectiveness of these methods is often limited when the nuances of the tasks' distribution cannot be captured by a single representation. In this work we overcome this issue by inferring a conditioning function, mapping the tasks' side information (such as the tasks' training dataset itself) into a representation tailored to the task at hand. We study environments in which our conditional strategy outperforms standard meta-learning, such as those in which tasks can be organized in separate clusters according to the representation they share. We then propose a meta-algorithm capable of leveraging this advantage in practice. In the unconditional setting, our method yields a new estimator enjoying faster learning rates and requiring less hyper-parameters to tune than current state-of-the-art methods. Our results are supported by preliminary experiments.

1 Introduction

Learning a shared representation among a class of machine learning problems is a well-established approach used both in multi-task learning [3, 20, 11] and meta-learning [18, 15, 5, 17, 35, 24, 30, 9, 7]. The idea behind this methodology is to consider two nested problem: at the within-task level an empirical risk minimization is performed on each task, using inputs transformed by the current representation, on the outer-task (meta-) level, such a representation is updated taking into account the errors of the within-task algorithm on previous tasks.

Such a technique was shown to be advantageous in contrast to solving each task independently when the tasks share a low dimensional representation, see e.g. [27, 25, 15, 24, 35, 5, 22, 9]. However, in real world applications we often deal with heterogeneous classes of learning tasks, which may overall be only loosely related. Consequently, the tasks' commonalities may not be captured well by a single representation shared among all the tasks. This is for instance the case in which the tasks can be organized in different groups (clusters), where only tasks belonging to the same cluster share the same low-dimensional representation.

In order to overcome this issue, previous authors developed non-convex methods (or convex relaxations) attempting at clustering the tasks, see e.g. [4, 26, 2, 20, 28, 40, 38, 31]. In this work, we follow

[*]Work done while the first author was with Istituto Italiano di Tecnologia (Italy).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the recent literature on heterogeneous meta-learning [37, 36, 32, 21, 10, 39, 14, 7] and propose a so-called *conditional meta-learning* approach for meta-learning a representation. Our algorithm learns a conditioning function mapping available tasks' side information into a *linear* representation that is tuned to that task at hand. Our approach borrows from [14], where the authors proposed a conditional meta-learning approach for fine tuning and biased regularization. In those cases however, the tasks' target vectors are assumed to be all close to a common bias vector rather than sharing the same low-dimensional linear representation, as instead explored in this work. As we explain in the following, working with the representation setting requires a significant contribution with respect to the bias one in order to give a new formulation of the problem, to consider a different meta-objective and a different interpretation of the results. In addition, the representation setting is known to be a more relevant and effective framework in many scenarios in comparison to the bias one (see e.g. [28]).

In this work, we propose for the first time an online conditional method for linear representation learning with strong theoretical guarantees. In particular, we show that the method is advantageous over standard (unconditional) representation learning methods used in meta-learning when the environment of observed tasks is heterogeneous.

Contributions and organization. The contributions of this work are the following. First, in Sec. 2, we design a conditional meta-learning approach to infer a linear representation that is tuned to the task at hand. Second, in Sec. 3, we formally characterize circumstances under which our conditional framework brings advantage with respect to the standard unconditional approach. In particular, we argue that this is the case when the tasks are organized in different clusters according to the support pattern or linear representation their target vectors' share. Third, in Sec. 4, we design a convex meta-algorithm providing a comparable gain as the number of the tasks it observes increases. In the unconditional setting, the proposed method is able to recover faster rates and it requires to tune one less hyper-parameter with respect to the state-of-the-art unconditional methods. Finally, in Sec. 5, we present numerical experiments supporting our theoretical claims. We conclude our work in Sec. 6 and we postpone the missing proofs to the supplementary material.

2 Conditional representation learning

In this section we introduce our conditional meta-learning setting for representation learning. Then, we proceed to identify the differences with respect to (with respect to) the standard unconditional counterpart. We begin our overview by first introducing the class of inner learning algorithms we use in this work.

Within-task algorithms. We consider the standard linear supervised learning setting over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$ input and output spaces, respectively. We denote by $\mathcal{P}(\mathcal{Z})$ the set of probability distributions (tasks) over $\mathcal{Z}$. For any task $\mu \in \mathcal{P}(\mathcal{Z})$ and a given loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we aim at finding a weight vector $w_\mu \in \mathbb{R}^d$ minimizing the *expected risk*

$$\min_{w \in \mathbb{R}^d} \mathcal{R}_\mu(w) \qquad \mathcal{R}_\mu(w) = \mathbb{E}_{(x,y)\sim\mu}\, \ell\big(\langle x, w\rangle, y\big), \tag{1}$$

where, $\langle \cdot, \cdot \rangle$ represents the Euclidean product in $\mathbb{R}^d$. In practice, μ is only partially observed through a dataset $Z = (x_i, y_i)_{i=1}^n \sim \mu^n$, namely, a collection of n identically independently distributed (i.i.d.) points sampled from μ. Thus, the goal becomes to use a learning algorithm in order to estimate a candidate weight vector with a small expected risk converging to the ideal $\mathcal{R}_\mu(w_\mu)$ as the sample size n grows. Specifically, in this work we will consider as candidate estimators, the family of regularized empirical risk minimizers for linear feature learning [3]. Formally, denoting by $\mathcal{D} = \bigcup_{n\in\mathbb{N}} \mathcal{Z}^n$ the space of all datasets on $\mathcal{Z}$, for a given $\theta \in \Theta$ in $\Theta = \mathbb{S}_+^d$ the set of positive definite $d \times d$ matrices, we will consider the following learning algorithms $A(\theta, \cdot) : \mathcal{D} \to \mathbb{R}^d$:

$$A(\theta, Z) = \operatorname*{argmin}_{w \in \mathrm{Ran}(\theta) \subset \mathbb{R}^d} \mathcal{R}_{Z,\theta}(w), \qquad R_{Z,\theta}(w) = \frac{1}{n}\sum_{i=1}^n \ell(\langle x_i, w\rangle, y_i) + \frac{1}{2}\big\langle w, \theta^\dagger w\big\rangle \tag{2}$$

where $\mathrm{Ran}(\theta)$ denotes the range of θ. Here $\theta^\dagger$ denotes the pseudoinverse of θ. Throughout this work we will denote by $\mathcal{R}_Z(\cdot) = 1/n \sum_{i=1}^n \ell(\langle x_i, \cdot\rangle, y_i)$ the empirical risk associated to Z. Here, θ plays the role of a linear feature representation that is learned during the meta-learning process (see [3]).

Remark 1 (Within-task regularization parameter)**.** *Differently to previous work, see e.g. [15], we do not impose any constraints on the trace of θ (e.g. $\mathrm{Tr}(\theta) \leq 1$). This allows us to absorb the regularization parameter λ typically used to control $\lambda \langle w, \theta^\dagger w \rangle$ in θ. As we will discuss later, this choice reduces the number of hyper-parameter to tune and it allows to enjoy faster learning rates.*

Remark 2 (Online variant of Eq. (2))**.** *Paying additional negligible logarithmic factors, our analysis and results extend also to the setting in which the minimizer in Eq. (2) is replaced by a pre-conditioned variant of online gradient descent on $\mathcal{R}_{Z,\theta}$ with starting point $w_0 = 0$ and appropriate step size:*

$$A(\theta, Z) = \frac{1}{n} \sum_{i=1}^{n} w_i, \quad w_{i+1} = w_i - \frac{\theta p_i}{i}, \quad p_i = s_i x_i + \theta^\dagger w_i, \quad s_i \in \partial \ell(\cdot, y_i)(\langle x_i, w_i \rangle). \tag{3}$$

Unconditional Meta-Learning. The standard unconditional meta-learning setting assumes there exist a meta-distribution $\rho \in \mathcal{P}(\mathcal{M})$ – also called *environment* in [6] – over a family $\mathcal{M} \subseteq \mathcal{P}(\mathcal{Z})$ of distributions (tasks) μ and it aims at selecting an inner algorithm in the family above that is well suited to solve tasks μ sampled from ρ. This target can be reformulated as finding a linear representation $\theta_\rho \in \Theta$ such that the corresponding algorithm $A(\theta_\rho, \cdot)$ minimizes the *transfer risk*

$$\min_{\theta \in \Theta} \mathcal{E}_\rho(\theta), \qquad \mathcal{E}_\rho(\theta) = \mathbb{E}_{\mu \sim \rho} \, \mathbb{E}_{Z \sim \mu^n} \, \mathcal{R}_\mu\big(A(\theta, Z)\big). \tag{4}$$

In practice, this stochastic problem is usually tackled by iteratively sampling a task $\mu \sim \rho$ and a corresponding dataset $Z \sim \mu^n$, and, then, performing a step of stochastic gradient descent on an empirical approximation of Eq. (4) computed from Z. This approach has proven effective for instance when the tasks of the environment share a simple common linear representation, see e.g. [18, 5, 22, 15, 16, 12, 17, 9]. However, when a single linear representation is not sufficient for the entire environment of tasks (e.g. multi-clusters), this homogeneous approach is expected to fail. In order to overcome this limitation, some recent works have adopted the following conditional approach to the problem, see e.g. [37, 36, 32, 21, 10, 39, 14].

Conditional Meta-learning. Analogously to [14], we assume that any task $\mu \sim \rho$ is provided of additional side information $s \in \mathcal{S}$. In such a case, we consider the environment ρ as a distribution $\rho \in \mathcal{P}(\mathcal{M}, \mathcal{S})$ over the set $\mathcal{M}$ of tasks and the set $\mathcal{S}$ of possible side information. Moreover, as usual, we assume ρ to decompose in $\rho(\cdot|s)\rho_{\mathcal{S}}(\cdot)$ and $\rho(\cdot|\mu)\rho_{\mathcal{M}}(\cdot)$ the conditional and marginal distributions with respect to $\mathcal{S}$ and $\mathcal{M}$. For instance, we observe that the side information s could contain descriptive features of the associated task, for example attributes in collaborative filtering [1], or additional information about the users in recommendation systems [19]). Moreover s could be formed by a portion of the dataset sampled from μ (see [37, 14]). Conditional meta-learning leverages this additional side information in order to adapt (or condition) the linear representation $\theta \in \Theta$ on the associated task at hand, by learning a linear-representation-valued function τ solving the problem

$$\min_{\tau \in \mathcal{T}} \mathcal{E}_\rho(\tau), \qquad \mathcal{E}_\rho(\tau) = \mathbb{E}_{(\mu,s) \sim \rho} \mathbb{E}_{Z \sim \mu^n} \mathcal{R}_\mu(A(\tau(s), Z)) \tag{5}$$

over the space $\mathcal{T}$ of measurable functions $\tau : \mathcal{S} \to \Theta$. Notice that we retrieve the unconditional meta-learning problem in Eq. (4) if we restrict Eq. (5) to the set of functions $\mathcal{T}^{\mathrm{const}} = \{\tau \mid \tau(\cdot) \equiv \theta, \ \theta \in \Theta\}$, mapping all the side information into the same constant linear representation.

In the next section, we will investigate the theoretical advantages of adopting such a conditional perspective and, then, we will introduce a convex meta-algorithm to tackle Eq. (5).

3 The advantage of conditional representation learning

In order to characterize the behavior of the optimal solution of Eq. (5) and to investigate the potential advantage of conditional meta-learning, we analyze the generalization properties of a given conditioning function τ. Formally, we compare the error $\mathcal{E}_\rho(\tau)$ with respect to the optimal minimum risk

$$\mathcal{E}^*_\rho = \mathbb{E}_{\mu \sim \rho} \, \mathcal{R}_\mu(w_\mu) \qquad w_\mu = \operatorname*{argmin}_{w \in \mathbb{R}^d} \mathcal{R}_\mu(w). \tag{6}$$

In order to do this, we first need to introduce the following standard assumptions used also in previous literature. Throughout this work we will denote by $\cdot^\top$ the standard transposition operation.

Assumption 1. *Let ℓ be a convex and L-Lipschitz loss function in the first argument. Additionally, there exist $R > 0$ such that $\|x\| \leq R$ for any $x \in \mathcal{X}$.*

Theorem 1 (Excess risk with generic conditioning function τ)**.** *Let Asm. 1 hold. For any $s \sim \rho_{\mathcal{S}}$, introduce the conditional covariance matrices $W(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)} w_\mu w_\mu^\top$ and $C(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)}\mathbb{E}_{x\sim\eta_\mu} xx^\top$, where, η_μ denotes the inputs' marginal distribution of the task μ. Let $\tau \in \mathcal{T}$ such that $\mathrm{Ran}(W(s)) \subseteq \mathrm{Ran}(\tau(s))$ for any $s \sim \rho_{\mathcal{S}}$ and let $A(\tau(s), \cdot)$ be the associated inner algorithm from Eq. (2). Then,*

$$\mathcal{E}_\rho(\tau) - \mathcal{E}_\rho^* \leq \frac{\mathbb{E}_{s\sim\rho_{\mathcal{S}}} \mathrm{Tr}\big(\tau(s)^\dagger W(s)\big)}{2} + \frac{2L^2 \mathbb{E}_{s\sim\rho_{\mathcal{S}}} \mathrm{Tr}\big(\tau(s) C(s)\big)}{n}. \tag{7}$$

Proof. For any $(\mu, s) \sim \rho$, consider the decomposition $\mathcal{E}_\rho(\tau) - \mathcal{E}_\rho^* = \mathbb{E}_{(\mu,s)\sim\rho}\big[\mathrm{B}_{\mu,s} + \mathrm{C}_{\mu,s}\big]$, with

$$\mathrm{B}_{\mu,s} = \mathbb{E}_{Z\sim\mu^n}\Big[\mathcal{R}_\mu(A(\tau(s), Z)) - \mathcal{R}_Z(A(\tau(s), Z))\Big]$$
$$\mathrm{C}_{\mu,s} = \mathbb{E}_{Z\sim\mu^n}\Big[\mathcal{R}_Z(A(\tau(s), Z)) - \mathcal{R}_\mu(w_\mu)\Big].$$

$\mathrm{B}_{\mu,s}$ is the generalization error of the inner algorithm $A(\tau(s), \cdot)$ on the task μ. Hence, applying stability arguments (see Prop. 6 in App. A), we can write $\mathrm{B}_{\mu,s} \leq 2L^2 \mathrm{Tr}\big(\tau(s)\mathbb{E}_{x\sim\eta_\mu} xx^\top\big) n^{-1}$. Regarding the term $\mathrm{C}_{\mu,s}$, for any conditioning function τ such that $w_\mu \in \mathrm{Ran}(\tau(s))$, we can write $\mathrm{C}_{\mu,s} \leq \mathbb{E}_{Z\sim\mu^n}\Big[\mathcal{R}_{Z,\tau(s)}(w_\mu) - \mathcal{R}_\mu(w_\mu)\Big] = 2^{-1}\mathrm{Tr}\big(\tau(s)^\dagger w_\mu w_\mu^\top\big)$, where, the inequality exploits the definition of the algorithm in Eq. (2) as minimum of the regularized empirical risk. The desired statement follows by combining the two bounds above and rewriting $\mathbb{E}_{(\mu,s)\sim\rho} = \mathbb{E}_{s\sim\rho_{\mathcal{S}}}\mathbb{E}_{\mu\sim\rho(\cdot|s)}$. □

Thm. 1 suggests that the conditioning function τ_ρ minimizing the right hand side of Eq. (7) is a good candidate to solve the meta-learning problem. The following result explores this question by showing that such a minimizer admits a closed form solution. The proof is reported in App. B. In the following, we will denote by $\|\cdot\|_F$ and $\|\cdot\|_*$ the Frobenius and trace norm of a matrix, respectively.

Proposition 2 (Best conditioning function in hindsight)**.** *The conditioning function minimizer and the minimum of the bound presented in Thm. 1 over the set $\{\tau \in \mathcal{T} \mid \mathrm{Ran}(W(s)) \subseteq \mathrm{Ran}(\tau(s)),\ \rho_{\mathcal{S}}$-almost surely$\}$, are respectively*

$$\tau_\rho(s) = (2L)^{-1}\, n^{1/2}\, C(s)^{\dagger/2}(C(s)^{1/2} W(s) C(s)^{1/2})^{1/2} C(s)^{\dagger/2}$$
$$\mathcal{E}_\rho(\tau_\rho) - \mathcal{E}_\rho^* \leq 2L \mathbb{E}_{s\sim\rho_{\mathcal{S}}} \big\| W(s)^{1/2} C(s)^{1/2} \big\|_* n^{-1/2}. \tag{8}$$

We observe that, in comparison to [14], the numerator term in the bound above describes a different kind of tasks' similarity assumption: the conditional variance term $\mathbb{E}_{(\mu,s)\sim\rho}\|w_\mu - \mathbb{E}_{\mu\sim\rho(\cdot|s)} w_\mu\|^2$ present in [14] is now substituted by the trace norm term $\mathbb{E}_{s\sim\rho_{\mathcal{S}}} \big\| W(s)^{1/2} C(s)^{1/2} \big\|_*$ above. Additionally, the bound above allows us to quantify the benefits of adopting the conditional feature learning strategy.

Conditional vs. unconditional Meta-Learning. Applying Prop. 2 to $\mathcal{T}^{\mathrm{const}}$, we obtain the optimal (constant) meta-parameter and the corresponding excess risk bound for unconditional meta-learning

$$\tau \equiv \theta_\rho = (2L)^{-1}\, n^{1/2}\, C_\rho^{\dagger/2}(C_\rho^{1/2} W_\rho C_\rho^{1/2})^{1/2} C_\rho^{\dagger/2}$$
$$\mathcal{E}_\rho(\theta_\rho) - \mathcal{E}_\rho^* \leq 2L \big\| W_\rho^{1/2} C_\rho^{1/2} \big\|_*\, n^{-1/2} \tag{9}$$

with unconditional covariance matrices $W_\rho = \mathbb{E}_{\mu\sim\rho} w_\mu w_\mu^\top$ and $C_\rho = \mathbb{E}_{\mu\sim\rho}\mathbb{E}_{x\sim\eta_\mu} xx^\top$. We observe that in the previous literature [13, 15] the authors restricted the unconditional problem over the smaller class of linear representation $\hat{\Theta} = \{\theta \in \mathbb{S}_+^d : \mathrm{Ran}(W_\rho) \subseteq \mathrm{Ran}(\theta), \mathrm{Tr}(\theta) \leq 1\}$ and they considered as the best unconditional representation, the matrix minimizing only a part of the previous bound, namely,

$$\hat{\theta}_\rho = \operatorname*{argmin}_{\theta\in\hat{\Theta}} \mathrm{Tr}\big(\theta^\dagger W_\rho\big) = W_\rho^{1/2}\big(\mathrm{Tr}\big(W_\rho^{1/2}\big)\big)^{-1}. \tag{10}$$

On the other hand, the unconditional oracle we introduce above in Eq. (9) allows us to recover a tighter bound which is able to recover the best performance between independent task learning (ITL) and the oracle considered in previous literature [15]. Indeed, by exploiting the duality between the trace norm $\|\cdot\|_*$ and the operator norm $\|\cdot\|_\infty$ of a matrix, we can upper bound the right-side-term in Eq. (9) by the quantity

$$2L \min\Big\{ \big\|W_\rho^{1/2}\big\|_* \big\|C_\rho^{1/2}\big\|_\infty, \big\|W_\rho^{1/2}\big\|_F \big\|C_\rho^{1/2}\big\|_F \Big\}\, n^{-1/2},$$

namely, the minimum between the bound for independent task learning and the bound for unconditional oracle obtained by previous authors. Notice that the unconditional quantity in Eq. (9) is always bigger than the conditional quantity in Eq. (8), since Eq. (9) coincides with the minimum over a smaller class of function. In order to quantify the gap between these two quantities – namely, the advantage in using the conditional approach with respect to the unconditional one – we have to compare the term $\left\|W_\rho^{1/2} C_\rho^{1/2}\right\|_*$ with the term $\mathbb{E}_{s\sim\rho_\mathcal{S}} \left\|C(s)^{1/2} W(s)^{1/2}\right\|_*$.

We report below a setting that can be considered illustrative for many real-world scenarios in which such a gap in performance is significant.

Example 1 (Clusters)**.** *Let $\mathcal{S} = \mathbb{R}^q$ be the side information space, for some integer $q > 0$. Let ρ be such that the side information marginal distribution $\rho_\mathcal{S}$ is given by a uniform mixture of m uniform distributions. More precisely, let $\rho_\mathcal{S} = \frac{1}{m}\sum_{i=1}^m \rho_\mathcal{S}^{(i)}$, with $\rho_\mathcal{S}^{(i)} = \mathcal{U}\big(\mathcal{B}(a_i, 1/2)\big)$ the uniform distribution on the ball of radius $1/2$ centered at $a_i \in \mathcal{S}$, characterizing the cluster i. For a given side information s, a task $\mu \sim \rho(\cdot|s)$ is sampled such that: 1) its inputs' marginal η_μ is a distribution with constant covariance matrix $C(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)}\mathbb{E}_{x\sim\eta_\mu} xx^\top = C$, for some $C \in \mathbb{S}_+^d$, 2) w_μ is sampled from a distribution with conditional covariance matrix $W(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)} w_\mu w_\mu^\top$, with $W(s)$ such that $(C^{1/2}W(s)C^{1/2})(C^{1/2}W(p)C^{1/2}) = 0$ if $s \neq p$. Then, $\mathbb{E}_{s\sim\rho_\mathcal{S}} \left\|C(s)^{1/2} W(s)^{1/2}\right\|_* = \frac{1}{\sqrt{m}} \left\|W_\rho^{1/2} C_\rho^{1/2}\right\|_*$.*

The inequality above tells us that, in the setting of Ex. 1, the conditional approach gains a $\sqrt{m}$ factor in comparison to the unconditional approach. Therefore, the larger the number of clusters is, the more pronounced the advantage of conditional approach with respect to the unconditional one will be. The $\sqrt{m}$ gain factor follows from the fact that the weight vectors w_μ sampled from the different clusters share disjoint supports (they share orthogonal representations). This allows us to rewrite the overall clusters weight vectors' covariance as the average of the intra clusters weight vectors' covariances. The $\sqrt{m}$ term comes from this rewriting and the quadratic behavior of the covariance matrix. We refer to App. C for more details and the deduction. We also observe that a particular case of the setting above could be that one in which $q = 1$ and the side information are *noisy* observations of the index of the cluster the tasks belong to. In our experiments, in Sec. 5, we consider a more interesting and realistic variant of the setting above, in which we will use as task's side information a training dataset sampled from that task. In the next section, we introduce a convex meta-algorithm mimicking this advantage also in practice.

4 Conditional representation Meta-Learning algorithm

To tackle conditional meta-learning in practice we consider a parametrization where the conditioning functions that are modeled with respect to a given feature map $\Phi : \mathcal{S} \to \mathbb{R}^k$ (with $k \in \mathbb{N}$) on the side information space. In other words, we consider $\tau : \mathcal{S} \to \mathbb{S}_+^d$,

$$\tau(\cdot) = \big(M\Phi(\cdot)\big)^\top M\Phi(\cdot) + C, \tag{11}$$

for some tensor $M \in \mathbb{R}^{p\times d\times k}$ ($p \in \mathbb{N}$) and matrix $C \in \mathbb{S}_+^d$. By construction, the above parametrization guarantees us to learn functions taking values in the set of positive semi-definite matrices. However, directly addressing the meta-learning problem poses two issues: first, dealing with tensorial structures might become computationally challenging in practice and second, such parametrization is quadratic in M and would lead to a non-convex optimization functional in practice. To tackle this issue, the following results shows that we can rewrite the conditioning function in the form of Eq. (11) by using a matrix in $\mathbb{S}_+^{dk}$. This will allows us to implement our method working with matrices in $\mathbb{S}_+^{dk}$, instead of tensors in $\mathbb{R}^{p\times d\times k}$. Throughout this work, we will denote by $\otimes$ the Kronecker product.

Proposition 3 (Matricial re-formulation of $\tau_M(s)$)**.** *Let τ be as in Eq. (11). Then,*

$$\tau(s) = \big(I_d \otimes \Phi(s)^\top\big) H_M \big(I_d \otimes \Phi(s)\big) + C, \tag{12}$$

where I_d is the identity in $\mathbb{R}^{d\times d}$ and H_M is the matrix in $\mathbb{R}^{dk\times dk}$ defined by the entries

$$\big(H_M\big)_{(i-1)k+h,(j-1)k+z} = \big\langle M(:,i,h), M(:,j,z)\big\rangle, \quad i,j = 1,\dots,d, \quad h,z = 1,\dots,k.$$

The arguments above motivate us to consider the following set of conditioning functions:

$$\mathcal{T}_\Phi = \left\{ \tau(\cdot) = \big(I_d \otimes \Phi(\cdot)^\top\big) H \big(I_d \otimes \Phi(\cdot)\big) + C \;\middle|\; \text{such that } H \in \mathbb{S}_+^{dk}, C \in \mathbb{S}_+^d \right\}. \tag{13}$$

To highlight the dependency of a function $\tau \in \mathcal{T}_\Phi$ with respect to its parameter H and C, we will denote $\tau = \tau_{H,C}$. Evidently, $\mathcal{T}_\Phi$ contains the space of all unconditional estimators $\mathcal{T}^{\text{const}}$. We consider $\mathcal{T}_\Phi$ equipped with the canonical norm $\|\tau_{H,C}\|^2 = \|(H,C)\|_F^2 = \|H\|_F^2 + \|C\|_F^2$, where, recall, $\|\cdot\|_F$ denotes the Frobenius norm. The following two standard assumptions will allow us to design and analyse our method.

Assumption 2. *The optimal function τ_ρ belongs to $\mathcal{T}_\Phi$, namely there exist $H_\rho \in \mathbb{S}_+^{dk}$ and $C_\rho \in \mathbb{S}_+^d$, such that $\tau_\rho(\cdot) = \tau_{H_\rho,C_\rho}(\cdot) = \big(I_d \otimes \Phi(\cdot)^\top\big) H_\rho \big(I_d \otimes \Phi(\cdot)\big) + C_\rho$.*

Assumption 3. *There exists $K > 0$ such that $\|\Phi(s)\| \leq K$ for any $s \in \mathcal{S}$.*

Asm. 2, known as well-specified setting assumption (see e.g. [33]), is a standard assumption in learning theory and it allows us to restrict the conditional meta-learning problem in Eq. (5) to $\mathcal{T}_\Phi$, rather than to the entire space $\mathcal{T}$ of measurable functions. Asm. 3 ensures that the meta-objective is Lipschitz (see below).

The convex surrogate problem. We start from observing that, exploiting the generalization properties of the within-task algorithm (see Prop. 6 in App. A), we can write the following

$$\mathcal{E}_\rho(\tau) \leq \mathbb{E}_{(\mu,s)\sim\rho}\, \mathbb{E}_{Z\sim\mu^n}\, F_Z(\tau(s)), \qquad F_Z(\theta) = \mathcal{R}_{Z,\theta}(A(\theta, Z)) + \frac{2L^2}{n} \mathrm{Tr}\Big(\theta \frac{X^\top X}{n}\Big)$$

where $X \in \mathbb{R}^{n\times d}$ is the matrix with the inputs vectors $(x_i)_{i=1}^n$ as rows. The inequality above suggests us to introduce the surrogate problem

$$\min_{\tau\in\mathcal{T}} \hat{\mathcal{E}}_\rho(\tau), \qquad \hat{\mathcal{E}}_\rho(\tau) = \mathbb{E}_{(\mu,s)\sim\rho}\, \mathbb{E}_{Z\sim\mu^n}\, F_Z(\tau(s)). \tag{14}$$

We stress that the surrogate problem we take here is different from the one considered in previous work [12, 15, 14, 9], where the authors considered as meta-objective only a part of the function above, namely, $\mathbb{E}_{(\mu,s)\sim\rho}\, \mathbb{E}_{Z\sim\mu^n} \big[\mathcal{R}_{Z,\tau(s)}(A(\tau(s), Z))\big]$. As we will see in the following, such a choice is more appropriate for the problem at hand, since, differently from the meta-objective used in previous literature, it will allow us to develop a conditional meta-learning method that is theoretically grounded also for linear representation learning.

Exploiting Asm. 2, the surrogate problem in Eq. (14) can be restricted to the class of linear functions $\mathcal{T}_\Phi$ in Eq. (13) and it can be rewritten more explicitly as

$$\min_{H\in\mathcal{S}^{dk}, C\in\mathbb{S}_+^d} \mathbb{E}_{(\mu,s)\sim\rho}\, \mathbb{E}_{Z\sim\mu^n}\, \mathcal{L}\big(H, C, s, Z\big), \qquad \mathcal{L}\big(H, C, s, Z\big) = F_Z\big(\tau_{H,C}(s)\big). \tag{15}$$

In the following proposition we outline some useful properties of the meta-loss $\mathcal{L}\big(\cdot,\cdot, s, Z\big)$ introduced above (such as convexity) supporting its choice as surrogate meta-loss.

Proposition 4 (Properties of the surrogate meta-loss $\mathcal{L}$)**.** *For any $Z \in \mathcal{D}$ and $s \in \mathcal{S}$, the function $\mathcal{L}\big(\cdot,\cdot, s, Z\big)$ is convex and one of its subgradients is given, for any $H \in \mathbb{S}_+^{dk}$ and $C \in \mathbb{S}_+^d$, by*

$$\nabla\mathcal{L}\big(H,\cdot, s, Z\big)(C) = \hat{\nabla}, \qquad \nabla\mathcal{L}\big(\cdot, C, s, Z\big)(H) = \big(I_d \otimes \Phi(s)\big)\hat{\nabla}\big(I_d \otimes \Phi(s)^\top\big),$$

where

$$\hat{\nabla} = -\frac{\lambda}{2} \tau_{H,C}(s)^\dagger w_{\tau_{H,C}(s)} w_{\tau_{H,C}(s)}^\top \tau_{H,C}(s)^\dagger + \frac{2L^2 X^\top X}{n^2}.$$

Moreover, by Asm. 1 and Asm. 3,

$$\big\|\nabla\mathcal{L}\big(\cdot,\cdot, s, Z\big)(H, C)\big\|_F \leq (1 + K^2)(LR)^2\big(2^{-1} + 2n^{-1}\big).$$

The proof of Prop. 4 is reported in App. D.2. It follows from combining results from [15] with the composition of the linear parametrization of the functions $\tau_{H,C} \in \mathcal{T}_\Phi$.

The conditional Meta-Learning estimator. The meta-learning strategy we propose consists in applying Stochastic Gradient Descent (SGD) on the surrogate problem in Eq. (15). Such a meta-algorithm is implemented in Alg. 1: we assume to observe a sequence of i.i.d. pairs $(Z_t, s_t)_{t=1}^T$ of training datasets and side information, and at each iteration we update the conditional parameters (H_t, C_t) by performing a step of constant size $\gamma > 0$ in the direction of $-\nabla\mathcal{L}(\cdot,\cdot, s_t, Z_t)(H_t, C_t)$ and a projection step on $\mathbb{S}_+^{dk} \times \mathbb{S}_+^d$. Finally, we output the conditioning function $\tau_{\bar{H},\bar{C}}$ parametrized by $(\bar{H}, \bar{C})$, the average across all the iterates $(H_t, C_t)_{t=1}^T$. The theorem below analyzes the generalization properties of such a conditioning function.

Algorithm 1 Meta-algorithm, SGD on Eq. (15)

Input $\gamma > 0$ meta-step size, $H_0 \in \mathbb{S}_+^{dk}$, $C_0 \in \mathbb{S}_+^d$
Initialization $H_1 = H_0 \in \mathbb{S}_+^{dk}$, $C = C_0 \in \mathbb{S}_+^d$
For $t = 1$ to T
Receive $(\mu_t, s_t) \sim \rho$ and $Z_t \sim \mu_t^n$
Let $\theta_t = \big(I_d \otimes \Phi(s_t)\big) H_t \big(I_d \otimes \Phi(s_t)^\top\big) + C_t$ and compute $w_{\theta_t} = A(\theta_t, Z_t)$ by Eq. (2)
Compute $\nabla \mathcal{L}(\cdot, \cdot, s_t, Z_t)(H_t, C_t)$ as in Prop. 4 with w_{θ_t}
Update $(H_{t+1}, C_{t+1}) = \mathrm{proj}_\Theta\big((H_t, C_t) - \gamma \nabla \mathcal{L}(\cdot, \cdot, s_t, Z_t)(H_t, C_t)\big)$
Return $\bar{H} = \dfrac{1}{T}\sum_{t=1}^T H_t$, $\bar{C} = \dfrac{1}{T}\sum_{t=1}^T C_t$

Theorem 5 (Excess risk bound for the conditioning function returned by Alg. 1)**.** *Let Asm. 1 and Asm. 3 hold. For any $s \sim \rho_{\mathcal{S}}$, recall the conditional covariance matrices $W(s)$ and $C(s)$ introduced in Thm. 1. Let $\tau_{H,C}$ be a fixed function in $\mathcal{T}_\Phi$ such that $\mathrm{Ran}(W(s)) \subseteq \mathrm{Ran}(\tau_{H,C}(s))$ for any $s \sim \rho_{\mathcal{S}}$. Let $\bar{H}$ and $\bar{C}$ be the outputs of Alg. 1 applied to a sequence $(Z_t, s_t)_{t=1}^T$ of i.i.d. pairs sampled from ρ with an appropriate meta-step size γ. Then, in expectation with respect to the sampling of $(Z_t, s_t)_{t=1}^T$,*

$$\mathbb{E}\, \mathcal{E}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* \;\le\; \frac{\mathbb{E}_{s\sim\rho_{\mathcal{S}}} \mathrm{Tr}\big(\tau_{H,C}(s)^\dagger W(s)\big)}{2} + \frac{2L^2 \mathbb{E}_{s\sim\rho_{\mathcal{S}}} \mathrm{Tr}\big(\tau_{H,C}(s) C(s)\big)}{n} + \Big(\frac{1}{2} + \frac{2}{n}\Big) \frac{(1+K^2)(LR)^2 \, \|(H - H_0, C - C_0)\|_F}{\sqrt{T}}.$$

Proof (Sketch). The detailed proof is reported in App. D.4. Exploiting the fact that, for any $\tau \in \mathcal{T}$, $\mathcal{E}_\rho(\tau) \le \hat{\mathcal{E}}_\rho(\tau)$ and adding $\pm \hat{\mathcal{E}}_\rho(\tau_{H,C})$, we can write the following

$$\begin{gathered} \mathbb{E}_{\mathbf{Z}}\, \mathcal{E}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* \le \mathrm{A}(\tau_{H,C}) + \mathrm{B}(\tau_{H,C}) \\ \mathrm{A}(\tau_{H,C}) = \mathbb{E}_{\mathbf{Z}}\, \hat{\mathcal{E}}_\rho(\tau_{\bar{H},\bar{C}}) - \hat{\mathcal{E}}_\rho(\tau_{H,C}) \qquad \mathrm{B}(\tau_{H,C}) = \hat{\mathcal{E}}_\rho(\tau_{H,C}) - \mathcal{E}_\rho^*. \end{gathered} \tag{16}$$

The term $\mathrm{A}(\tau_{H,C})$ can be controlled according to the convergence properties of the meta-algorithm in Alg. 1 as described in Prop. 12. Regarding the term $\mathrm{B}(\tau_{H,C})$, exploiting the definition of the within-task algorithm in Eq. (2) as minimum, for any $\tau \in \mathcal{T}$ such that $\mathrm{Ran}(\mathbb{E}_{\mu\sim\rho(\cdot|s)} w_\mu w_\mu^\top) \subseteq \mathrm{Ran}(\tau(s))$ for any $s \sim \rho_{\mathcal{S}}$, we can rewrite

$$\mathrm{B}(\tau) \le \frac{\mathbb{E}_{(\mu,s)\sim\rho} \mathrm{Tr}\big(\tau(s)^\dagger w_\mu w_\mu^\top\big)}{2} + \frac{2L^2 \mathbb{E}_{(\mu,s)\sim\rho} \mathrm{Tr}\big(\tau(s) \mathbb{E}_{x\sim\eta_\mu} x x^\top\big)}{n}.$$

The desired statement then derives from combining the two parts above and optimizing with respect to γ. □

Remark 3 (Online variant of Eq. (2))**.** *Similarly to the bias regularization framework in [14], for the online inner family in Rem. 2, we approximate the meta-subgradient in Prop. 4 by replacing the batch minimizer $A(\tau_{H,C}(s), Z)$ in Eq. (2) with the last iterate of the online algorithm in Eq. (3).*

Proposed vs. optimal conditioning function. Specializing the bound in Thm. 5 to the best conditioning function τ_ρ in Prop. 2, thanks to Asm. 2, we get, up to contants, the following bound for our estimator,

$$\mathbb{E}_{s\sim\rho_{\mathcal{S}}} \big\|W(s)^{1/2} C(s)^{1/2}\big\|_* \, n^{-1/2} + \|(H_\rho - H_0, C_\rho - C_0)\|_F \, T^{-1/2}.$$

From such a bound, we can state that our proposed meta-algorithm achieves comparable performance to the best conditioning function τ_ρ in hindsight, when the number of observed tasks is sufficiently large. Moreover, recalling the unconditional oracle $\hat{\theta}_\rho$ in Eq. (10) used in previous literature, regarding the second term vanishing with T, we observe that our conditional meta-learning approach incurs a cost of $\|(H_\rho - H_0, C_\rho - C_0)\|_F T^{-1/2}$ as opposed to the cost of $\|\hat{\theta}_\rho - \theta_0\| T^{-1/4}$ associated to state-of-the-art unconditional meta-learning approaches (see [15, 5, 22, 9]). Thus, our conditional approach presents a faster convergence rate with respect to T than such unconditional methods, but a complexity term that is expected to be larger due to the larger complexity of the class of functions we

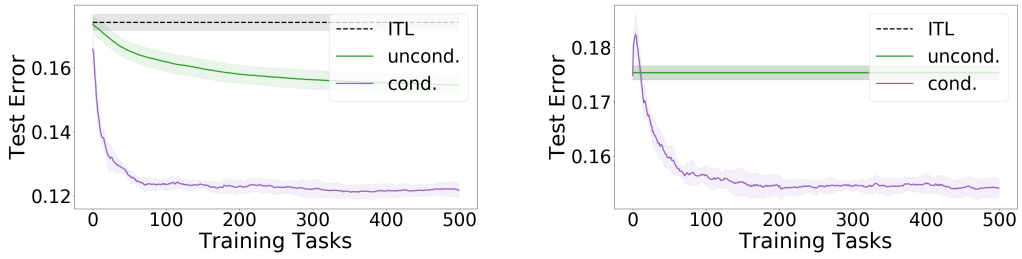

Figure 1: Mean test error (5 generations) on synthetic data. 2 (Left) and 6 (Right) clusters.

are working with. Such a faster rate with respect to T is essentially due to our formulation of the problem on the entire set of positive-semidefinite matrices (with no trace constraints). This in fact allows us to incorporate the within-task regularization parameter λ directly in the linear representation and to gain a $\sqrt{T}$ order that was lost in previous literature when tuning with respect to the parameter λ. At the same time, this allows us to develop also a method requiring to tune just one hyper-parameter, while previous unconditional approaches requires to tune two hyper-parameters.

Comparison to unconditional Meta-Learning. Specializing Thm. 5 to the best unconditional estimator $\tau_{H,C} \equiv \theta_\rho$ we introduced in Eq. (9), the bound for our estimator becomes, up to constants,

$$\left\|W_\rho^{1/2} C_\rho^{1/2}\right\|_* n^{-1/2} + \|\theta_\rho - C_0\| \, T^{-1/2}.$$

From the bound above, we can conclude that the conditional approach provides, at least, the same guarantees as its unconditional counterpart. Moreover, we stress again that the bound above presents a faster rate with respect to T in comparison to the state-of-the-art unconditional methods.

5 Experiments

We now present preliminary experiments in which we compare the proposed conditional meta-learning approach in Alg. 1 (cond.) with the unconditional counterpart (uncond.) and solving the tasks independently (ITL, namely, running the inner algorithm separately across the tasks with the constant linear representation $\theta = I_d \in \mathbb{S}_+^d$). We considered regression problems and we evaluated the errors by ℓ the absolute loss. We implemented the online variant of the within-task algorithm introduced in Eq. (3). The hyper-parameter γ was chosen by (meta-)cross validation on separate $T_{\rm tr}$, $T_{\rm va}$ and $T_{\rm te}$ respectively meta-train, -validation and -test sets. Each task is provided with a training dataset $Z_{\rm tr}$ of $n_{\rm tr}$ points and a test dataset $Z_{\rm te}$ of $n_{\rm te}$ points used to evaluate the performance of the within-tasks algorithm. In App. E we report the details of this process in our experiments.

Synthetic clusters. We considered two variants of the setting described in Ex. 1 with side information corresponding to the training datasets $Z_{\rm tr}$ associated to each task. In both settings, we sampled $T_{\rm tot} = 900$ tasks from a uniform mixture of m clusters. For each task μ, we generated the target vector $w_\mu \in \mathbb{R}^d$ with $d = 20$ as $w_\mu = P(j_\mu)\tilde{w}_\mu$, where, $j_\mu \in \{1, \dots, m\}$ denotes the cluster from which the task μ was sampled and with the components of $\tilde{w}_\mu \in \mathbb{R}^{d/(10)}$ sampled from the Gaussian distribution $\mathcal{G}(0,1)$ and then $\tilde{w}_\mu$ normalized to have unit norm, with $P(j_\mu) \in \mathbb{R}^{d \times d/(10)}$ a matrix with orthonormal columns. We then generated the corresponding dataset $(x_i, y_i)_{i=1}^{n_{\rm tot}}$ with $n_{\rm tot} = 80$ according to the linear equation $y = \langle x, w_\mu \rangle + \epsilon$, with x sampled uniformly on the unit sphere in $\mathbb{R}^d$ and ϵ sampled from a Gaussian distribution, $\epsilon \sim \mathcal{G}(0, 0.1)$. In this setting, the operator norm of the inputs' covariance matrix is small (equal to $1/d$) and the weight vectors' covariance matrix of each single cluster is low-rank (its rank is $d/(10) = 2$). We implemented our conditional method using the feature map $\Phi : \mathcal{D} \to \mathbb{R}^{2d}$ defined by $\Phi(Z) = \frac{1}{n_{\rm tr}} \sum_{i=1}^{n_{\rm tr}} \phi(z_i)$, with $\phi(z_i) = \mathrm{vec}\big(x_i (y_i, 1)^\top\big)$, where, for any matrix $A = [a_1, a_2] \in \mathbb{R}^{d \times 2}$ with columns $a_1, a_2 \in \mathbb{R}^d$, $\mathrm{vec}(A) = (a_1, a_2)^\top \in \mathbb{R}^{2d}$. In Fig. 1, we report the results we got on an environment of tasks generated as above with $m = 2$ (Left) and $m = 6$ (Right) clusters, respectively. As we can see, when the clusters are two, the unconditional approach outperforms ITL (as predicted from previous literature), but the unconditional method is in turn outperformed by our conditional counterpart. When the number of clusters raises to six, the performance of unconditional meta-learning degrades to the same performance of ITL,

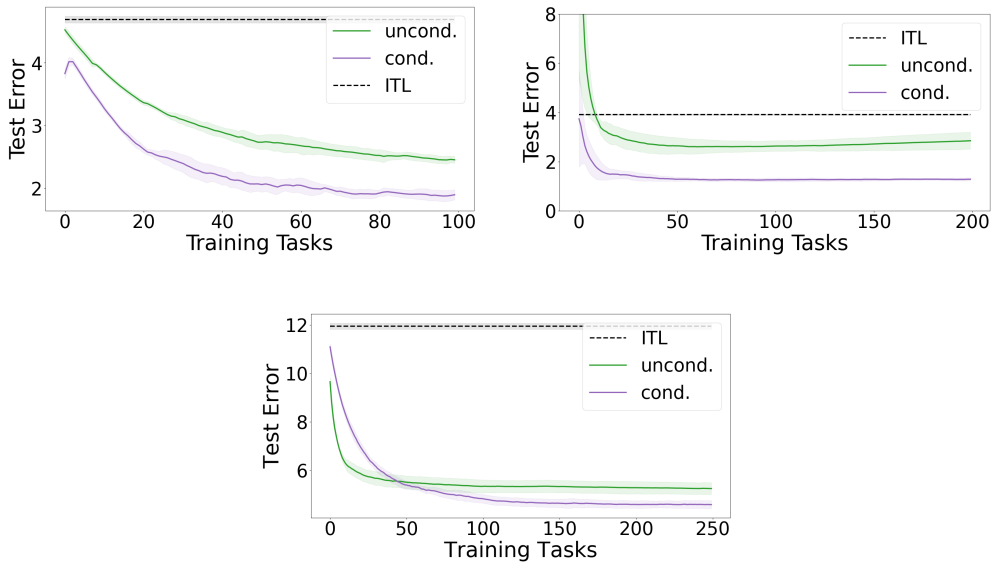

Figure 2: Mean test error (5 splits) on Lenk (Top-Left), Movielens-100k (Top-Right), Jester-1 (Bottom) dataset.

while conditional meta-learning outperforms both methods. Summarizing, the more the heterogeneity of the environment (number of clusters) is significant, the more the conditional approach brings advantage with respect to the unconditional one. This is in line with Ex. 1.

Real datasets. We tested the performance of the methods also on the regression problem on the computer survey data from [23] (see also [28]). $T_{\text{tot}} = 180$ people (tasks) rated the likelihood of purchasing one of $n_{\text{tot}} = 20$ computers. The input represents $d = 13$ computers' characteristics and the label is a rate in $\{0, \dots, 10\}$. In this case, we used as side information the training datapoints $Z = (z_i)_{i=1}^{n_{\text{tr}}}$ and the feature map $\Phi : \mathcal{D} \to \mathbb{R}^{d+1}$ defined by $\Phi(Z) = w_Z$, with w_Z the solution of Tikhonov regularization with the squared loss, namely, the vector satisfying $(\hat{X}^\top \hat{X} + I_{d+1}) w_Z = \hat{X}^\top y$, where, $\hat{X} \in \mathbb{R}^{(d+1)\times n}$ is the matrix obtained by adding to the matrix $X \in \mathbb{R}^{n\times d}$ one column of ones at the end. Fig. 2 (Top-Left) shows that also in this case, the unconditional approach outperforms ITL, but the performance of its conditional counterpart is much better.

We also tested the performance of the methods on the Movielens-100k and Jester-1 real-world datasets, containing ratings of users (tasks) to movies and jokes (points), respectively. Recommendation system settings with d items can be interpreted within the meta-learning setting by considering each data point (x, y) to have input $x \in \mathbb{R}^d$ to be the one-hot encoding of the current item to be rated (e.g. a movie or a joke) and $y \in \mathbb{R}$ the corresponding score, see e.g. [12] for more details. We restricted the original dataset to the $n_{\text{tot}} = 20$ most voted movies/jokes (as a consequence, by formulation, $d = 20$). We guaranteed each user voted at least 5 movies/jokes, which led to a total of $T_{\text{tot}} = 400/450$ tasks (i.e. users). In both cases, we used as side information the training datapoints $Z = (z_i)_{i=1}^{n_{\text{tr}}}$. For the Movielens-100k dataset we used the same feature map described for the synthetic clusters experiments in Fig. 1. For the Jester-1 dataset, let M and m denote the maximum and minimum rating value that can be assigned to a joke. We adopted the feature map $\Phi : \mathcal{D} \to \mathbb{R}^{2d+1}$ such that, for any dataset $Z = (x_i, y_i)_{i=1}^n$, we have $\Phi(Z) = \big(\text{vec}(\widetilde{\Phi}(Z)); 1\big)$, where vec denotes the vectorization operator (i.e. mapping a matrix in the vector concatenating all its columns) and $\widetilde{\Phi} : Z \to \mathbb{R}^{d\times 2}$ is such that $\widetilde{\Phi}(Z) = \big(\cos(\alpha(Z)), \sin(\alpha(Z))\big) \odot (\sum_{i=1}^n x_i)$ with $\alpha(Z) = \sum_{i=1}^n x_i \left(\frac{\pi}{4} \frac{M - y_i}{M - m}\right)$ and $\odot$ denoting the Hadamard (entry-wise) product broad-casted across both columns. The rationale behind this feature map is to represent as similar vectors those users with similar scores for the same movies. In particular, each item-score pair observed in training is represented as a unitary vector in $\mathbb{R}^2_{++}$, with the angle depending on the score attributed to that item (the vector corresponds to zero if that movie was not observed at the training time). As it can be noticed in Fig. 2 (Top Right and Bottom), the

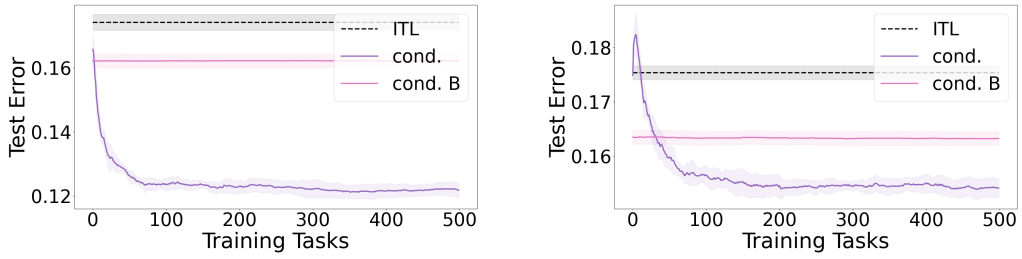

Figure 3: Mean test error (5 generations) of our method and [14] on 2 (Left) and 6 (Right) clusters.

proposed approach performs significantly better than ITL and its unconditional counterpart also on these two benchmarks. This suggests that groups of users might rely each on similar features (but different from those of other groups) to rate an item in the dataset (respectively a movie or a joke).

Comparison with [14]. We conclude the experimental section by comparing the performance of our method with the conditional meta-learning approach for biased regularization proposed in [14]. In that case, the tasks' target vectors are assumed to be all close to a common bias vector rather than sharing the same low-dimensional linear representation, as instead in our method. The representation setting is known to be a more relevant and effective framework in many scenarios in comparison to the bias one (see e.g. [28]). This is confirmed in Fig. 3 where our conditional representation learning method ('cond.') significantly outperforms the one in [14] ('cond. B') in the synthetic settings used in Fig. 1, since the tasks' similarity leveraged by [14] is not appropriate in these cases.

6 Conclusion

We proposed a conditional meta-learning approach aiming at learning a function mapping task's side information into a linear representation that is well suited for the task at hand. We theoretically and experimentally showed that the proposed conditional approach is advantageous with respect to the standard unconditional counterpart when the observed tasks share heterogeneous linear representations. As a consequence of our analysis we also developed a new unconditional meta-learning variant requiring tuning less hyper-parameters and relying on faster rates with respect to state-of-the-art unconditional approaches. We identify two future directions addressing the limitations of our method. A first question left opened is how to design a suitable feature map Φ when the tasks' training data is used as side information. Following [32, 37], we adopted a mean embedding representation. However, given the importance played by such feature map in Thm. 5, it will be worth investigating better alternatives. Secondly, it will be valuable to investigate how to predict non-linear conditioning functions (similarly to e.g. [7, 16, 32]) and use less expensive algorithms to update the positive matrices, such as the Frank-Wolfe algorithm used in [9] for unconditional settings. In this last case, according to our analysis, applying standard convergence rates for Frank-Wolfe algorithm, we expect the meta-learning algorithm based on Frank-Wolfe iteration to incur a slower rate of order $T^{-1/4}$ (instead of $T^{-1/2}$ for the SGD method proposed here) in Thm. 5, paying the computational benefits in terms of statistical performance.

Acknowledgments and Disclosure of Funding

This work was supported in part by SAP SE and EPSRC Grant N. EP/P009069/1. C.C. acknowledges the support of the Royal Society (grant SPREM RGS\R1\201149) and Amazon.com Inc. (Amazon Research Award – ARA). G.D. acknowledges Leonardo SpA for funding her participation to the conference.

References

[1] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(Mar):803–826, 2009.

[2] Andreas Argyriou, Stéphan Clémençon, and Ruocong Zhang. Learning the graph of relations among multiple tasks. 2013.

[3] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[4] Andreas Argyriou, Andreas Maurer, and Massimiliano Pontil. An algorithm for transfer learning in a heterogeneous environment. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer, 2008.

[5] Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433, 2019.

[6] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.*, 12(149–198):3, 2000.

[7] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[8] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[9] Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *Algorithmic Learning Theory*, pages 235–246, 2019.

[10] T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Weighted message passing and minimum energy flow for heterogeneous stochastic block models with side information. *Journal of Machine Learning Research*, 21(11):1–34, 2020.

[11] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[12] Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575, 2019.

[13] Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees. In *Proc. 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[14] Giulia Denevi, Massimiliano Pontil, and Carlo Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, and Massimiliano Pontil. Online-within-online meta-learning. In *Advances in Neural Information Processing Systems*, pages 13089–13099, 2019.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.

[17] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *International Conference on Learning Representations*, 2018.

[18] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930, 2019.

[19] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[20] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752, 2009.

[21] Ghassen Jerfel, Erin Grant, Tom Griffiths, and Katherine A Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. In *Advances in Neural Information Processing Systems*, pages 9119–9130, 2019.

[22] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, pages 5915–5926, 2019.

[23] Peter J Lenk, Wayne S DeSarbo, Paul E Green, and Martin R Young. Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.

[24] Andreas Maurer. Transfer bounds for linear feature learning. *Machine Learning*, 75(3):327–350, 2009.

[25] Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *International Conference on Machine Learning*, 2013.

[26] Andreas Maurer and Massimiliano Pontil. Transfer learning in a heterogeneous environment. In *2012 3rd International Workshop on Cognitive Information Processing (CIP)*, pages 1–6. IEEE, 2012.

[27] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

[28] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. New perspectives on k-support and cluster norms. *Journal of Machine Learning Research*, 17(155):1–38, 2016.

[29] Charles A Micchelli, Jean M Morales, and Massimiliano Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489, 2013.

[30] Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999, 2014.

[31] Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *International Conference on Machine Learning*, pages 1444–1452. PMLR, 2013.

[32] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2018.

[33] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[34] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[35] Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

[36] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Advances in Neural Information Processing Systems*, pages 1–12, 2019.

[37] Ruohan Wang, Yiannis Demiris, and Carlo Ciliberto. A structured prediction approach for conditional meta-learning. *Advances in Neural Information Processing Systems*, 2020.

[38] Kishan Wimalawarne, Masashi Sugiyama, and Ryota Tomioka. Multitask learning meets tensor factorization: task imputation via convex optimization. In *NIPS*, pages 2825–2833. Citeseer, 2014.

[39] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. *arXiv preprint arXiv:1905.05301*, 2019.

[40] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Clustered multi-task learning via alternating structure optimization. *Advances in neural information processing systems*, 2011:702, 2011.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See the abstract and Lines $44 - 54$.
   (b) Did you describe the limitations of your work? [Yes] See Sec. 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work does not imply any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Asm. 1, Asm. 2, Asm. 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sec. 3, Sec. 4 and Supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Sec. 5, App. E and Supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. 5, App. E and Supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Sec. 5, App. E and Supplementary material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See App. E and Supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Fig. 3 we compared with the method in [14] and we cited that work.
   (b) Did you mention the license of the assets? [Yes] See Fig. 3.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We attached our code in the Supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We used benchmark data and we cited the original source where one can find all the details.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data we are using does not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing or conducted research with human subjects.

Appendix

The supplementary material is organized as follows. In App. A we give the bound on the generalization error of the algorithm in Eq. (2) that we used in various proofs. In App. B we report the proof to get the closed form of the best conditioning function τ_ρ outlined in Prop. 2. In App. C we report the proof of the statement in Ex. 1. In App. D, we report the proofs of the statements we used in Sec. 4 in order to prove the expected excess risk bound in Thm. 5 for Alg. 1. Finally, in App. E we report the experimental details we missed in the main body.

A Generalization bound of the within-task algorithm

We now study the generalization error of the within-task algorithm in Eq. (2), i.e. the discrepancy between the (true) risk and the empirical risk of the corresponding estimator. This is done in the following result where we exploit stability arguments, more precisely the so-called hypothesis stability, see [8, Def. 3].

Proposition 6 (Generalization error of the within-task algorithm in Eq. (2))**.** *Let Asm. 1 hold. For a distribution $\mu \sim \rho$, fix a dataset $Z = (x_i, y_i)_{i=1}^n \sim \mu^n$. For any $\theta \in \Theta$, let $w_\theta(Z)$ be the corresponding RERM in Eq. (2) over Z. Then, the following generalization error bound holds for $w_\theta(Z)$:*

$$\mathbb{E}_{Z\sim\mu^n}\left[\mathcal{R}_\mu(w_\theta(Z)) - \mathcal{R}_Z(w_\theta(Z))\right] \leq \frac{2L^2}{n}\operatorname{Tr}\left(\mathbb{E}_{z\sim\mu}\,\theta x x^\top\right). \tag{17}$$

Proof. During this proof, we need to make explicit the dependency of the RERM (Regularized Empirical Risk Minimizer) w_θ in Eq. (2) with respect to the dataset Z. For any $i \in \{1, \dots, n\}$, consider the dataset $Z^{(i)}$, a copy of the original dataset Z in which we exchange the point $z_i = (x_i, y_i)$ with a new i.i.d. point $z_i' = (x_i', y_i')$. For a fixed $\theta \in \Theta$, we analyze how much this perturbation affects the outputs of the RERM algorithm in Eq. (2). In other words, we study the discrepancy between $w_\theta(Z)$ and $w_\theta(Z^{(i)})$. We start from observing that, since by Asm. 1 $\mathcal{R}_{Z,\theta}$ is 1-strongly convex with respect to $\|\cdot\|_\theta = \sqrt{\langle \cdot, \theta^\dagger \cdot \rangle}$, by growth condition and the definition of the RERM algorithm, we can write the following

$$\begin{aligned}
\frac{1}{2}\left\|w_\theta(Z^{(i)}) - w_\theta(Z)\right\|_\theta^2 &\leq \mathcal{R}_{Z,\theta}(w_\theta(Z^{(i)})) - \mathcal{R}_{Z,\theta}(w_\theta(Z)) \\
\frac{1}{2}\left\|w_\theta(Z^{(i)}) - w_\theta(Z)\right\|_\theta^2 &\leq \mathcal{R}_{Z^{(i)},\theta}(w_\theta(Z)) - \mathcal{R}_{Z^{(i)},\theta}(w_\theta(Z^{(i)})).
\end{aligned} \tag{18}$$

Hence, summing the two inequalities above, we get

$$\begin{aligned}
\left\|w_\theta(Z^{(i)}) - w_\theta(Z)\right\|_\theta^2 &\leq \mathcal{R}_{Z,\theta}(w_\theta(Z^{(i)})) - \mathcal{R}_{Z^{(i)},\theta}(w_\theta(Z^{(i)})) + \mathcal{R}_{Z^{(i)},\theta}(w_\theta(Z)) - \mathcal{R}_{Z,\theta}(w_\theta(Z)) \\
&= \frac{\mathrm{A} + \mathrm{B}}{n},
\end{aligned} \tag{19}$$

where we have introduced the terms

$$\begin{aligned}
\mathrm{A} &= \ell(\langle x_i', w_\theta(Z)\rangle, y_i') - \ell(\langle x_i', w_\theta(Z^{(i)})\rangle, y_i') \\
\mathrm{B} &= \ell(\langle x_i, w_\theta(Z^{(i)})\rangle, y_i) - \ell(\langle x_i, w_\theta(Z)\rangle, y_i).
\end{aligned} \tag{20}$$

Now, introducing the subgradients $s_{\theta,i}' \in \partial\ell(\cdot, y_i')(\langle x_i', w_\theta(Z)\rangle)$ and $s_{\theta,i} \in \partial\ell(\cdot, y_i)(\langle x_i, w_\theta(Z^{(i)})\rangle)$ and applying Holder's inequality, we can write

$$\begin{aligned}
\mathrm{A} &\leq \left\langle x_i' s_{\theta,i}', w_\theta(Z) - w_\theta(Z^{(i)})\right\rangle \leq \left\|x_i' s_{\theta,i}'\right\|_{\theta,*}\left\|w_\theta(Z^{(i)}) - w_\theta(Z)\right\|_\theta \\
\mathrm{B} &\leq \left\langle x_i s_{\theta,i}, w_\theta(Z^{(i)}) - w_\theta(Z)\right\rangle \leq \left\|x_i s_{\theta,i}\right\|_{\theta,*}\left\|w_\theta(Z^{(i)}) - w_\theta(Z)\right\|_\theta,
\end{aligned} \tag{21}$$

where $\|\cdot\|_{\theta,*} = \sqrt{\langle \cdot, \theta \cdot \rangle}$ is the dual norm of $\|\cdot\|_\theta$. Combining these last two inequalities with Eq. (19) and simplifying, we get the following

$$\left\|w_\theta(Z^{(i)}) - w_\theta(Z)\right\|_\theta \leq \frac{1}{n}\left(\left\|x_i' s_{\theta,i}'\right\|_{\theta,*} + \left\|x_i s_{\theta,i}\right\|_{\theta,*}\right). \tag{22}$$

Hence, combining the first row in Eq. (21) with Eq. (22), we can write

$$\ell(\langle x_i', w_\theta(Z)\rangle, y_i') - \ell(\langle x_i', w_\theta(Z^{(i)})\rangle, y_i') \leq \frac{1}{n}\Big(\big\|x_i' s_{\theta,i}'\big\|_{\theta,*}^2 + \big\|x_i' s_{\theta,i}'\big\|_{\theta,*}\big\|x_i s_{\theta,i}\big\|_{\theta,*}\Big). \quad (23)$$

Now, taking the expectation with respect to $Z \sim \mu^n$ and $z_i' \sim \mu$ of the left side member above, according to [8, Lemma 7], we get

$$\mathbb{E}_{Z\sim\mu^n}\,\mathbb{E}_{z_i'\sim\mu}\Big[\ell(\langle x_i', w_\theta(Z)\rangle, y_i') - \ell(\langle x_i', w_\theta(Z^{(i)})\rangle, y_i')\Big] = \mathbb{E}_{Z\sim\mu^n}\Big[\mathcal{R}_\mu(w_\theta(Z)) - \mathcal{R}_Z(w_\theta(Z))\Big].$$

Finally, taking the expectation of the right side member, exploiting the fact that the points are i.i.d. according μ, we get

$$\mathbb{E}_{Z\sim\mu^n}\,\mathbb{E}_{z_i'\sim\mu}\,\frac{1}{n}\left(\big\|x_i' s_{\theta,i}'\big\|_{\theta,*}^2 + \big\|x_i' s_{\theta,i}'\big\|_{\theta,*}\big\|x_i s_{\theta,i}\big\|_{\theta,*}\right) \leq \frac{2}{n}\,\mathbb{E}_{Z\sim\mu^n}\,\mathbb{E}_{z_i'\sim\mu}\,\big\|x_i' s_{\theta,i}'\big\|_{\theta,*}^2, \quad (24)$$

where we recall that $s_{\theta,i}' \in \partial\ell(\cdot, y_i')(\langle x_i', w_\theta(Z)\rangle)$. Combining the two last statements above, we get

$$\mathbb{E}_{Z\sim\mu^n}\,\big[\mathcal{R}_\mu(w_\theta(Z)) - \mathcal{R}_Z(w_\theta(Z))\big] \leq \frac{2}{n}\,\mathbb{E}_{Z\sim\mu^n}\,\mathbb{E}_{z_i'\sim\mu}\,\big\|x_i' s_{\theta,i}'\big\|_{\theta,*}^2. \quad (25)$$

Finally, substituting the close form of $\|\cdot\|_{\theta,*}$ and observing that, by Asm. 1 we have $\big\|x_i' s_{\theta,i}'\big\|_{\theta,*}^2 \leq L^2\big\|x_i'\big\|_{\theta,*}^2$, we get the desired statement:

$$\mathbb{E}_{Z\sim\mu^n}\,\big[\mathcal{R}_\mu(w_\theta(Z)) - \mathcal{R}_Z(w_\theta(Z))\big] \leq \frac{2L^2}{n}\,\mathbb{E}_{z_i'\sim\mu}\,\big\langle x_i', \theta x_i'\big\rangle = \frac{2L^2}{n}\,\mathrm{Tr}\big(\mathbb{E}_{z\sim\mu}\,\theta x x^\top\big). \quad (26)$$

□

B Proof of Prop. 2

In this section we report the proof to get the closed form of the best conditioning function τ_ρ outlined in Prop. 2. In order to do this, we need the following results.

Lemma 7. *For any $\mu \sim \rho_{\mathcal{M}}$, define the inputs' covariance matrix $C_\mu = \mathbb{E}_{x\sim\eta_\mu} xx^\top$. Then, for any $w_\mu \in \operatorname{argmin}_{w\in\mathbb{R}^d} \mathcal{R}_\mu(w)$, the projection $w_{0,\mu} = C_\mu^\dagger C_\mu w_\mu$ of w_μ onto the range of C_μ is still a minimizer of $\mathcal{R}_\mu$.*

Proof. Consider the decomposition of w_μ with respect to the range of C_μ:

$$w_\mu = w_{0,\mu} + w^\perp \quad (27)$$

with $w_{0,\mu} = C_\mu^\dagger C_\mu w_\mu$ and $w^\perp \in \mathbb{R}^d$ such that $C_\mu w^\perp = 0$. We note that, almost surely with respect to the points $x \in \mathbb{R}^d$ sampled from μ, we have $\langle w^\perp, x\rangle = 0$. This follows by noting that by the orthogonality between C_μ and $w^\perp$, we have

$$0 = \big\langle w^\perp, C_\mu w^\perp\big\rangle = \mathbb{E}_{x\sim\eta_\mu}\,\big\langle w^\perp, xx^\top w^\perp\big\rangle = \mathbb{E}_{x\sim\eta_\mu}\,\big\langle x, w^\perp\big\rangle^2, \quad (28)$$

that can hold only if $\big\langle x, w^\perp\big\rangle^2 = 0$ almost surely (a.s.) with respect to η_μ. We conclude that $\langle w_\mu, x\rangle = \langle w_{0,\mu}, x\rangle + \big\langle w^\perp, x\big\rangle = \langle w_{0,\mu}, x\rangle$ a.s. with respect to μ and, consequently, $\mathcal{R}_\mu(w_\mu) = \mathcal{R}_\mu(w_{0,\mu})$. □

Corollary 8. *For any $s \in \mathcal{S}$, recall the conditional covariance matrices in Thm. 1. Then, $\mathrm{Ran}(W(s)) \subset \mathrm{Ran}(C(s))$, namely the range of the task-vector conditional covariance $W(s)$ is always contained in the range of the input conditional covariance $C(s)$.*

Proof. The corollary is a direct consequence of the previous Lemma 7. The result above guarantees that for any $\mu \sim \rho_{\mathcal{M}}$, the rank-one operator $W_\mu = w_\mu w_\mu^\top$ has range contained in the range of C_μ. Taking the conditional expectations $W(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)} W_\mu$ and $C(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)} C_\mu$ maintains this relation unaltered, giving the desired statement. □

Lemma 9. *Let $P \in \mathbb{S}^d_+$ be an orthogonal projector, namely such that $P = P^2$. Then, for any positive definite matrix $\theta \in \mathbb{S}^d_{++}$, we have $P\theta^{-1}P \succeq (P\theta P)^\dagger$.*

Proof. The proof is essentially a corollary of Schur's complement. Let consider the decomposition

$$\theta = \underbrace{P\theta P}_{A} + \underbrace{P\theta(I-P)}_{B} + \underbrace{(I-P)\theta P}_{B^\top} + \underbrace{(I-P)\theta(I-P)}_{C} \tag{29}$$

where $A, C \in \mathbb{S}^d_+$, $B \in \mathbb{R}^{d\times d}$ and $CB = B^\top C = 0$ since $(I-P)P = P(I-P) = P - P^2 = P - P = 0$. Additionally, since $C^\dagger = CC^\dagger C^\dagger = C^\dagger C^\dagger C$, we have that also $AC^\dagger = ACC^\dagger C^\dagger = 0$ and analogously $C^\dagger B = B^\top C^\dagger = 0$. Note that since θ is invertible, both A and C are full rank. We now observe a few relevant interactions between the objects above. In particular, we observe that $CC^\dagger B^\top = B^\top$. To see this, first note that

$$CC^\dagger B^\top = (I-P)\theta(I-P)\big((I-P)\theta(I-P)\big)^\dagger(I-P)\theta P. \tag{30}$$

By taking $D = (I-P)\theta^{1/2}$ and using the properties of the pseudoinverse (e.g. $D = D^\top(DD^\top)^\dagger$), we have

$$\begin{aligned} CC^\dagger B^\top &= DD^\top(DD^\top)^\dagger D\theta^{1/2}P \\ &= DD^\dagger D\theta^{1/2}P \\ &= D\theta^{1/2}P \\ &= B^\top. \end{aligned} \tag{31}$$

We now derive an alternative characterization of θ in terms of A, B, C. By adding and removing a term $BC^\dagger B$ to θ, we have

$$\begin{aligned} \theta &= A + B + B^\top + C \\ &= A - BC^\dagger B^\top + BC^\dagger B^\top + B + B^\top + C \\ &= A - BC^\dagger B^\top + B + C + (B+C)(C^\dagger B^\top) \\ &= A - BC^\dagger B^\top + B + C + (A - BC^\dagger B^\top + B + C)(C^\dagger B) \\ &= (A - BC^\dagger B^\top + B + C)(I + C^\dagger B^\top), \end{aligned} \tag{32}$$

where we have first used the equality $CC^\dagger B^\top = B^\top$ and then the ortogonality $AC^\dagger = B^\top C^\dagger = 0$. Following a similar reasoning

$$\begin{aligned} A - BC^\dagger B^\top + B + C &= A - BC^\dagger B^\top + C + BC^\dagger C \\ &= A - BC^\dagger B^\top + C + BC^\dagger(A - BC^\dagger B^\top + C) \\ &= (I + BC^\dagger)(A - BC^\dagger B^\top + C) \end{aligned} \tag{33}$$

since $BC^\dagger C = C$ (following the same reasoning used for $B^\top = CC^\dagger B^\top$) and $AC^\dagger = C^\dagger B = 0$. We conclude that

$$\theta = (I + BC^\dagger)(A - BC^\dagger B^\top + C)(I + C^\dagger B^\top). \tag{34}$$

We now show that all terms in the equation above are invertible. First note that $(I + BC^\dagger)^{-1} = (I - BC^\dagger)$ and $(I + C^\dagger B^\top)^{-1} = (I + C^\dagger B^\top)$. Moreover, since $\theta \succ 0$ and $C(A - BC^\dagger B^\top) = 0$, then also $A - BC^\dagger B^\top \succ 0$. We have

$$\theta^{-1} = (I - C^\dagger B^\top)(A - BC^\dagger B^\top + C)^{-1}(I - BC^\dagger), \tag{35}$$

from which we conclude

$$\begin{aligned} P\theta^{-1}P &= P(A - BC^\dagger B^\top + C)^{-1}P \\ &= P\Big((A - BC^\dagger B^\top)^\dagger + C^\dagger\Big)P \\ &= P(A - BC^\dagger B^\top)^\dagger P \\ &= (A - BC^\dagger B^\top)^\dagger. \end{aligned} \tag{36}$$

Since $BC^\dagger B^\top \succeq 0$, we have $A - BC^\dagger B^\top \preceq A$ and therefore $(A - BC^\dagger B^\top)^\dagger \succeq A^\dagger$ from which we have

$$P\theta^{-1}P = (A - BC^\dagger B^\top)^\dagger \succeq A^\dagger = (P\theta P)^\dagger, \tag{37}$$

as desired. □

Proposition 10. *Consider two matrices $A, B \in \mathbb{S}^d_+$ such that* $\mathrm{Ran}(A) \subseteq \mathrm{Ran}(B)$ *and consider the following associated problem:*

$$\min_{\theta \in \mathbb{S}^d_+,\ \mathrm{Ran}(A) \subseteq \mathrm{Ran}(\theta)} \mathrm{Tr}(\theta^{-1} A) + \mathrm{Tr}(\theta B). \tag{38}$$

Then, a minimizer and the corresponding minimum of the problem above are given by

$$\theta_* = B^{-1/2}(B^{1/2} A B^{1/2})^{1/2} B^{-1/2} \qquad 2\left\| B^{1/2} A^{1/2} \right\|_*. \tag{39}$$

Moreover θ_* is the unique minimizer such that $\mathrm{Ran}(\theta_*) \subset \mathrm{Ran}(B)$.

Proof. Let $\Theta = \{\theta \in \mathbb{S}^d_+ \mid \mathrm{Ran}(A) \subset \mathrm{Ran}(\theta)\}$ and denote by $F : \Theta \to \mathbb{R}$ the objective functional of the problem in Eq. (38), such that for any $\theta \in \Theta$

$$F(\theta) = \mathrm{Tr}(\theta^{-1} A) + \mathrm{Tr}(\theta B). \tag{40}$$

Note that the sign of inverse is well defined since $\mathrm{Ran}(A) \subset \mathrm{Ran}(\theta)$. We begin the proof by showing that the Eq. (38) is equivalent to

$$\min_{\theta \in \mathbb{S}^d_+,\ \mathrm{Ran}(A) \subset \mathrm{Ran}(\theta) \subset \mathrm{Ran}(B)} \mathrm{Tr}(\theta^{-1} A) + \mathrm{Tr}(\theta B). \tag{41}$$

To see this, let $P = BB^\dagger$ the orthogonal projector onto the range of B. By hypothesis, $A = PAP$ and $B = PBP$. Therefore, for any $\theta \in \mathbb{S}^d_{++}$

$$\begin{aligned} F(\theta) &= \mathrm{Tr}(\theta^{-1} A) + \mathrm{Tr}(\theta B) \\ &= \mathrm{Tr}(P\theta^{-1} P A) + \mathrm{Tr}(P\theta P B) \\ &\geq \mathrm{Tr}((P\theta P)^\dagger A) + \mathrm{Tr}(P\theta P B) \\ &= F(P\theta P), \end{aligned}$$

where we have applied the fact that $P\theta^{-1}P \succeq (P\theta P)^\dagger$ from Lemma 9 and the positive semidefiniteness of A. The inequality above implies the equivalence between Eq. (38) and Eq. (41). Indeed, let $\theta_* \in \Theta$ be a minimizer of Eq. (38) and consider a sequence $(\theta_n)_{n \in \mathbb{N}}$ such that $\theta_n \in \mathbb{S}^d_{++}$ for any $n \in \mathbb{N}$ and $\theta_n \to \theta_*$. By continuity of F we have also that $F(\theta_n) \to F(\theta_*)$. Clearly, $F(\theta_*) \leq F(P\theta_n P) \leq F(\theta_n)$ and therefore also $F(P\theta_n P) \to F(\theta_*)$. By continuity of F over Θ, this also implies that the limit $\lim_{n \to +\infty} P\theta_n P = P\theta_* P$ is a minimizer for Eq. (38) (and one such that $\mathrm{Ran}(\theta_*) \subset \mathrm{Ran}(B)$). We consider now the set $\Theta_B = \{\theta \in \mathbb{S}^d_+ \mid \mathrm{Ran}(\theta) = \mathrm{Ran}(B)\}$ of all positive semidefinite matrices with same range as B, hence invertible on $\mathrm{Ran}(B)$. Note that Θ_B is an open subset of Θ and its closure in Θ corresponds to Θ itself. By definition, any $\theta \in \Theta_B$ is such that $\theta = B^{\dagger/2} X B^{\dagger/2}$ with $\mathrm{Ran}(X) = \mathrm{Ran}(B)$. This implies in particular that $XB^\dagger B = X$ and $\theta^\dagger = B^{\dagger/2} X^\dagger B^{\dagger/2}$. Therefore,

$$\begin{aligned} F(\theta) &= \mathrm{Tr}(\theta^\dagger A) + \mathrm{Tr}(\theta B) && (42) \\ &= \mathrm{Tr}(X^\dagger B^{1/2} A B^{1/2}) + \mathrm{Tr}(X), && (43) \end{aligned}$$

and $\mathrm{Ran}(B^{1/2} A B^{1/2}) \subseteq \mathrm{Ran}(B) = \mathrm{Ran}(X)$. We can now minimize the problem with respect to X, namely

$$\min_{X \in \mathbb{S}^d_+,\ \mathrm{Ran}(B^{1/2} A B^{1/2}) \subseteq \mathrm{Ran}(X)} \mathrm{Tr}(X^\dagger B^{1/2} A B^{1/2}) + \mathrm{Tr}(X). \tag{44}$$

The minimization corresponds to the variational form of the trace norm of $B^{1/2} A B^{1/2}$ [29] and has solution $X_* = (B^{1/2} A B^{1/2})^{1/2}$, with minimum corresponding to $2\mathrm{Tr}((B^{1/2} A B^{1/2})^{1/2}) = 2\left\| B^{1/2} A^{1/2} \right\|_*$. To conclude the proof, let $G : \{X \in \mathbb{S}^d_+ \mid \mathrm{Ran}(B^{1/2} A B^{1/2}) \subseteq \mathrm{Ran}(X)\} \to \mathbb{R}$ be the objective functional in Eq. (44) such that $G(X) = \mathrm{Tr}(X^\dagger B^{1/2} A B^{1/2}) + \mathrm{Tr}(X)$. Let now $X_* \in \mathbb{S}^d_+$ be a minimzer for G and $(X_n)_{n \in \mathbb{N}}$ be a minimizing sequence with $\mathrm{Ran}(X_n) = \mathrm{Ran}(B)$ for each $n \in \mathbb{N}$ and $X_n \to X_*$. Let $(\theta_n)_{n \in \mathbb{N}}$ such that $\theta_n = B^{\dagger/2} X B^{\dagger/2}$ for any $n \in \mathbb{N}$. Then we have $\theta_n \to B^{\dagger/2} X_* B^{\dagger/2}$ and by continuity $F(B^{\dagger/2} X_* B^{\dagger/2}) = G(X_*)$, hence $\min_X G(X) \leq \min_\theta F(\theta)$. Note that $B^{\dagger/2} X_* B^{\dagger/2}$ is a minimizer for F, since F and G have same minimum value. To see this it is sufficient to show that, given a minimizing sequence $(\theta_n)_{n \in \mathbb{N}}$ such that $\mathrm{Ran}(\theta_n) = \mathrm{Ran}(B)$ for any $n \in \mathbb{N}$ and $\theta_n \to \theta_*$, we have $X_n = B^{1/2} \theta_n B^{1/2} \to B^{1/2} \theta_n B^{1/2}$

and thus $F(\theta_*) = G(B^{1/2}\theta_n B^{1/2})$. We have shown that $\min_\theta F(\theta) \geq \min_X G(X)$. Therefore $\theta_* = B^{\dagger/2} X_* B^{\dagger/2} = B^{\dagger/2}(B^{1/2}AB^{1/2})^{1/2}B^{\dagger/2}$ is a minimizer of Eq. (38) as desired. The uniqueness of θ_* follows from the uniqueness of X_* from the standard results on the variational form of the trace norm [29]. □

We now have all the ingredients necessary to prove Prop. 2.

Proposition 2 (Best conditioning function in hindsight)**.** *The conditioning function minimizer and the minimum of the bound presented in Thm. 1 over the set* $\{\tau \in \mathcal{T} \mid \text{Ran}(W(s)) \subseteq \text{Ran}(\tau(s)),\ \rho_{\mathcal{S}}\text{-almost surely}\}$*, are respectively*

$$\tau_\rho(s) = (2L)^{-1}\, n^{1/2}\, C(s)^{\dagger/2}(C(s)^{1/2}W(s)C(s)^{1/2})^{1/2}C(s)^{\dagger/2}$$

$$\mathcal{E}_\rho(\tau_\rho) - \mathcal{E}_\rho^* \leq 2L\mathbb{E}_{s\sim\rho_{\mathcal{S}}}\left\|W(s)^{1/2}C(s)^{1/2}\right\|_* n^{-1/2}. \tag{8}$$

Proof. We aim to minimize

$$\min_{\substack{\tau:\mathcal{S}\to\Theta \\ \text{Ran}(W(s))\subseteq\text{Ran}(\tau(s))}} \mathbb{E}_{s\sim\rho_{\mathcal{S}}}\, \varphi(s,\tau(s)) \qquad \text{with} \qquad \varphi(s,\theta) = \frac{\text{Tr}\big(\theta^\dagger W(s)\big)}{2} + \frac{2L^2\text{Tr}\big(\theta C(s)\big)}{n}. \tag{45}$$

over the set of all measurable functions $\tau : \mathcal{S} \to \Theta$. Note that from Cor. 8, for any $s \in \mathcal{S}$ we have $\text{Ran}(W(s)) \subset \text{Ran}(C(s))$. Therefore we can apply Prop. 10 to have that for any $s \in \mathcal{S}$, the problem

$$\min_{\theta\in\mathbb{S}_+^d,\ \text{Ran}(W(s))\subseteq\text{Ran}(\theta)} \varphi(s,\theta) \tag{46}$$

has solution

$$\tau_\rho(s) = \frac{\sqrt{n}}{2L}\, C(s)^{\dagger/2}(C(s)^{1/2}W(s)C^{1/2})^{1/2}C(s)^{\dagger/2}. \tag{47}$$

Therefore, for any $\tau : \mathcal{S} \to \Theta$ we have

$$\mathbb{E}_{s\sim\rho_{\mathcal{S}}}\, \varphi(\tau_\rho(s), s) \leq \mathbb{E}_{s\sim\rho_{\mathcal{S}}}\, \varphi(\tau(s), s), \tag{48}$$

and therefore $\mathbb{E}_{s\sim\rho_{\mathcal{S}}}\, \varphi(\tau_\rho(s), s) \leq \min_\tau \mathbb{E}_{s\sim\rho_{\mathcal{S}}}\, \varphi(\tau)$. To conclude the proof we need to show that τ_ρ is measurable. This follows immediately by applying Aumann's measurable selection principle, see for instance the formulation in [34, Lemma A.3.18]. Under the notation of [34], we can apply the result by taking $h(s,\theta) = (\theta\theta^\dagger - I)W(s)$, the set $A = \{0\} \subset Y = \mathbb{S}_+^d$. This guarantees the existence of a measurable function $\tau_0 : \mathcal{S} \to \Theta$ such that it minimizes pointwise $\varphi(s,\cdot)$ for any $s \in \mathcal{S}$ on the set $\{\theta \in \mathbb{S}_+^d \mid \text{Ran}(W(s)) \subset \text{Ran}(\theta)\}$. The uniqueness of $\tau_\rho(s)$ for each $s \in \mathcal{S}$ guarantees that $\tau_\rho = \tau_0$ is measurable as desired. □

C Proof of Ex. 1

In this section we report the proof of the statement in Ex. 1.

Example 1 (Clusters)**.** *Let $\mathcal{S} = \mathbb{R}^q$ be the side information space, for some integer $q > 0$. Let ρ be such that the side information marginal distribution $\rho_{\mathcal{S}}$ is given by a uniform mixture of m uniform distributions. More precisely, let $\rho_{\mathcal{S}} = \frac{1}{m}\sum_{i=1}^m \rho_{\mathcal{S}}^{(i)}$, with $\rho_{\mathcal{S}}^{(i)} = \mathcal{U}\big(\mathcal{B}(a_i, 1/2)\big)$ the uniform distribution on the ball of radius $1/2$ centered at $a_i \in \mathcal{S}$, characterizing the cluster i. For a given side information s, a task $\mu \sim \rho(\cdot|s)$ is sampled such that: 1) its inputs' marginal η_μ is a distribution with constant covariance matrix $C(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)}\mathbb{E}_{x\sim\eta_\mu} xx^\top = C$, for some $C \in \mathbb{S}_+^d$, 2) w_μ is sampled from a distribution with conditional covariance matrix $W(s) = \mathbb{E}_{\mu\sim\rho(\cdot|s)} w_\mu w_\mu^\top$, with $W(s)$ such that $(C^{1/2}W(s)C^{1/2})(C^{1/2}W(p)C^{1/2}) = 0$ if $s \neq p$. Then, $\mathbb{E}_{s\sim\rho_{\mathcal{S}}}\left\|C(s)^{1/2}W(s)^{1/2}\right\|_* = \frac{1}{\sqrt{m}}\left\|W_\rho^{1/2}C_\rho^{1/2}\right\|_*$.*

Proof. According to the setting described in the example, we can rewrite the following:

$$\begin{aligned}
\mathbb{E}_{s\sim\rho_{\mathcal{S}}}\left\|C(s)^{1/2}W(s)^{1/2}\right\|_* &= \mathbb{E}_{s\sim\rho_{\mathcal{S}}}\left\|C^{1/2}W(s)^{1/2}\right\|_* \\
&= \mathbb{E}_{s\sim\rho_{\mathcal{S}}}\mathrm{Tr}\Big(\big(C^{1/2}W(s)C^{1/2}\big)^{1/2}\Big) \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{s\sim\rho_{\mathcal{S}}^{(i)}}\mathrm{Tr}\Big(\big(C^{1/2}W(s)C^{1/2}\big)^{1/2}\Big) \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathrm{Tr}\Big(\big(C^{1/2}W(a_i)C^{1/2}\big)^{1/2}\Big) \\
&= \frac{1}{m}\mathrm{Tr}\left(\sum_{i=1}^{m}\big(C^{1/2}W(a_i)C^{1/2}\big)^{1/2}\right) \\
&= \frac{1}{m}\mathrm{Tr}\left(\Big(\sum_{i=1}^{m}C^{1/2}W(a_i)C^{1/2}\Big)^{1/2}\right),
\end{aligned} \tag{49}$$

where, in the first equality we have exploited the fact that $C(s)$ is a constant matrix C, in the second equality we have applied the definition of the rewriting of the trace norm of a matrix A as $\|A\|_* = \mathrm{Tr}\big((AA^\top)^{1/2}\big)$, in the third and fourth equality we have exploited the assumption on $\rho_{\mathcal{S}}$, and finally, in the last equality, by point 2), we managed to apply the fact that, for two matrices $A, B \in \mathbb{S}_+^d$ such that $A^{1/2}B^{1/2} = B^{1/2}A^{1/2} = 0$, we have

$$(A^{1/2}+B^{1/2})(A^{1/2}+B^{1/2}) = A+B \implies (A+B)^{1/2} = A^{1/2}+B^{1/2}. \tag{50}$$

On the other hand, we observe that we can also write the following:

$$\begin{aligned}
\left\|C_\rho^{1/2}W_\rho^{1/2}\right\|_* &= \left\|C^{1/2}W_\rho^{1/2}\right\|_* \\
&= \mathrm{Tr}\Big(\big(C^{1/2}W_\rho C^{1/2}\big)^{1/2}\Big) \\
&= \mathrm{Tr}\Big(\big(C^{1/2}\mathbb{E}_{s\sim\rho_{\mathcal{S}}}W(s)C^{1/2}\big)^{1/2}\Big) \\
&= \mathrm{Tr}\left(\Big(C^{1/2}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{s\sim\rho_{\mathcal{S}}^{(i)}}W(s)C^{1/2}\Big)^{1/2}\right) \\
&= \frac{1}{\sqrt{m}}\mathrm{Tr}\left(\Big(\sum_{i=1}^{m}C^{1/2}W(a_i)C^{1/2}\Big)^{1/2}\right) \\
&= \frac{1}{\sqrt{m}}\mathrm{Tr}\left(\Big(C^{1/2}\sum_{i=1}^{m}W(a_i)C^{1/2}\Big)^{1/2}\right),
\end{aligned} \tag{51}$$

where, in the first equality we have exploited the fact that $C(s)$ is a constant matrix C, in the second equality we have applied the definition of the rewriting of the trace norm of a matrix A as $\|A\|_* = \mathrm{Tr}\big((AA^\top)^{1/2}\big)$ and in the fourth and fifth equality we have exploited the assumption on $\rho_{\mathcal{S}}$. The desired statement directly derives from combining Eq. (49) and Eq. (51). □

D Proofs of the statements in Sec. 4

In this section we report the proofs of the statements we used in Sec. 4 in order to prove the expected excess risk bound for Alg. 1 in Thm. 5. We start from proving the matricial rewriting of Prop. 3 in App. D.1. We then prove in App. D.2 the properties of the surrogate functions in Prop. 4. Then, in App. D.3, we prove the convergence rate of Alg. 1 on the surrogate problem in Eq. (15).

D.1 Proof of Prop. 3

We start from proving the matricial rewriting of Prop. 3..

Proposition 3 (Matricial re-formulation of $\tau_M(s)$)**.** *Let τ be as in Eq. (11). Then,*

$$\tau(s) = \big(I_d \otimes \Phi(s)^\top\big) H_M \big(I_d \otimes \Phi(s)\big) + C, \tag{12}$$

where I_d is the identity in $\mathbb{R}^{d\times d}$ and H_M is the matrix in $\mathbb{R}^{dk\times dk}$ defined by the entries

$$\big(H_M\big)_{(i-1)k+h,(j-1)k+z} = \big\langle M(:,i,h), M(:,j,z)\big\rangle, \quad i,j = 1,\dots,d, \quad h,z = 1,\dots,k.$$

Proof. We start from observing that for any $i,j = 1,\dots,d$, we can rewrite the following

$$\begin{aligned}
\Big(\big(M\Phi(s)\big)^\top M\Phi(s)\Big)_{i,j} &= \big\langle \big(M\Phi(s)\big)^\top(i,:), \big(M\Phi(s)\big)(:,j)\big\rangle \\
&= \big\langle \big(M\Phi(s)\big)(:,i), \big(M\Phi(s)\big)(:,j)\big\rangle \\
&= \sum_{q=1}^{m} \big(M\Phi(s)\big)(:,i)_q \big(M\Phi(s)\big)(:,j)_q \\
&= \sum_{q=1}^{m} \left(\sum_{h=1}^{k} M_{q,i,h}\Phi(s)_h\right)\left(\sum_{z=1}^{k} M_{q,j,z}\Phi(s)_z\right) \\
&= \sum_{q=1}^{m}\sum_{h=1}^{k}\sum_{z=1}^{k} M_{q,i,h}M_{q,j,z}\Phi(s)_h\Phi(s)_z \\
&= \sum_{h=1}^{k}\sum_{z=1}^{k} \Phi(s)_h\Phi(s)_z \sum_{q=1}^{m} M_{q,i,h}M_{q,j,z} \\
&= \sum_{h=1}^{k}\sum_{z=1}^{k} \Phi(s)_h\Phi(s)_z \left(\sum_{q=1}^{m} M_{q,i,h}M_{q,j,z}\right) \\
&= \sum_{h=1}^{k}\sum_{z=1}^{k} \Phi(s)_h\Phi(s)_z \big\langle M(:,i,h), M(:,j,z)\big\rangle.
\end{aligned} \tag{52}$$

We now observe that for any $i,j = 1,\dots,d$, we can rewrite the following

$$\begin{aligned}
\Big(\big(I_d \otimes \Phi(s)^\top\big) H_M \big(I_d \otimes \Phi(s)\big)\Big)_{i,j} &= \big\langle \big(I_d \otimes \Phi(s)^\top\big)(i,:), \big(H_M\big(I_d \otimes \Phi(s)\big)\big)(:,j)\big\rangle \\
&= \big\langle \big(I_d \otimes \Phi(s)\big)(:,i), \big(H_M\big(I_d \otimes \Phi(s)\big)\big)(:,j)\big\rangle \\
&= \sum_{n=1}^{kd} \big(I_d \otimes \Phi(s)\big)_{n,i} \big(H_M\big(I_d \otimes \Phi(s)\big)\big)_{n,j} \\
&= \sum_{n=1}^{kd} \big(I_d \otimes \Phi(s)\big)_{n,i} \big\langle H_M(n,:), \big(I_d \otimes \Phi(s)\big)(:,j)\big\rangle \\
&= \sum_{n=1}^{kd} \big(I_d \otimes \Phi(s)\big)_{n,i} \sum_{p=1}^{kd} \big(H_M\big)_{n,p} \big(I_d \otimes \Phi(s)\big)_{p,j} \\
&= \sum_{n=1}^{kd}\sum_{p=1}^{kd} \big(I_d \otimes \Phi(s)\big)_{n,i} \big(H_M\big)_{n,p} \big(I_d \otimes \Phi(s)\big)_{p,j} \\
&= \sum_{n=1}^{kd}\sum_{p=1}^{kd} \Phi(s)_h\, \delta_{n,(i-1)k+h} \big(H_M\big)_{n,p} \Phi(s)_z\, \delta_{p,(j-1)k+z} \\
&= \sum_{h=1}^{k}\sum_{z=1}^{k} \Phi(s)_h\Phi(s)_z \big(H_M\big)_{(i-1)k+h,(j-1)k+z},
\end{aligned} \tag{53}$$

where, in the seventh equality we have exploited the fact that, by definition,

$$\big(I_d \otimes \Phi(s)\big)_{n,i} = \begin{cases} \Phi(s)_r & \text{if } r = n-(i-1)k \\ 0 & \text{otherwise} \end{cases} = \Phi(s)_r \; \delta_{n,r+(i-1)k}. \tag{54}$$

and in the last equality we have defined the new indexes $h, z = 1, \dots, k$ as

$$h = n-(i-1)k \qquad z = p-(j-1)k \tag{55}$$

and, as consequence, we have rewritten

$$n = (i-1)k + h \qquad p = (j-1)k + z. \tag{56}$$

As, a consequence, if we define H_M as the matrix in $\mathbb{R}^{dk \times dk}$ with entries

$$\big(H_M\big)_{(i-1)k+h,(j-1)k+z} = \big\langle M(:,i,h), M(:,j,z)\big\rangle, \tag{57}$$

with $i, j = 1, \dots, d$ and $h, z = 1, \dots, k$, then, Eq. (52):

$$\Big(\big(M\Phi(s)\big)^\top M\Phi(s)\Big)_{i,j} = \sum_{h=1}^{k}\sum_{z=1}^{k} \Phi(s)_h \Phi(s)_z \big\langle M(:,i,h), M(:,j,z)\big\rangle \tag{58}$$

and Eq. (53):

$$\Big(\big(I_d \otimes \Phi(s)^\top\big) H_M \big(I_d \otimes \Phi(s)\big)\Big)_{i,j} = \sum_{h=1}^{k}\sum_{z=1}^{k} \Phi(s)_h \Phi(s)_z \big(H_M\big)_{(i-1)k+h,(j-1)k+z} \tag{59}$$

coincide. This coincides with the first desired statement. In order to prove the statement $H_M \in \mathbb{S}_+^{dk}$, we show that $H_M = A_M^\top A_M$, where A_M is the matrix in $\mathbb{R}^{m \times dk}$ defined as

$$A_M(:,(i-1)k+h) = M(:,i,h). \tag{60}$$

We start from recalling that, by definition of H_M, we have

$$\big(H_M\big)_{(i-1)k+h,(j-1)k+z} = \big\langle M(:,i,h), M(:,j,z)\big\rangle. \tag{61}$$

Moreover, we observe that, for any $p, q = 1, \dots, kd$,

$$\big(A_M^\top A_M\big)_{p,q} = \big\langle \big(A_M^\top\big)(p,:), A_M(:,q)\big\rangle_{\mathbb{R}^m} = \big\langle A_M(:,p), A_M(:,q)\big\rangle. \tag{62}$$

As a consequence, the desired statement is satisfied if we define

$$\big(A_M\big)(:,(i-1)k+h) = M(:,i,h). \tag{63}$$

We now prove the last statement. Let $(e_i)_{i=1}^d$ be the canonical basis in $\mathbb{R}^d$. By the definition of the trace and the rewriting of $\tau(s)$ in Prop. 3, denoting by vec the vectorization operation, we can rewrite

$$\begin{aligned}
\mathrm{Tr}\big(\tau(s)\big) &= \sum_{i=1}^{d} \big\langle e_i, \tau(s) e_i\big\rangle \\
&= \sum_{i=1}^{d} \big\langle e_i, \big(I_d \otimes \Phi(s)^\top\big) H_M \big(I_d \otimes \Phi(s)\big) e_i\big\rangle \\
&= \sum_{i=1}^{d} e_i^\top \big(I_d \otimes \Phi(s)^\top\big) H_M \big(I_d \otimes \Phi(s)\big) e_i \\
&= \sum_{i=1}^{d} \Big(\big(I_d \otimes \Phi(s)\big) e_i\Big)^\top H_M \big(I_d \otimes \Phi(s)\big) e_i \\
&= \sum_{i=1}^{d} \Big(\mathrm{vec}\big(\Phi(s) e_i^\top\big)\Big)^\top H_M \mathrm{vec}\big(\Phi(s) e_i^\top\big) \\
&= \mathrm{Tr}\Big(H_M \sum_{i=1}^{d} \mathrm{vec}\big(\Phi(s) e_i^\top\big) \mathrm{vec}\big(\Phi(s) e_i^\top\big)^\top\Big) \\
&\leq \mathrm{Tr}\big(H_M\big) \left\| \sum_{i=1}^{d} \mathrm{vec}\big(\Phi(s) e_i^\top\big) \mathrm{vec}\big(\Phi(s) e_i^\top\big)^\top \right\|_\infty \\
&= \mathrm{Tr}\big(H_M\big) \big\|\Phi(s)\big\|_{\mathbb{R}^k}^2,
\end{aligned} \tag{64}$$

where, in the fifth equality, we have applied the relation

$$\big(C^\top \otimes A\big)\mathrm{vec}(B) = \mathrm{vec}(ABC) \tag{65}$$

with $A = \Phi(s)$, $B = e_i^\top$ and $C = I_d$, i.e.

$$\big(I_d \otimes \Phi(s)\big)e_i = \mathrm{vec}\big(\Phi(s)e_i^\top\big), \tag{66}$$

in the inequality we have applied Holder's inequality and in the last equality we have applied the following proposition. □

Proposition 11. *For any $i = 1, \dots, d$, define*

$$v_i = \mathrm{vec}\big(\Phi(s)e_i^\top\big) \tag{67}$$

Then,

$$\left\| \sum_{i=1}^d \mathrm{vec}\big(\Phi(s)e_i^\top\big)\mathrm{vec}\big(\Phi(s)e_i^\top\big)^\top \right\|_\infty = \left\| \sum_{i=1}^d v_i v_i^\top \right\|_\infty = \big\|\Phi(s)\big\|^2. \tag{68}$$

Proof. We start from observing that, for any $i, j = 1, \dots, d$, we have

$$\begin{aligned} v_i^\top v_j &= \mathrm{vec}\big(\Phi(s)e_i^\top\big)^\top \mathrm{vec}\big(\Phi(s)e_j^\top\big) \\ &= \mathrm{Tr}\big(e_i \Phi(s)^\top \Phi(s) e_j^\top\big) \\ &= \mathrm{Tr}\big(\Phi(s)^\top \Phi(s) e_j^\top e_i\big) \\ &= \Phi(s)^\top \Phi(s) e_j^\top e_i \\ &= \big\|\Phi(s)\big\|^2 \delta_{i,j}, \end{aligned} \tag{69}$$

where, in the second equality, we have used the property of the operator vec:

$$\mathrm{vec}(A)^\top \mathrm{vec}(B) = \mathrm{Tr}\big(A^\top B\big) \tag{70}$$

with

$$A = \Phi(s)e_i^\top \qquad B = \Phi(s)e_j^\top. \tag{71}$$

As a consequence, the vectors

$$\tilde{v_i} = \frac{v_i}{\|v_i\|} = \frac{v_i}{\big\|\Phi(s)\big\|} \qquad i = 1, \dots, d, \tag{72}$$

form an orthonormal basis of the space. Moreover, we can rewrite the operator above as follows

$$\sum_{i=1}^d \mathrm{vec}\big(\Phi(s)e_i^\top\big)\mathrm{vec}\big(\Phi(s)e_i^\top\big)^\top = \sum_{i=1}^d v_i v_i^\top = \sum_{i=1}^d \big\|\Phi(s)\big\|^2 \tilde{v_i}\tilde{v_i}^\top. \tag{73}$$

The rewriting above coincides with the eigenvalues' decomposition of the operator: the vectors $\tilde{v_i}$ are the eigenvectors with associated constant eigenvalues $\big\|\Phi(s)\big\|^2$. As a consequence, we can conclude that

$$\left\| \sum_{i=1}^d \mathrm{vec}\big(\Phi(s)e_i^\top\big)\mathrm{vec}\big(\Phi(s)e_i^\top\big)^\top \right\|_\infty = \big\|\Phi(s)\big\|^2. \tag{74}$$

□

D.2 Proof of Prop. 4

We now prove the properties of the surrogate functions in Prop. 4.

Proposition 4 (Properties of the surrogate meta-loss $\mathcal{L}$)**.** *For any $Z \in \mathcal{D}$ and $s \in \mathcal{S}$, the function $\mathcal{L}\big(\cdot, \cdot, s, Z\big)$ is convex and one of its subgradients is given, for any $H \in \mathbb{S}_+^{dk}$ and $C \in \mathbb{S}_+^d$, by*

$$\nabla\mathcal{L}\big(H, \cdot, s, Z\big)(C) = \hat{\nabla}, \qquad \nabla\mathcal{L}\big(\cdot, C, s, Z\big)(H) = \big(I_d \otimes \Phi(s)\big)\hat{\nabla}\big(I_d \otimes \Phi(s)^\top\big),$$

where

$$\hat{\nabla} = -\frac{\lambda}{2}\tau_{H,C}(s)^\dagger w_{\tau_{H,C}(s)} w_{\tau_{H,C}(s)}^\top \tau_{H,C}(s)^\dagger + \frac{2L^2 X^\top X}{n^2}.$$

Moreover, by Asm. 1 and Asm. 3,

$$\big\|\nabla\mathcal{L}\big(\cdot, \cdot, s, Z\big)(H, C)\big\|_F \leq (1 + K^2)(LR)^2\big(2^{-1} + 2n^{-1}\big).$$

Proof. We are interested in studying the properties of the surrogate function $\mathcal{L}\big(\cdot,\cdot,s,Z\big) : \mathbb{S}_+^{dk} \times \mathbb{S}_+^d \to \mathbb{R}$ in Eq. (15). We start from observing that, such a function coincides with the composition of the function

$$\theta \in \mathbb{S}_+^d \mapsto \Delta(\theta, Z) = F(\theta, Z) + G(\theta, Z) \in \mathbb{R}$$

$$F(\theta, Z) = \min_{w \in \mathbb{R}^d} \mathcal{R}_{Z,\theta}(w) \qquad \mathcal{R}_{Z,\theta}(w) = \frac{1}{n}\sum_{i=1}^n \ell(\langle x_i, w\rangle, y_i) + \frac{\lambda}{2}\big\langle w, \theta^\dagger w\big\rangle + \iota_{\mathrm{Ran}(\theta)}(w) \tag{75}$$

$$G(\theta, Z) = \frac{2L^2}{n}\mathrm{Tr}\Big(\theta\frac{X^\top X}{n}\Big).$$

with the linear transformation

$$s \in \mathcal{S} \mapsto \tau_{H,C}(s) = \big(I_d \otimes \Phi(s)^\top\big) H \big(I_d \otimes \Phi(s)\big) + C \in \mathbb{S}_+^d. \tag{76}$$

In other words, for any $H \in \mathbb{S}_+^{dk}$ and $C \in \mathbb{S}_+^d$, we can write

$$\mathcal{L}\big(H, C, s, Z\big) = \Delta(\tau_{H,C}(s), Z) = F(\tau_{H,C}(s), Z) + G(\tau_{H,C}(s), Z). \tag{77}$$

We now observe that both the functions $F(\cdot, Z)$ and $G(\cdot, Z)$ are both convex ($F(\cdot, Z)$ is convex since it is the minimum of a jointly convex function see [15] and $G(\cdot, Z)$ is a linear function). As a consequence, the function $\Delta(\cdot, Z)$ is convex over $\mathbb{S}_+^d$. This implies the convexity of the surrogate function $\mathcal{L}\big(\cdot,\cdot,s,Z\big)$ over $\mathbb{S}_+^{dk} \times \mathbb{S}_+^d$ (composition of a convex function with a linear transformation). In order to get the closed form of the gradient in Prop. 4 we proceed in a similar way as in [14]. More precisely, we start from recalling that, as already observed in [15], thanks to strong duality in the within-task problem, for any $\theta \in \mathbb{S}_+^d$, we can rewrite

$$F(\theta, Z) = \min_{w \in \mathrm{Ran}(\theta)} \mathcal{R}_{Z,\theta}(w) = \max_{\alpha \in \mathbb{R}^n} \Big\{ -\frac{1}{n}\sum_{i=1}^n \ell_i^*(\alpha_i) - \frac{1}{2n^2}\mathrm{Tr}\big(\theta X^\top \alpha\alpha^\top X\big)\Big\}, \tag{78}$$

where, $\ell_i^*(\cdot)$ denotes the Fenchel conjugate of $\ell_i(\cdot) = \ell(\cdot, y_i)$ and $\alpha \in \mathbb{R}^n$ coicides with the dual variable. As a consequence, we can rewrite

$$\begin{aligned}
\Delta(\theta, Z) &= F(\theta, Z) + G(\theta, Z) \\
&= \max_{\alpha \in \mathbb{R}^n} \Big\{ -\frac{1}{n}\sum_{i=1}^n \ell_i^*(\alpha_i) - \frac{1}{2n^2}\mathrm{Tr}\big(\theta X^\top \alpha\alpha^\top X\big)\Big\} + \frac{2L^2}{n}\mathrm{Tr}\Big(\theta\frac{X^\top X}{n}\Big) \\
&= \max_{\alpha \in \mathbb{R}^n} \Big\{ -\frac{1}{n}\sum_{i=1}^n \ell_i^*(\alpha_i) + \mathrm{Tr}\Big(\theta\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\Big)\Big\}.
\end{aligned} \tag{79}$$

As a consequence, we have

$$\begin{aligned}
\Delta(\tau_{H,C}, Z) &= \max_{\alpha \in \mathbb{R}^n} \Big\{ -\frac{1}{n}\sum_{i=1}^n \ell_i^*(\alpha_i) + \mathrm{Tr}\Big(\big(I_d \otimes \Phi(s)^\top\big) H \big(I_d \otimes \Phi(s)\big)\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\Big) \\
&\qquad + \mathrm{Tr}\Big(C\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\Big)\Big\} \\
&= \max_{\alpha \in \mathbb{R}^n} \Big\{ -\frac{1}{n}\sum_{i=1}^n \ell_i^*(\alpha_i) + \mathrm{Tr}\Big(H\big(I_d \otimes \Phi(s)\big)\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\big(I_d \otimes \Phi(s)^\top\big)\Big) \\
&\qquad + \mathrm{Tr}\Big(C\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\Big)\Big\} \\
&= \max_{\alpha \in \mathbb{R}^n} Q(\alpha, H, C, s, Z),
\end{aligned} \tag{80}$$

where we have introduced the function

$$\begin{aligned}
Q(\alpha, H, C, s, Z) = &-\frac{1}{n}\sum_{i=1}^n \ell_i^*(\alpha_i) + \mathrm{Tr}\Big(H\big(I_d \otimes \Phi(s)\big)\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\big(I_d \otimes \Phi(s)^\top\big)\Big) \\
&+ \mathrm{Tr}\Big(C\Big(-\frac{X^\top \alpha\alpha^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}\Big)\Big).
\end{aligned} \tag{81}$$

Hence, applying [15, Lemma 44], we know that, once computed a maximizer $\alpha_{\tau_{H,C}(s)}$ of the function above $\alpha \in \mathbb{R}^n \mapsto Q(\alpha, H, C, s, Z)$,

$$\nabla Q(\alpha_{\tau_{H,C}(s)}, \cdot, \cdot, s, Z)(H, C) \in \frac{\partial \Delta(\tau_{H,C}(s), Z)}{\partial (H, C)} = \frac{\partial \mathcal{L}\big(H, C, s, Z\big)}{\partial (H, C)}. \tag{82}$$

As a consequence, since for a given matrix A, $\nabla \mathrm{Tr}\big(\,\cdot\, A\big)(H) = A$, we get that

$$\nabla \mathcal{L}\big(\cdot, \cdot, s, Z\big)(H, C) = \Big(\big(I_d \otimes \Phi(s)\big)\hat{\nabla}\big(I_d \otimes \Phi(s)^\top\big), \hat{\nabla}\Big) \in \frac{\partial \mathcal{L}\big(H, C, s, Z\big)}{\partial (H, C)}, \tag{83}$$

with

$$\hat{\nabla} = -\frac{X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X}{2n^2} + \frac{2L^2 X^\top X}{n^2}. \tag{84}$$

Finally, in order to get the desired closed form in Prop. 4, we just need to observe that, according to the optimality conditions of the within-task problem in (see [15, Lemma 44]) with $\theta \in \mathbb{S}_+^d$, we have that

$$X^\top \alpha_\theta = -n\theta^\dagger w_\theta. \tag{85}$$

As a consequence, we can rewrite Eq. (84) as follows by using the primal solution of the within-task problem:

$$\hat{\nabla} = -\frac{\lambda}{2}\tau_{H,C}(s)^\dagger w_{\tau_{H,C}(s)} w_{\tau_{H,C}(s)}^\top \tau_{H,C}(s)^\dagger + \frac{2L^2 X^\top X}{n^2}. \tag{86}$$

Finally, we observe that, by the closed form in Prop. 4,

$$\big\|\nabla \mathcal{L}\big(\cdot, \cdot, s, Z\big)(H, C)\big\|_F \leq \big\|\nabla \mathcal{L}\big(\cdot, \cdot, s, Z\big)(H, C)\big\|_* \leq \mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{D} \tag{87}$$

with

$$\begin{aligned}
\mathrm{A} &= \left\| \big(I_d \otimes \Phi(s)\big) \frac{X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X}{2n^2} \big(I_d \otimes \Phi(s)^\top\big) \right\|_* \\
\mathrm{B} &= \left\| \big(I_d \otimes \Phi(s)\big) \frac{2L^2 X^\top X}{n^2} \big(I_d \otimes \Phi(s)^\top\big) \right\|_* \\
\mathrm{C} &= \left\| \frac{X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X}{2n^2} \right\|_* \\
\mathrm{D} &= \left\| \frac{2L^2 X^\top X}{n^2} \right\|_*.
\end{aligned} \tag{88}$$

We now observe that all the matrices inside the trace norms above are positive semidefinite (as a matter of fact, if a matrix Q is positive semidefinite, then, $P^\top Q P$ is positive semidefinite for any matrix P). As a consequence, all the trace norms above coincide with the trace of the corresponding matrices, namely,

$$\begin{aligned}
\mathrm{A} &= \mathrm{Tr}\left(\big(I_d \otimes \Phi(s)\big) \frac{X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X}{2n^2} \big(I_d \otimes \Phi(s)^\top\big) \right) \\
\mathrm{B} &= \mathrm{Tr}\left(\big(I_d \otimes \Phi(s)\big) \frac{2L^2 X^\top X}{n^2} \big(I_d \otimes \Phi(s)^\top\big) \right) \\
\mathrm{C} &= \mathrm{Tr}\left(\frac{X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X}{2n^2} \right) \\
\mathrm{D} &= \mathrm{Tr}\left(\frac{2L^2 X^\top X}{n^2} \right).
\end{aligned} \tag{89}$$

We now observe that, proceeding as above in Eq. (64) and exploiting Asm. 3, we can write

$$\begin{aligned} \mathrm{A} &\leq \left\|\Phi(s)\right\|^2 \mathrm{Tr}\left(\frac{X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X}{2n^2}\right) = \left\|\Phi(s)\right\|^2 \mathrm{C} \leq K^2 \mathrm{C} \\ \mathrm{B} &\leq \left\|\Phi(s)\right\|^2 \mathrm{Tr}\left(\frac{2L^2 X^\top X}{n^2}\right) = \left\|\Phi(s)\right\|^2 \mathrm{D} \leq K^2 \mathrm{D}. \end{aligned} \tag{90}$$

Hence, combining everything in Eq. (87), we get

$$\left\|\nabla \mathcal{L}\big(\cdot, \cdot, s, Z\big)(H, C)\right\|_F \leq (1 + K^2)\big(\mathrm{C} + \mathrm{D}\big). \tag{91}$$

The desired statement derives from observing that, since, by Asm. 1, $\mathrm{Tr}\big(X^\top \alpha_{\tau_{H,C}(s)} \alpha_{\tau_{H,C}(s)}^\top X\big) \leq (nLR)^2$ (see [15, Lemma 44]) and $\mathrm{Tr}\big(X^\top X\big) = \mathrm{Tr}\big(X X^\top\big) = \sum_{i=1}^n \|x_i\|^2 \leq nR^2$, then

$$\mathrm{C} \leq \frac{(LR)^2}{2\lambda} \qquad \mathrm{D} \leq \frac{2(LR)^2}{n}. \tag{92}$$

□

D.3 Convergence rate of Alg. 1 on the surrogate problem in Eq. (15)

We now give the convergence rate of Alg. 1 on the surrogate problem in Eq. (15).

Proposition 12 (Convergence rate on the surrogate problem in Eq. (15))**.** *Let $\bar{H}$ and $\bar{C}$ be the average of the iterations obtained from the application of Alg. 1 over the training data $(Z_t, s_t)_{t=1}^T$ with constant meta-step size $\gamma > 0$. Then, under Asm. 1 and Asm. 3, for any $\tau_{H,C} \in \mathcal{T}_\Phi$, in expectation with respect to the sampling of $(Z_t, s_t)_{t=1}^T$,*

$$\mathbb{E}\, \hat{\mathcal{E}}_\rho\big(\tau_{\bar{H},\bar{C}}\big) - \hat{\mathcal{E}}_\rho\big(\tau_{H,C}\big) \leq \frac{\gamma(1 + K^2)^2 (LR)^4}{2\lambda^2} \Big(\frac{1}{2} + \frac{2}{n}\Big)^2 + \frac{\left\|(H - H_0, C - C_0)\right\|_F^2}{2\gamma T}. \tag{93}$$

Proof. We observe that Alg. 1 coincides with projected Stochastic Gradient Descent applied to the convex and Lipschitz (see Prop. 4) surrogate problem in Eq. (15):

$$\min_{H \in \mathbb{S}_+^{dk}, C \in \mathbb{S}_+^d} \hat{\mathcal{E}}_\rho(\tau_{H,C}) \qquad \hat{\mathcal{E}}_\rho(\tau_{H,C}) = \mathbb{E}_{(\mu,s)\sim\rho}\, \mathbb{E}_{Z\sim\mu^n}\, \mathcal{L}\big(H, C, s, Z\big). \tag{94}$$

As a consequence, by standard arguments (see e.g. [33, Lemma 14.1, Thm. 14.8] and references therein), for any $\tau_{H,C} \in \mathcal{T}_\Phi$, we have

$$\mathbb{E}\, \hat{\mathcal{E}}_\rho\big(\tau_{\bar{H},\bar{C}}\big) - \hat{\mathcal{E}}_\rho\big(\tau_{H,C}\big) \leq \frac{\gamma}{2T} \sum_{t=1}^T \mathbb{E} \left\|\nabla \mathcal{L}\big(\cdot, \cdot, s, Z_t\big)(H_t, C_t)\right\|_F^2 + \frac{\left\|(H - H_0, C - C_0)\right\|_F^2}{2\gamma T}. \tag{95}$$

The desired statement derives from combining this bound with the bound on the norm of the meta-subgradients in Prop. 4. □

D.4 Proof of Thm. 5

We now have all the ingredients necessary to prove Thm. 5.

Theorem 5 (Excess risk bound for the conditioning function returned by Alg. 1)**.** *Let Asm. 1 and Asm. 3 hold. For any $s \sim \rho_{\mathcal{S}}$, recall the conditional covariance matrices $W(s)$ and $C(s)$ introduced in Thm. 1. Let $\tau_{H,C}$ be a fixed function in $\mathcal{T}_\Phi$ such that $\mathrm{Ran}(W(s)) \subseteq \mathrm{Ran}(\tau_{H,C}(s))$ for any $s \sim \rho_{\mathcal{S}}$. Let $\bar{H}$ and $\bar{C}$ be the outputs of Alg. 1 applied to a sequence $(Z_t, s_t)_{t=1}^T$ of i.i.d. pairs sampled from ρ with an appropriate meta-step size γ. Then, in expectation with respect to the sampling of $(Z_t, s_t)_{t=1}^T$,*

$$\begin{aligned} \mathbb{E}\, \mathcal{E}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* \;\leq\; & \frac{\mathbb{E}_{s\sim\rho_{\mathcal{S}}} \mathrm{Tr}\big(\tau_{H,C}(s)^\dagger W(s)\big)}{2} + \frac{2L^2 \mathbb{E}_{s\sim\rho_{\mathcal{S}}} \mathrm{Tr}\big(\tau_{H,C}(s) C(s)\big)}{n} \\ & + \Big(\frac{1}{2} + \frac{2}{n}\Big) \frac{(1 + K^2)(LR)^2 \left\|(H - H_0, C - C_0)\right\|_F}{\sqrt{T}}. \end{aligned}$$

Proof. We start from observing thta, in expectation with respect to the meta-training set, for any fixed conditioning function $\tau_{H,C} \in \mathcal{T}_\Phi$, we can write the following decomposition

$$\begin{aligned}\mathbb{E}\,\mathcal{E}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* &\leq \mathbb{E}\,\hat{\mathcal{E}}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* \\ &= \mathbb{E}\,\hat{\mathcal{E}}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* \pm \hat{\mathcal{E}}_\rho(\tau_{H,C}) \\ &= \underbrace{\mathbb{E}\,\hat{\mathcal{E}}_\rho(\tau_{\bar{H},\bar{C}}) - \hat{\mathcal{E}}_\rho(\tau_{H,C})}_{\mathrm{A}(\tau_{H,C})} + \underbrace{\hat{\mathcal{E}}_\rho(\tau_{H,C}) - \mathcal{E}_\rho^*}_{\mathrm{B}(\tau_{H,C})},\end{aligned} \tag{96}$$

where in the inequality above we have exploited the fact that, for any $\tau \in \mathcal{T}$, $\mathcal{E}_\rho(\tau) \leq \hat{\mathcal{E}}_\rho(\tau)$ We now observe that the term $\mathrm{A}(\tau_{H,C})$ can be controlled according to the convergence properties of the meta-algorithm in Alg. 1 as described in Prop. 12:

$$\mathbb{E}\,\hat{\mathcal{E}}_\rho\big(\tau_{\bar{H},\bar{C}}\big) - \hat{\mathcal{E}}_\rho\big(\tau_{H,C}\big) \leq \frac{\gamma(1+K^2)^2(LR)^4}{2}\Big(\frac{1}{2}+\frac{2}{n}\Big)^2 + \frac{\big\|(H-H_0, C-C_0)\big\|_F^2}{2\gamma T}. \tag{97}$$

Regarding the term $\mathrm{B}(\tau_{H,C})$, we observe that, for any τ, we can rewrite

$$\begin{aligned}\mathrm{B}(\tau) &= \hat{\mathcal{E}}_\rho(\tau) - \mathcal{E}_\rho^* \\ &= \mathbb{E}_{(\mu,s)\sim\rho}\,\mathbb{E}_{Z\sim\mu^n}\Big[\mathcal{R}_{Z,\tau(s)}(A(\tau(s),Z)) - \mathcal{R}_\mu(w_\mu)\Big] + \frac{2L^2\mathbb{E}_{(\mu,s)\sim\rho}\,\mathrm{Tr}\big(\tau(s)\mathbb{E}_{x\sim\eta_\mu} xx^\top\big)}{n} \\ &\leq \frac{\mathbb{E}_{(\mu,s)\sim\rho}\,\mathrm{Tr}\big(\tau(s)^\dagger w_\mu w_\mu^\top\big)}{2} + \frac{2L^2\mathbb{E}_{(\mu,s)\sim\rho}\,\mathrm{Tr}\big(\tau(s)\mathbb{E}_{x\sim\eta_\mu} xx^\top\big)}{n},\end{aligned} \tag{98}$$

where in the inequality we have exploited the fact that, thanks to the definition of the algorithm, for any $(\mu, s) \sim \rho$, we can write

$$\mathbb{E}_{Z\sim\mu^n}\Big[\mathcal{R}_{Z,\tau(s)}(A(\tau(s),Z)) - \mathcal{R}_\mu(w_\mu)\Big] \leq \frac{\mathrm{Tr}\big(\tau(s)^\dagger w_\mu w_\mu^\top\big)}{2}. \tag{99}$$

Combining the bounds on the two terms above in Eq. (96), we get

$$\begin{aligned}\mathbb{E}\,\mathcal{E}_\rho(\tau_{\bar{H},\bar{C}}) - \mathcal{E}_\rho^* \leq& \frac{\mathbb{E}_{(\mu,s)\sim\rho}\,\mathrm{Tr}\big(\tau_{H,C}(s)^\dagger w_\mu w_\mu^\top\big)}{2} + \frac{2L^2\mathbb{E}_{(\mu,s)\sim\rho}\,\mathrm{Tr}\big(\tau_{H,C}(s)\mathbb{E}_{x\sim\eta_\mu} xx^\top\big)}{n} \\ &+ \frac{\gamma(1+K^2)^2(LR)^4}{2}\Big(\frac{1}{2}+\frac{2}{n}\Big)^2 + \frac{\big\|(H-H_0, C-C_0)\big\|_F^2}{2\gamma T}.\end{aligned} \tag{100}$$

Optimizing with respect to the hyper-parameter $\gamma > 0$, for

$$\gamma \;=\; \frac{\|(H-H_0, C-C_0)\|_F}{(1+K^2)(LR)^2}\,\Big(\frac{1}{2}+\frac{2}{n}\Big)^{-1}\,\frac{1}{\sqrt{T}}, \tag{101}$$

we get the desired statement. □

E Experimental details

We now report the experimental details we missed in the main body. Specifically, we report the details regarding the tuning of the hyper-parameter γ and the characteristics of the machine we used for running our experiments.

Synthetic clusters. In order to tune the hyper-parameter γ we applied the procedure above with 14 candidates values for γ in the range $[10^{-5}, 10^5]$ with logarithmic spacing and we evaluated the performance of the estimated meta-parameters (linear representations) by using $T = T_{\mathrm{tr}} = 500$, $T_{\mathrm{va}} = 300$, $T_{\mathrm{te}} = 100$ of the available tasks for meta-training, meta-validation and meta-testing, respectively. In order to train and to test the inner algorithm, we splitted each within-task dataset into $n = n_{\mathrm{tr}} = 50\%\ n_{\mathrm{tot}}$ for training and $n_{\mathrm{te}} = 50\%\ n_{\mathrm{tot}}$ for test.

Lenk dataset. In order to tune the hyper-parameter γ we applied the procedure above with 14 candidates values for γ in the range $[10^{-5}, 10^5]$ with logarithmic spacing and we evaluated the

performance of the estimated meta-parameters (linear representations) by using $T = T_{\rm tr} = 100$, $T_{\rm va} = 40$, $T_{\rm te} = 30$ of the available tasks for meta-training, meta-validation and meta-testing, respectively. In order to train and to test the inner algorithm, we splitted each within-task dataset into $n = n_{\rm tr} = 16$ for training and $n_{\rm te} = 4$ for test.

Movieles-100k dataset. In order to tune the hyper-parameter γ we applied the procedure above with 14 candidates values for γ in the range $[10^{-5}, 10^{5}]$ with logarithmic spacing and we evaluated the performance of the estimated meta-parameters (linear representations) by using $T = T_{\rm tr} = 200$, $T_{\rm va} = 100$, $T_{\rm te} = 100$ of the available tasks for meta-training, meta-validation and meta-testing, respectively. In order to train and to test the inner algorithm, we splitted each within-task dataset into $n = n_{\rm tr} = 15$ for training and $n_{\rm te} = 5$ for test.

Jester-1 dataset. In order to tune the hyper-parameter γ we applied the procedure above with 14 candidates values for γ in the range $[10^{-5}, 10^{5}]$ with logarithmic spacing and we evaluated the performance of the estimated meta-parameters (linear representations) by using $T = T_{\rm tr} = 250$, $T_{\rm va} = 100$, $T_{\rm te} = 100$ of the available tasks for meta-training, meta-validation and meta-testing, respectively. In order to train and to test the inner algorithm, we splitted each within-task dataset into $n = n_{\rm tr} = 15$ for training and $n_{\rm te} = 5$ for test.

All the experiments were conducted on a workstation with 4 Intel Xeon E5-2697 V3 2.60Ghz CPUs and 256GB RAM.