Comparison bewteen multi-task and single-task oracle risks in kernel ridge regression

Matthieu Solnon

Comparison bewteen multi-task and single-task oracle risks in kernel ridge regression

Matthieu Solnon

ENS; Sierra Project-team
Département d'Informatique de l'École Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France
e-mail: matthieu.solnon@ens.fr

Abstract: In this paper we study multi-task kernel ridge regression and try to understand when the multi-task procedure performs better than the single-task one, in terms of averaged quadratic risk. In order to do so, we compare the risks of the estimators with perfect calibration, the *oracle risk*. We are able to give explicit settings, favorable to the multi-task procedure, where the multi-task oracle performs better than the single-task one. In situations where the multi-task procedure is conjectured to perform badly, we also show the oracle does so. We then complete our study with simulated examples, where we can compare both oracle risks in more natural situations. A consequence of our result is that the multi-task ridge *estimator* has a lower risk than any single-task estimator, in favorable situations.

MSC 2010 subject classifications: Primary 62H05; secondary 62C25, 62G08, 62J07, 68Q32.
Keywords and phrases: Kernel methods, Multi-task, Oracle risk, Ridge regression.

Contents

1. Introduction

Increasing the sample size is the most common way to improve the performance of statistical estimators. In some cases (see, for instance, the experiments of Evgeniou et al. [13] on customer data analysis or those of Jacob et al. [18] on molecule binding problems), having access to some new data may be impossible, often due to experimental limitations. One way to circumvent those constraints is to use datasets from several related (and, hopefully, "similar") problems, as if it gave additional (in some sense) observations on the initial problem. The statistical methods using this heuristic are called "multi-task" techniques, as opposed to "single-task" techniques, where every problem is treated one at a time. In this paper, we study kernel ridge regression in a multi-task framework and try to understand when multi-task can improve over single-task.

The first trace of a multi-task estimator can be found in the work of Stein [28]. In this article, Charles Stein showed that the usual maximum-likelihood

estimator of the mean of a Gaussian vector (of dimension larger than 3, every dimension representing here a task) is not admissible—that is, there exists another estimator that has a lower risk for every parameter. He showed the existence of an estimator that uniformly attains a lower quadratic risk by shrinking the estimators along the different dimensions towards an arbitrary point. An explicit form of such an estimator was given by James and Stein [19], yielding the famous James-Stein estimator. This phenomenon, now known as the "Stein's paradox", was widely studied in the following years and the behaviour of this estimator was confirmed by empirical studies, in particular the one from Efron and Morris [12]. This first example clearly shows the goals of the multi-task procedure: an advantage is gained by borrowing information from different tasks (here, by shrinking the estimators along the different dimensions towards a common point), the improvement being scored by the global (averaged) squared risk. Therefore, this procedure does not guarantee individual gains on every task, but a global improvement on the sum of those task-wise risks.

We consider here $p \geq 2$ different regression tasks, a framework we refer to as "multi-task" regression, and where the performance of the estimators is measured by the fixed-design quadratic risk. Kernel ridge regression is a classical framework to work with and comes with a natural norm, which often has desirable properties (such as, for instance, links with regularity). This norm is also a natural "similarity measure" between the regression functions. Evgeniou et al. [13] showed how to extend kernel ridge regression to a multi-task setting, by adding a regularization term that binds the regression functions along the different tasks together. One of the main questions that is asked is to assert whether the multi-task estimator has a lower risk than any single-task estimator. It was recently proved by Solnon et al. [27] that a fully data-driven calibration of this procedure is possible, given some assumptions on the set of matrices used to regularize—which correspond to prior knowledge on the tasks. Under those assumptions, the estimator is showed to verify an *oracle inequality*, that is, its risk matches (up to constants) the best possible one, the *oracle risk*. Thus, it suffices to compare the oracle risks for the multi-task procedure and the single-task one to provide an answer to this question.

The multi-task regression setting, which could also be called "multivariate regression", has already been studied in different papers. It was first introduced by Brown and Zidek [9] in the case of ridge regression, and then adapted by Evgeniou et al. [13] in its kernel form. Another view of the meaning of "task similarity" is that the functions all share a few common features, and can be expressed by a similar regularization term. This idea was expressed in a linear set up (also known as group lasso) by Obozinski et al. [25] and Lounici et al. [23], in multiple kernel learning by Koltchinskii and Yuan [21] or in semi-supervised learning by Ando and Zhang [1]. The kernel version of this was also studied [2, 18], a convex relaxation leading to a trace norm regularization and allowing the calibration of parameters. Another point of view was brought by Ben-David and Schuller [8], defining a multi-task framework in classification, two classification

problems being similar if, given a group of permutations of the input set, a dataset of the one can be permuted in a dataset of the other. They followed the analysis of Baxter [7], which shows very general bounds on the risk of a multi-task estimator in a model-selection framework, the sets of all models reflecting the insight the statistician has on the multi-task setting.

Advantages of the multi-task procedure over the single task one were first shown experimentally in various situations by, for instance, Thrun and O'Sullivan [29], Caruana [11] or Bakker and Heskes [6]. For classification, Ben-David and Schuller [8] compare upper bounds on multi-task and single-task classification errors, and showed that the multi-task estimator could, in some settings, need less training data to reach the same upper bounds. The low dimensional linear regression setting was analysed by Rohde and Tsybakov [26], who showed that, under sparsity assumptions, restricted isometry conditions and using the trace-norm regularization, the multi-task estimator achieves the rates of a single-task estimator with a np-sample. Liang et al. [22] also obtained a theoretical criterion, applicable to the linear regression setting and unfortunately non observable, which tells when the multi-task estimator asymptotically has a lower risk than the lower one. A step was recently carried by Feldman et al. [14] in a kernel setting where every function is estimated by a constant. They give a closed-form expression of the oracle for two tasks and run simulations to compare the risk of the multi-task estimator to the risk of the single-task estimator.

In this article we study the oracle multi-task risk and compare it to the oracle single-task risk. We then find situations where the multi-task oracle is proved to have a lower risk than the single-task oracle. This allows us to better understand which situation favors the multi-task procedure and which does not. After having defined our model (Section 2), we write down the risk of a general multi-task ridge estimator and see that it admits a convenient decomposition using two key elements: the mean of the tasks and the resulting variance (Section 3). This decomposition allows us to optimize this risk and get a precise estimation of the oracle risk, in settings where the ridge estimator is known to be minimax optimal (Section 4). We then explore several repartitions of the tasks that give the latter multi-task rates, study their single-task oracle risk (Section 5) and compare it to their respective multi-task rates. This allows us to discriminate several situations, depending whether the multi-task oracle either outperforms its single-task counterpart, underperforms it or whether both behave similarly (Section 6). We also show that, in the cases favorable to the multi-task oracle detailed in the previous sections, the estimator proposed by Solnon et al. [27] behaves accordingly and achieves a lower risk than the single-task oracle (Section 7). We finally study settings where we can no longer explicitly study the oracle risk, by running simulations, and we show that the multi-task oracle continues to retain the same virtues and disadvantages as before (Section 8).

We now introduce some notations, which will be used throughout the article.

- The integer n is the sample size, the integer p is the number of tasks.

- For any $n \times p$ matrix Y, we define

$$y = \mathrm{vec}(Y) := (Y_{1,1}, \ldots, Y_{n,1}, Y_{1,2}, \ldots, Y_{n,2}, \ldots, Y_{1,p}, \ldots, Y_{n,p}) \in \mathbb{R}^{np},$$

  that is, the vector in which the columns $Y^j := (Y_{i,j})_{1 \leq i \leq n}$ are stacked.
- $\mathcal{M}_n(\mathbb{R})$ is the set of all real square matrices of size n.
- $\mathcal{S}_p(\mathbb{R})$ is the set of symmetric matrices of size p.
- $\mathcal{S}_p^+(\mathbb{R})$ is the set of symmetric positive-semidefinite matrices of size p.
- $\mathcal{S}_p^{++}(\mathbb{R})$ is the set of symmetric positive-definite matrices of size p.
- $\mathbf{1}$ is the vector of size p whose components are all equal to 1.
- $\|\cdot\|_2$ is the usual Euclidean norm on $\mathbb{R}^k$ for any $k \in \mathbb{N}$: $\forall u \in \mathbb{R}^k$, $\|u\|_2^2 := \sum_{i=1}^k u_i^2$.
- For two real sequences (u_n) and (v_n) we write $u_n \asymp v_n$ if there exists positive constants ℓ and L such that, for a large enough n, $\ell v_n \leq u_n \leq L v_n$.
- For $(a, b) \in \mathbb{R}^2$, $a \wedge b$ denotes the minimum of a and b.

2. Kernel ridge regression in a multi-task setting

We consider here that each task is treated as a kernel ridge-regression problem and we will then extend the single-task ridge-regression estimator in a multi-task setting.

2.1. Model and estimator

Let Ω be a set, $\mathcal{A}$ be a σ-algebra on Ω and $\mathbb{P}$ be a probability measure on $\mathcal{A}$. We observe $\mathcal{D}_n = (X_i, Y_i^1, \ldots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$. For each task $j \in \{1, \ldots, p\}$, $\mathcal{D}_n^j = (X_i, y_i^j)_{i=1}^n$ is a sample with distribution $\mathcal{P}^j$, whose first marginal distribution is $\mathcal{P}$, for which a simple regression problem has to be solved.

We assume that for every $j \in \{1, \ldots, p\}$, $F^j \in L^2(\mathbb{P})$, Σ is a symmetric positive-definite matrix of size p such that the vectors $(\varepsilon_i^j)_{j=1}^p$ are independent and identically distributed (i.i.d.) with normal distribution $\mathcal{N}(0, \Sigma)$, with mean zero and covariance matrix Σ, and

$$\forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, p\}, \; y_i^j = F^j(X_i) + \varepsilon_i^j \; . \tag{1}$$

We suppose here, for simplicity, that $\Sigma = \sigma^2 I_p$, with $\sigma^2 \in \mathbb{R}_+^\star$.

Remark 1. *This implies that the outputs of every task are independent, which slightly simplifies the setting but allow lighter calculations. It is to be noted, though, that the analysis carried afterwards can still take place without this assumption. This can be dealt by diagonalizing Σ, majoring the quantities of interest using the largest eigenvalue of Σ and minoring those quantities by its smallest eigenvalue. The comparisons shown in Section 6 are still valid, only being enlarged by the condition number of Σ. A fully data-driven estimation of Σ was proposed by Solnon et al. [27].*

We consider here a fixed-design setting, that is, we consider the input points as fixed and want to predict the output of the functions F^j on those input points only. The analysis could be transfered to the random-design setting by using tools developped by Hsu et al. [17].

For an estimator $(\widehat{F}^1, \dots, \widehat{F}^p)$, the natural quadratic risk to consider is

$$\mathbb{E}\left[\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}(\widehat{F}^j(X_i) - F^j(X_i))^2 | (X_1, \dots, X_n)\right] \quad .$$

For the sake of simplicity, all the expectations that follow will implicitly be written conditional on $(X_1, \dots, X_n)$. This corresponds to the fixed-design setting, which treats the input points as fixed.

Remark 2. *We will use the following notations from now on :*

$$f = \mathrm{vec}\left((f^j(X_i))_{i,j}\right), \quad f^j = \mathrm{vec}\left((f^j(X_i))_{i=1}^n\right) \quad \textit{and} \quad y = \mathrm{vec}\left((Y_i^j)_{i,j}\right) \quad ,$$

so that, when using such vectorized notations, the elements are stacked task by task, the elements refering to the first task always being stored in the first entries of the vector, and so on.

We want to estimate f using elements of a particular function set. Let $\mathcal{F} \subset L^2(\mathbb{P})$ be a reproducing kernel Hilbert space (RKHS) [4], with kernel k and feature map $\Phi : \mathcal{X} \to \mathcal{F}$, which give us the positive semidefinite kernel matrix $K = (k(X_i, X_\ell))_{1 \leq i,\ell \leq n} \in \mathcal{S}_n^+(\mathbb{R})$.

As done by Solnon et al. [27] we extend the multi-task estimators generalizing the ridge-regression used in Evgeniou et al. [13]. Given a positive-definite matrix $M \in \mathcal{S}_p^{++}(\mathbb{R})$, we consider the estimator

$$\widehat{F}_M \in \operatorname*{argmin}_{g \in \mathcal{F}^p} \left\{ \underbrace{\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2}_{\text{Empirical risk}} + \underbrace{\sum_{j=1}^{p}\sum_{\ell=1}^{p} M_{j,l}\langle g^j, g^\ell\rangle_{\mathcal{F}}}_{\text{Regularization term}} \right\} . \qquad (2)$$

This leads to the fixed-design estimator

$$\widehat{f}_M = A_M y \in \mathbb{R}^{np} \quad ,$$

with

$$A_M = A_{M,K} := \widetilde{K}_M(\widetilde{K}_M + npI_{np})^{-1} = (M^{-1} \otimes K)\left((M^{-1} \otimes K) + npI_{np}\right)^{-1} \quad ,$$

where $\otimes$ denotes the Kronecker product (see the textbook of Horn and Johnson [16] for simple properties of the Kronecker product).

Remark 3. *This setting also captures the single-task setting. Taking $j \in \{1, \dots, p\}$, $f^j = (f^j(X_1), \dots, f^j(X_n))^\top$ being the target-signal for the jth task and $y^j = (y_1^j, \dots, y_n^j)^\top$ being the observed output of the jth task, the single-task estimator for the jth task becomes (for $\lambda \in \mathbb{R}_+$)*

$$\widehat{f}_\lambda^j = A_\lambda y^j = K(K + n\lambda I_n)^{-1} y^j \quad .$$

2.2. Two regularization terms for one problem

A common hypothesis that motivates the use of multi-task estimators is that all the target functions of the different tasks lie in a single cluster (that is, the p functions that are estimated are all close with respect to the norm defined on $\mathcal{F}$). Two different regularization terms are usually considered in this setting:

- one that penalizes the norms of the p function and their differences, introduced by Evgeniou et al. [13], leading to the criterion (with $(g^j)_{j=1}^p \in \mathcal{F}^p,\ (\alpha,\beta) \in (\mathbb{R}_+)^2$)

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2 + \frac{\alpha}{p}\sum_{j=1}^{p}\left\|g^j\right\|_{\mathcal{F}}^2 + \frac{\beta}{2p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left\|g^j - g^k\right\|_{\mathcal{F}}^2 \quad ; \quad (3)$$

- one that penalizes the norms of the average of the p functions and the resulting variance, leading to the criterion (with $(g^j)_{j=1}^p \in \mathcal{F}^p,\ (\lambda,\mu) \in (\mathbb{R}_+)^2$)

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2 + \lambda\left\|\frac{\sum_{j=1}^{p} g^j}{p}\right\|_{\mathcal{F}}^2 + \mu\left[\frac{\sum_{j=1}^{p}\left\|g^j\right\|_{\mathcal{F}}^2}{p} - \left\|\frac{\sum_{j=1}^{p} g^j}{p}\right\|_{\mathcal{F}}^2\right] . \quad (4)$$

As we will see, those two penalties are closely related. Lemma 1 indeed shows that the two former penalties can be obtained as a special case of Equation (2), the matrix M being respectively

$$M_{\mathrm{SD}}(\alpha,\beta) := \frac{\alpha}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\alpha+p\beta}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right)$$

and

$$M_{\mathrm{AV}}(\lambda,\mu) := \frac{\lambda}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) .$$

Thus, we see that those two criteria are related, since $M_{\mathrm{SD}}(\alpha,\beta) = M_{\mathrm{AV}}(\alpha,\alpha+p\beta)$ for every (α,β). Minimizing Equations (3) and (4) over $\mathcal{F}^p$ respectly give the ridge estimators $\widehat{f}_{SD}(\alpha,\beta) = A_{M_{\mathrm{SD}}(\alpha,\beta)}Y$ and $\widehat{f}_{\mathrm{AV}}(\lambda,\mu) = A_{M_{\mathrm{AV}}(\lambda,\mu)}Y$.

Remark 4. *We can now see that the regularization terms used in Equations (3) and (4) are equivalent when the parameters are not constrained to be positive. However, if one desires to use the regularization (3) (that is, with $\lambda = \alpha$ and $\mu = \alpha + p\beta$) and seeks to calibrate those parameters by taking them to be non-negative (which is to be expected if they are seen as regularization parameters), the following problems could occur:*

- *if the optimization is carried over (λ,μ), then the selected parameter $\beta = \frac{\mu-\lambda}{p}$ may be negative;*
- *conversely, if the risk of the estimator defined by Equation (3) is optimized over the parameters $(\alpha,\alpha+p\beta)$ with the constraints $\alpha \geq 0$ and $\beta \geq 0$, then the infimum over $\mathbb{R}_+^2$ could never be approached.*

We will also show in the next section that the risk of $\widehat{f}_{\mathrm{AV}}(\lambda,\mu)$ nicely decomposes in two parts, the first part depending only on λ and the second only on μ, which is not the case for $\widehat{f}_{SD}(\alpha,\beta)$ because of the aforementioned phenomenon. This makes us prefer the second formulation and use the matrices M_{AV} instead of the matrices M_{SD}.

3. Decomposition of the risk

A fully data-driven selection of the hyper-parameters was proposed by Arlot and Bach [3], for the single-task ridge estimator, and by Solnon et al. [27] for the multi-task estimator. The single-task estimator is shown to have a risk which is close to the single-task oracle-risk (with a fixed-design)

$$\mathfrak{R}^{\star}_{\mathrm{ST}} = \inf_{(\lambda^1,\dots,\lambda^p)\in\mathbb{R}^p_+} \left\{ \frac{1}{np}\mathbb{E}\left[\sum_{j=1}^{p}\left\|\widehat{f}^{j}_{\lambda^j} - f^j\right\|_2^2\right]\right\} \ ,$$

while the multi-task estimator is shown to have a risk which is close to the multi-task oracle risk

$$\mathfrak{R}^{\star}_{\mathrm{MT}} = \inf_{(\lambda,\mu)\in\mathbb{R}^2_+} \left\{ \frac{1}{np}\mathbb{E}\left[\left\|\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)} - f\right\|_2^2\right]\right\} \ .$$

The purpose of this paper is to closely study both oracle risks and, ultimately, to compare them. We show in this section how to decompose the risk of an estimator obtained by minimizing Equation (4) over $(g^j)_{j=1}^p \in \mathcal{F}^p$. A key point of this analysis is that the matrix $M_{\mathrm{AV}}(\lambda,\mu)$ naturally decomposes over two orthogonal vector-subspaces of $\mathbb{R}^p$. By exploiting this decomposition we can simply use the classical bias-variance decomposition to analyse the Euclidean risk of those linear estimators.

3.1. Eigendecomposition of the matrix $M_{\mathrm{AV}}(\lambda,\mu)$

In this section we show that all the matrices $M_{\mathrm{AV}}(\lambda,\mu)$ have the same eigenvectors, which gives us a simple decomposition of the matrices $A_{M_{\mathrm{AV}}(\lambda,\mu)}$. Let us denote by $(e_1,\dots,e_p)$ the canonical basis of $\mathbb{R}^p$. The eigenspaces of $p^{-1}\mathbf{1}\mathbf{1}^\top$ are orthogonal and correspond to:

- span $\{e_1+\dots+e_p\}$ associated to eigenvalue 1,
- span $\{e_2-e_1,\dots,e_p-e_1\}$ associated to eigenvalue 0.

Thus, with $(\lambda,\mu)\in(R^+)^2$, we can diagonalize in an orthonormal basis any matrix $M_{\mathrm{AV}}(\lambda,\mu)$ as $M = M_{\mathrm{AV}}(\lambda,\mu) = P^\top D_{\frac{\lambda}{p},\frac{\mu}{p}} P$, with $D = \mathrm{Diag}\{\frac{\lambda}{p},\frac{\mu}{p},\dots,\frac{\mu}{p}\} = D_{\frac{\lambda}{p},\frac{\mu}{p}}$. Let us also diagonalize K in an orthonormal basis : $K = Q^\top \Delta Q$, $\Delta = \mathrm{Diag}\{\gamma_1,\dots,\gamma_n\}$. Then

$$A_M = A_{M_{\mathrm{AV}}(\lambda,\mu)} = (P^\top\otimes Q^\top)\left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta)+npI_{np}\right)^{-1}\right](P\otimes Q) \ .$$

We can then note that $(D^{-1} \otimes \Delta)\left((D^{-1} \otimes \Delta) + npI_{np}\right)^{-1}$ is a diagonal matrix, whose diagonal entry of index $(j-1)n+i$ ($i \in \{1,\dots,n\}$, $j \in \{1,\dots,p\}$) is

$$\begin{cases} \frac{\gamma_i}{\gamma_i + n\lambda} \text{ if } j = 1 \ , \\ \frac{\gamma_i}{\gamma_i + n\mu} \text{ if } j > 1 \ . \end{cases}$$

In the following section we will use the following notations :

- for every $j \in \{1,\dots,p\}$, $(h_i^j)_{i=1}^n$ denotes the coordinates of $(f^j(X_i))_{i=1}^n$ in the basis that diagonalizes K,
- for every $i \in \{1,\dots,n\}$, $(\nu_i^j)_{j=1}^p$ denotes the coordinates of $(h_i^j)_{j=1}^p$ in the basis that diagonalizes M.

Or, to sum up, we have :

$$\forall j \in \{1,\dots,p\}, \quad \begin{pmatrix} h_1^j \\ \vdots \\ h_n^j \end{pmatrix} = Q \begin{pmatrix} f^j(X_1) \\ \vdots \\ f^j(X_n) \end{pmatrix}$$

and

$$\forall i \in \{1,\dots,n\}, \quad \begin{pmatrix} \nu_i^1 \\ \vdots \\ \nu_i^p \end{pmatrix} = P \begin{pmatrix} h_i^1 \\ \vdots \\ h_i^p \end{pmatrix} \ .$$

With the usual notation $\nu^j = (\nu_1^j,\dots,\nu_n^j)^\top$ and f , we get, by using elementary properties of the Kronecker product,

$$\nu = \begin{pmatrix} \nu^1 \\ \vdots \\ \nu^p \end{pmatrix} = (P \otimes Q) f \ .$$

3.2. Bias-variance decomposition

We now use a classical bias-variance decomposition of the risk of $\widehat{f}_{\mathrm{AV}}(\lambda,\mu)$ and show that the quantities introduced above allow a simple expression of this risk. For any matrix $M \in \mathcal{S}_p^{++}(\mathbb{R})$, the classical bias-variance decomposition for the linear estimator $\widehat{f}_M = A_M y$ is

$$\begin{aligned} \frac{1}{np}\mathbb{E}\left[\left\|\widehat{f}_M - f\right\|_2^2\right] &= \frac{1}{np}\|(A_M - I_{np})f\|_2^2 + \frac{1}{np}\operatorname{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \\ &= \underbrace{\frac{1}{np}\|(A_M - I_{np})f\|_2^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2}{np}\operatorname{tr}(A_M^\top A_M)}_{\text{Variance}} \ . \end{aligned}$$

We can now compute both bias and variance of the estimator $\widehat{f}_{\mathrm{AV}}(\lambda,\mu)$ by decomposing $A_{M_{\mathrm{AV}}(\lambda,\mu)}$ on the eigenbasis introduced in the previous section.

$np\times$**Variance :**

$$\begin{aligned}
& \sigma^2 \operatorname{tr}(A_M^\top A_M) \\
= \; & \sigma^2 \operatorname{tr}\left((P\otimes Q)^\top \left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta)+npI_{np}\right)^{-1}\right]^2 (P\otimes Q)\right) \\
= \; & \sigma^2 \operatorname{tr}\left(\left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta)+npI_{np}\right)^{-1}\right]^2 \right) \\
= \; & \sigma^2 \sum_{i=1}^n \left[\left(\frac{\gamma_i}{\gamma_i+n\lambda}\right)^2 + (p-1)\left(\frac{\gamma_i}{\gamma_i+n\mu}\right)^2 \right] .
\end{aligned}$$

$np\times$**Bias :**

$$\begin{aligned}
& \|(A_M - I_{np})f\|_2^2 \\
= \; & \|(P\otimes Q)^\top \left[(D^{-1}\otimes K)\left((D^{-1}\otimes K)+npI_{np}\right)^{-1} - I_{np}\right](P\otimes Q)f\|_2^2 \\
= \; & \| \left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta)+npI_{np}\right)^{-1} - I_{np}\right]\nu\|_2^2 \\
= \; & (n\lambda)^2 \sum_{i=1}^n \frac{(\nu_i^1)^2}{(\gamma_i+n\lambda)^2} + (n\mu)^2 \sum_{i=1}^n\sum_{j=2}^p \frac{(\nu_i^j)^2}{(\gamma_i+n\mu)^2} \\
= \; & (n\lambda)^2 \sum_{i=1}^n \frac{(\nu_i^1)^2}{(\gamma_i+n\lambda)^2} + (n\mu)^2 \sum_{i=1}^n \frac{\sum_{j=2}^p (\nu_i^j)^2}{(\gamma_i+n\mu)^2} .
\end{aligned}$$

Thus, the risk of $\widehat{f}_{\mathrm{AV}}(\lambda,\mu)$ becomes

$$\begin{aligned}
n\lambda^2 \sum_{i=1}^n \frac{\frac{(\nu_i^1)^2}{p}}{(\gamma_i+n\lambda)^2} + \frac{\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i+n\lambda}\right)^2 \\
+n\mu^2 \sum_{i=1}^n \frac{\frac{\sum_{j=2}^p (\nu_i^j)^2}{p}}{(\gamma_i+n\mu)^2} + \frac{\sigma^2(p-1)}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i+n\mu}\right)^2 .
\end{aligned} \tag{5}$$

This decomposition has two direct consequences:

- the oracle risk of the multi-task procedure can be obtained by optimizing Equation (5) independently over λ and μ;
- the estimator $\widehat{f}_{\mathrm{AV}}$ can be calibrated by independently calibrating two parameters.

It is now easy to optimize over the quantities in Equation (5). An interesting fact is that both sides have a natural and interesting interpretation, which we give now.

3.3. Remark

To avoid further ambiguities and to simplify the formulas we introduce the following notations for every $i \in \{1, \dots, n\}$:

$$\mu_i = \nu_i^1 = \frac{h_i^1 + \cdots + h_i^p}{\sqrt{p}}$$

and

$$\varsigma_i^2 = \frac{\sum_{j=1}^p (h_i^j)^2}{p} - \left(\frac{\sum_{j=1}^p h_i^j}{p}\right)^2 = \frac{1}{p}\sum_{j=1}^p \left(h_i^j - \frac{\sum_{j=1}^p h_i^j}{p}\right)^2 \ ,$$

so that

$$p\varsigma_i^2 = \sum_{j=2}^p (\nu_i^j)^2 \ .$$

Remark 5. *We can see that for every $i \in \{1, \dots, n\}$, $\mu_i/\sqrt{p}$ is the average of the p target functions f^j, expressed on the basis diagonalizing K. Likewise, ς_i^2 can be seen as the variance between the p target functions f^j (which does not come from the noise).*

Henceforth, the risk of $\widehat{f}_{\mathrm{AV}}(\lambda, \mu)$ over (λ, μ) is decoupled into two parts.

- With the parameter λ, a part which corresponds to the risk of a single-task ridge estimator, which regularizes the mean of the tasks functions, with a noise variance σ^2/p:

$$n\lambda^2 \sum_{i=1}^n \frac{\frac{\mu_i^2}{p}}{(\gamma_i + n\lambda)^2} + \frac{\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\lambda}\right)^2 \ . \tag{6}$$

- With the parameter μ, a part which corresponds to the risk of a single-task ridge estimator, which regularizes the variance of the tasks functions, with a noise variance $(p-1)\sigma^2/p$:

$$n\mu^2 \sum_{i=1}^n \frac{\varsigma_i^2}{(\gamma_i + n\mu)^2} + \frac{(p-1)\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\mu}\right)^2 \ . \tag{7}$$

Remark 6. *Our analysis can also be used on any set of positive semi-definite matrices $\mathcal{M}$ that are jointly diagonalizable on an orthonormal basis, as was $\{M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}_+^2\}$. The element of interest then becomes the norms of the projections of the input tasks on the different eigenspaces (here, the mean and the resulting variance of the p tasks). An example of such a set is when the tasks are known to be split into several clusters, the assignement of each task to its cluster being known to the statistician. The matrices that can be used then regularize the mean of the tasks and, for each cluster, the variance of the tasks belonging to this cluster.*

4. Precise analysis of the multi-task oracle risk

In the latter section we showed that, in order to obtain the multi-task risk, we just had to optimize several functions, which have the form of the risk of a kernel ridge estimator. The risk of those estimators has already been widely studied. Johnstone [20] (see also the article of Caponnetto and De Vito [10] for random design) showed that, for a single-task ridge estimator, if the coefficients of the decomposition of the input function on the eigenbasis of the kernel decrease as $i^{-2\delta}$, with $2\delta > 1$, then the minimax rates for the estimation of this imput function is of order $n^{1/2\delta-1}$. The kernel ridge estimator is then known to be minimax optimal, under certain regularity assumptions (see the work of Bach [5] for more details). If the eigenvalues of the kernel are known to decrease as $i^{-2\beta}$, then a single-task ridge estimator is minimax optimal under the following assumption:

$$1 < 2\delta < 4\beta + 1 \ . \qquad (\mathbf{H_M}(\beta, \delta))$$

The analysis carried in the former section shows that the key elements to express this risk are the components of the average of the signals (μ_i) and the components of the variance of the signals (ς_i) on the basis that diagonalises the kernel matrix K, together with the eigenvalues of this matrix (γ_i). It is then natural to impose the same natural assumptions that make the single-task ridge estimator optimal on those elements. We first suppose that the eigenvalues of the kernel matrix have a polynomial decrease rate:

$$\forall i \in \{1, \dots, n\},\ \gamma_i = n i^{-2\beta} \ . \qquad (\mathbf{H_K}(\beta))$$

Then, we assume that the the components of the average of the signals and the variance of the signals also have a polynomial decrease rate:

$$\forall i \in \{1, \dots, n\},\ \left\{ \begin{array}{rcl} \frac{\mu_i^2}{p} & = & C_1 n i^{-2\delta} \\ \varsigma_i^2 & = & C_2 n i^{-2\delta} \end{array} \right. . \qquad (\mathbf{H_{AV}}(\delta, C_1, C_2))$$

Remark 7. *We assume for simplicity that both Assumptions* $(\mathbf{H_K}(\beta))$ *and* $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ *hold in equality, although the equivalence* $\asymp$ *is only needed.*

Example 1. *This example, related to Assumptions* $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ *and* $(\mathbf{H_K}(\beta))$ *by taking* $\beta = m$ *and* $2\delta = k + 2$*, is detailed by Wahba [30] and by Gu [15]. Let* $\mathcal{P}(2\pi)$ *the set of all square-integrable* 2π*-periodic functions on* $\mathbb{R}$*,* $m \in \mathbb{N}^\star$ *and define* $\mathcal{H} = \left\{ f \in \mathcal{P}(2\pi),\ f^{(m)}_{|[0,2\pi]} \in L^2[0, 2\pi] \right\}$*. This set* $\mathcal{H}$ *has a RKHS structure, with a reproducing kernel having the Fourier base functions as eigenvectors. The* i*-th eigenvalue of this kernel is* i^{-2m}*. For any function* $f \in \mathcal{P}[0, 2\pi] \cap \mathcal{C}^k[0, 2\pi]$*, then its Fourier coefficient are* $O(i^{-k})$*. For instance, if* $f \in \mathcal{P}[0, 2\pi]$ *such that* $\forall x \in [-\pi, \pi],\ f^{(k)}(x) = |x|$*, then its Fourier coefficients are* $\asymp i^{-(k+2)}$*.*

Under Assumptions ($\mathbf{H_K}(\beta)$) and ($\mathbf{H_{AV}}(\delta, C_1, C_2)$), we can now more precisely express the risk of a multi-task estimator. Equation (6) thus becomes

$$\begin{aligned}
& n\lambda^2 \sum_{i=1}^{n} \frac{\frac{\mu_i^2}{p}}{(\gamma_i + n\lambda)^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{\gamma_i}{\gamma_i + n\lambda} \right)^2 \\
= \; & n\lambda^2 \sum_{i=1}^{n} \frac{C_1 n i^{-2\delta}}{(n i^{-2\beta} + n\lambda)^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{n i^{-2\beta}}{n i^{-2\beta} + n\lambda} \right)^2 \\
= \; & C_1 \lambda^2 \sum_{i=1}^{n} \frac{i^{4\beta - 2\delta}}{(1 + \lambda i^{2\beta})^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1 + \lambda i^{2\beta})^2} \\
= \; & R(n, p, \sigma^2, \lambda, \beta, \delta, C_1) \ ,
\end{aligned}$$

while Equation (7) becomes

$$\begin{aligned}
& n\mu^2 \sum_{i=1}^{n} \frac{\varsigma_i^2}{(\gamma_i + n\mu)^2} + \frac{(p-1)\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{\gamma_i}{\gamma_i + n\mu} \right)^2 \\
= \; & n\mu^2 \sum_{i=1}^{n} \frac{C_2 n i^{-2\delta}}{(n i^{-2\beta} + n\mu)^2} + \frac{(p-1)\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{n i^{-2\beta}}{n i^{-2\beta} + n\mu} \right)^2 \\
= \; & C_2 \mu^2 \sum_{i=1}^{n} \frac{i^{4\beta - 2\delta}}{(1 + \mu i^{2\beta})^2} + \frac{(p-1)\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1 + \mu i^{2\beta})^2} \\
= \; & R(n, p, (p-1)\sigma^2, \mu, \beta, \delta, C_2) \ ,
\end{aligned}$$

with

$$R(n, p, \sigma^2, x, \beta, \delta, C) = Cx^2 \sum_{i=1}^{n} \frac{i^{4\beta - 2\delta}}{(1 + x i^{2\beta})^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1 + x i^{2\beta})^2} \ . \tag{8}$$

Remark 8. *It is to be noted that the function R corresponds to the risk of a single-task ridge estimator when the decomposition of the input function on the eigenbasis of K has $i^{-2\delta}$ for coefficients and when $p = 1$. Thus, studying R will allow us to derive both single-task and multi-task oracle rates.*

4.1. Study of the optimum of $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$

We just showed that the function $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$ was suited to derive both single-task and multi-task oracle risk. Bach [5] showed how to obtain a majoration on the function $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$, so that its infimum was showed to match the minimax rates under Assumption ($\mathbf{H_M}(\beta, \delta)$).

In this section, we first propose a slightly more precise upper bound of this risk function. We then show how to obtain a lower bound on this infimum that matches the aforementioned upper bound. This will be done by precisely localizing the parameter minimizing $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$.

Let us first introduce the following notation:

Definition 1.

$$R^\star(n,p,\sigma^2,\beta,\delta,C) = \inf_{\lambda\in\mathbb{R}_+} \left\{R(n,p,\sigma^2,\lambda,\beta,\delta,C)\right\} \ .$$

We now give the upper bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$. For simplicity, we will denote by $\kappa(\beta,\delta)$ a constant, defined in Equation (24), which only depends on β and δ.

Property 1. *Let n and p be positive integers, σ, β and δ positive real numbers such that ($\mathbf{H_M}(\beta,\delta)$), ($\mathbf{H_K}(\beta)$) and ($\mathbf{H_{AV}}(\delta,C_1,C_2)$) hold. Then,*

$$R^\star(n,p,\sigma^2,\beta,\delta,C) \leq \left(2^{1/2\delta}\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} C^{1/2\delta}\kappa(\beta,\delta)\right) \wedge \frac{\sigma^2}{p} \ . \tag{9}$$

Proof. Property 1 is proved in Section C of the appendix. □

In the course of showing Property 1, we obtained an upper bound on the risk function R that holds uniformly on $\mathbb{R}_+$. Obtaining a similar (up to multiplicative constants) lower bound that also holds uniformly on $\mathbb{R}_+$ is unrealistic. However, we will be able to lower bound $R^\star$ by showing that R is minimized by an optimal parameter $\lambda^\star$ that goes to 0 as n goes to $+\infty$.

Property 2. *If Assumption ($\mathbf{H_M}(\beta,\delta)$) holds, the risk $R(n,p,\sigma^2,\cdot,\beta,\delta,C)$ attains its global minimum over $\mathbb{R}_+$ on $[0,\varepsilon\left(\frac{np}{\sigma^2}\right)]$, with*

$$\varepsilon\left(\frac{np}{\sigma^2}\right) = \sqrt{C^{(1/2\delta)-1}2^{1/2\delta}\kappa(\beta,\delta)} \times \frac{1}{\left(\frac{np}{\sigma^2}\right)^{1/2-(1/4\delta)}}\left(1+\eta\left(\frac{np}{\sigma^2}\right)\right) \ ,$$

where $\eta(x)$ goes to 0 as x goes to $+\infty$.

Proof. Property 2 is shown in Section D of the appendix. □

Remark 9. *Thanks to the assumption made on δ, $\frac{1}{2\delta}-1<0$ so that $\left(\frac{np}{\sigma^2}\right)^{\frac{1}{2\delta}-1}$ goes to 0 as $\frac{np}{\sigma^2}$ goes to $+\infty$. This allows us to state that, if the other parameters are constant, $\lambda^\star$ goes to 0 as the quantity $\frac{np}{\sigma^2}$ goes to $+\infty$.*

We can now give a lower bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$. We will give two versions of this lower bound. First, we state a general result.

Property 3. *For every (C,β,δ) such that $1<2\delta<4\beta$ holds, there exits an integer N and a constant $\alpha\in(0,1)$ such that, for every for every (n,p,σ^2) verifying $\frac{np}{\sigma^2}\geq N$, we have*

$$R^\star(n,p,\sigma^2,\beta,\delta,C) \geq \left(\alpha\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} C^{1/2\delta}\kappa(\beta,\delta)\right) \wedge \frac{\sigma^2}{4p} \ . \tag{10}$$

Proof. Property 3 is proved in Section E.3 of the appendix. □

Remark 10. *It is to be noted that N and α only depend on β and δ. We can also remark that α can be taken arbitrarily close to*

$$\frac{\int_0^1 \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2}du} \wedge \frac{\int_0^1 \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du} \quad .$$

Numerical computations show that, by taking $\beta = \delta = 2$, this constant is larger than 0.33*.*

Remark 11. *The assumption made on β and δ is slighlty more restrictive than ($\mathbf{H_M}(\beta,\delta)$), under which the upper bound is shown to hold and under which the single-task estimator is shown to be minimax optimal.*

We are now ensured that R attains its global minimum on $\mathbb{R}_+$, thus we can give the following definition.

Definition 2. For every n, p, σ^2, δ, β and C, under the assumption of Property 2, we introduce

$$\lambda_R^\star \in \operatorname*{argmin}_{\lambda \in \mathbb{R}_+} \left\{ R(n,p,\sigma^2,\lambda,\beta,\delta,C) \right\} \quad .$$

We now give a slightly refined version of Property 3, by discussing whether this optimal parameter $\lambda_R^\star$ is larger or lower than the threshold $n^{-2\beta}$. This allows us to better understand the effect of regularizarion on the oracle risk $R^\star$.

Property 4. *For every (β,δ) such that $4\beta > 2\delta > 1$, integers N_1 and N_2 exist such that*

1. *for every (n,p,σ^2) verifying $\frac{np}{\sigma^2} \geq N_1$ and $n^{1-2\delta} \times \frac{p}{\sigma^2} \leq \frac{1}{N_2}$, then*

$$\lambda_R^\star \geq \frac{1}{n^{2\beta}}$$

and

$$R^\star(n,p,\sigma^2,\beta,\delta,C) \asymp \left(\frac{\sigma^2}{np}\right)^{1-1/2\delta} \quad .$$

2. *for every (n,p,σ^2) verifying $\frac{np}{\sigma^2} \geq N_1$ and $n^{1-2\delta} \times \frac{p}{\sigma^2} \geq N_2$, then*

$$\lambda_R^\star \leq \frac{1}{n^{2\beta}}$$

and

$$R^\star(n,p,\sigma^2,\beta,\delta,C) \asymp R(n,p,\sigma^2,0,\beta,\delta,C) \asymp \frac{\sigma^2}{p} \quad ;$$

Proof. Property 4 is proved in Section E.4 of the appendix. □

Remark 12. *If $p \leq n\sigma^2$ and $\delta > 1$ then we are in the first case, for a large enough n. This is a case where regularization has to be employed in order to obtain optimal convergence rates.*

Remark 13. *If σ^2 and n are fixed and p goes to $+\infty$ then we are in the second case. It is then useless to regularize the risk, since the risk can only be lowered by a factor 4. This also corresponds to a single-task setting where the noise variance σ^2 is very small and where the estimation problem becomes trivial.*

4.2. Multi-task oracle risk

We can now use the upper and lower bounds on $R^\star$ to control the oracle risk of the multi-task estimator. We define

$$\lambda^\star \in \operatorname*{argmin}_{\lambda\in\mathbb{R}_+} \left\{R(n,p,\sigma^2,\lambda,\beta,\delta,C_1)\right\}$$

and

$$\mu^\star \in \operatorname*{argmin}_{\mu\in\mathbb{R}_+} \left\{R(n,p,(p-1)\sigma^2,\mu,\beta,\delta,C_2)\right\} \ .$$

Property 2 ensures that $\lambda^\star$ and $\mu^\star$ exist, even though they are not necessarily unique. The oracle risk then is

$$\mathfrak{R}^\star_{\mathrm{MT}} = \inf_{(\lambda,\mu)\in\mathbb{R}_+^2} \left\{\frac{1}{np}\mathbb{E}\left[\left\|\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)} - f\right\|_2^2\right]\right\} = \frac{1}{np}\mathbb{E}\left[\left\|\widehat{f}_{M_{\mathrm{AV}}(\lambda^\star,\mu^\star)} - f\right\|_2^2\right] \ .$$

We now state the main result of this paper, which simply comes from the analysis of $R^\star$ performed above.

Theorem 1. *For every n, p, C_1, C_2, σ^2, β and δ such that Assumption ($\mathbf{H_M}(\beta,\delta)$) holds, we have*

$$\mathfrak{R}^\star_{\mathrm{MT}} \leq 2^{1/2\delta}\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1}\kappa(\beta,\delta)\left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)}C_2^{1/2\delta}\right] \ . \tag{11}$$

Furthermore, constants N and $\alpha \in (0,1)$ exist such that, if $n \geq N$, $p/\sigma^2 \leq n$ and $2 < 2\delta < 4\beta$, we have

$$\mathfrak{R}^\star_{\mathrm{MT}} \geq \alpha\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1}\kappa(\beta,\delta)\left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)}C_2^{1/2\delta}\right] \ . \tag{12}$$

Proof. The risk of the multi-task estimator $\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)}$ can be written as

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C_1) + R(n,p,(p-1)\sigma^2,\mu,\beta,\delta,C_2) \ .$$

We then apply Properties 1 and 3, since $p/\sigma^2 \leq n$ implies that $p/(p-1)\sigma^2 \leq n$. The assumption $\delta > 1$ ensures that the first setting of Property 4 holds. □

Remark 14. *An interesting fact is that the oracle multi-task risk is of the order $(np/\sigma^2)^{1/2\delta-1}$. This corresponds to the risk of a single-task ridge estimator with sample size np.*

Remark 15. *As noted before, the assumption under which the lower bound holds is slightly stronger than Assumption ($\mathbf{H_M}(\beta,\delta)$).*

5. Single-task oracle risk

In the former section we obtained a precise approximation of the multi-task oracle risk $\mathfrak{R}^\star_{\mathrm{MT}}$. We would now like to obtain a similar approximation for the single-task oracle risk $\mathfrak{R}^\star_{\mathrm{ST}}$. In the light of Section 3, the only element we need to obtain the oracle risk of task $j \in \{1,\dots,p\}$ is the expression of $(h_i^j)_{i=1}^n$, that is, the coordinates of $(f^j(X_i))_{i=1}^n$ on the eigenbasis of K. Unfortunately, Assumption ($\mathbf{H_{AV}}(\delta, C_1, C_2)$) does not correspond to one set of task functions $(f^1,\dots,f^p)$. Thus, since several single-task settings can lead to the same multi-task oracle risk, we now explicitly define two repartitions of the task functions $(f^1,\dots,f^p)$, for which the single-task oracle risk will be computed.

- "2 points": suppose, for simplicity, that p is even and that

$$f^1 = \cdots = f^{p/2} \quad \text{and} \quad f^{p/2+1} = \cdots = f^p \ . \tag{2Points}$$

- "1 outlier":

$$f^1 = \cdots = f^{p-1} \ . \tag{1Out}$$

Both assumptions correspond to settings in which the multi-task procedure would legitimately be used. Assumption (2Points) models the fact that all the functions lie in a cluster of small radius. It supposes that the functions are split into two groups of equal size, in order to be able to explicitly derive the single-task oracle risk. Assumption (1Out) supposes that all the functions are grouped in one cluster, with one outlier. In order to make the calculations possible, all the functions in one group are assumed to be equal. Since this is not a fully convincing situation to study the behaviour of the multi-task oracle, simulation experiments were also run on less restrictive settings. The results of those experiments are shown in Section 8.

Remark 16. *The hypotheses* (2Points) *and* (1Out) *made on the functions* f^j *can be expressed on* (h_i^j). *Assumption* (2Points) *becomes*

$$\forall i \in \{1,\dots,n\},\ h_i^1 = \cdots = h_i^{p/2} \quad \textit{and} \quad h_i^{p/2+1} = \cdots = h_i^p \ ,$$

while Assumption (1Out) *becomes*

$$\forall i \in \{1,\dots,n\},\ h_i^1 = \cdots = h_i^{p-1} \ .$$

Under those hypotheses we now want to derive an expression of $(h_i^1,\dots,h_i^p)$ given (μ_i, ς_i) so that we can exactly compute the single-task oracle risk. Remember we defined for every $i \in \{1,\dots,n\}$,

$$\mu_i = \frac{1}{\sqrt{p}} \sum_{j=1}^p h_i^j$$

and

$$\varsigma_i^2 = \frac{1}{p}\sum_{j=1}^p \left(h_i^j\right)^2 - \frac{\mu_i^2}{p} = \frac{1}{p}\sum_{j=1}^p \left(h_i^j - \frac{\mu_i}{\sqrt{p}}\right)^2 \ .$$

We also re-introduce the single-task oracle risk:

$$\mathfrak{R}^\star_{\mathrm{ST}} = \inf_{(\lambda^1,\dots,\lambda^p)\in\mathbb{R}_+^p} \left\{ \frac{1}{np}\sum_{j=1}^p \left\| \widehat{f}^j_{\lambda^j} - f^j \right\|_2^2 \right\} \ .$$

We now want to closely study this single-task oracle risk, in both settings.

5.1. *Analysis of the oracle single-task risk for the "2 points" case* (2Points)

In this section we write the single-task oracle risk when Assumption (2Points) holds. As shown in Lemma 8, the risk of the estimator $\widehat{f}^j_\lambda = A_\lambda y^j$ for the jth task, which we denote by $R^j(\lambda)$, verifies

$$R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}-\sqrt{C_2}\right)^2) \leq R^j(\lambda) \leq R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}+\sqrt{C_2}\right)^2) \ .$$

Both upper and lower parts eventually behave similarly. In order to simplify notations and to avoid having to constantly write two risks, we will assume that half of the tasks have a risk equal to the right-hand side of the later inequality and the other half a risk equal to the left-hand side of this inequality. This leads to the following assumption:

$$\forall i \in \{1,\dots,n\}, \ \left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta}(\sqrt{C_1}+\sqrt{C_2}) \\ h_i^p & = & \sqrt{n} i^{-\delta}(\sqrt{C_1}-\sqrt{C_2}) \end{array} \right. \ . \qquad (\mathbf{H_{2Points}})$$

This minor change does not affect the convergence rates of the estimator. Consequently, if $1 \leq j \leq p/2$ the risk for task j is $R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}+\sqrt{C_2}\right)^2)$ so that the oracle risk for task j is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \times \left(\sqrt{C_1}+\sqrt{C_2}\right)^{1/\delta} \ ,$$

and if $p/2+1 \leq j \leq p$ the risk for task j is $R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}-\sqrt{C_2}\right)^2)$ so that the oracle risk for task j is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \times \left|\sqrt{C_1}-\sqrt{C_2}\right|^{1/\delta} \ ,$$

Remark 17. *We can remark that* ($\mathbf{H_{2Points}}$) *implies* (2Points) *and that* ($\mathbf{H_{2Points}}$) *implies* ($\mathbf{H_{AV}}(\delta,C_1,C_2)$)*, as shown in Lemma 10. Consequently, if* ($\mathbf{H_{2Points}}$) *holds, we have, for every* $i \in \{1,\dots,n\}$, $h_i^1 = \frac{\mu_i}{\sqrt{p}} + \varsigma_i$ *and* $h_i^p = \frac{\mu_i}{\sqrt{p}} - \varsigma_i$.

Corollary 1. *For every* n, p, C_1, C_2, σ^2, β *and* δ *such that* $2 < 2\delta < 4\beta$ *and* $n\sigma^2 > 1$ *and that Assumptions* ($\mathbf{H_{2Points}}$) *and* ($\mathbf{H_K}(\beta)$) *hold, then*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} \frac{\kappa(\beta,\delta)}{2} \times p^{1-1/2\delta} \left[\left(\sqrt{C_1}+\sqrt{C_2}\right)^{1/\delta} + \left|\sqrt{C_1}-\sqrt{C_2}\right|^{1/\delta} \right] \ . \qquad (13)$$

5.2. Analysis of the oracle single-task risk for the "1 outlier" case (1Out)

In this section we suppose that Assumption (1Out) holds. As shown in Lemma 9, we can lower and upper bound the risks of the single-tasks estimators by functions of the shape $R(n,p,\sigma^2,\lambda,\beta,\delta,C)$. As in the latter section, to avoid the burden of writing two long risk terms at every step, and since all those risks have the same convergence rates, we suppose from now on the new assumption:

$$\forall i \in \{1,\dots,n\} \left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1} + \frac{1}{\sqrt{p-1}} \sqrt{C_2} \right) \\ h_i^p & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1} - \sqrt{p-1} \sqrt{C_2} \right) \end{array} \right. . \qquad (\mathbf{H_{1Out}})$$

This minor change does not affect the convergence rates of the estimator. Consequently, if $1 \leq j \leq p-1$ the risk for task j is $R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^2)$ so that the oracle risk for task j is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \times \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}} \right)^{1/\delta} ,$$

while the risk for task p is $R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} - \sqrt{(p-1)C_2}\right)^2)$ so that the oracle risk for task p is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \times \left| \sqrt{C_1} - \sqrt{(p-1)C_2} \right|^{1/\delta} .$$

Remark 18. *We can also remark here that* ($\mathbf{H_{1Out}}$) *implies* (1Out) *and that* ($\mathbf{H_{1Out}}$) *implies* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$)*, as shown in Lemma 9. Consequently, if* ($\mathbf{H_{1Out}}$) *holds, we have, for every* $i \in \{1,\dots,n\}$, $h_i^1 = \frac{\mu_i}{\sqrt{p}} + \frac{1}{\sqrt{p-1}}\varsigma_i$ *and* $h_i^p = \frac{\mu_i}{\sqrt{p}} - \sqrt{p-1}\varsigma_i$.

oracle

Corollary 2. *For every* n, p, C_1, C_2, σ^2, β *and* δ *such that* $2 < 2\delta < 4\beta$ *and* $n\sigma^2 > 1$ *and that Assumptions* ($\mathbf{H_{1Out}}$) *and* ($\mathbf{H_K}(\beta)$) *hold, then*

$$\begin{aligned} \mathfrak{R}^\star_{\mathrm{ST}} &\asymp \left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \\ &\quad \times p^{1-1/2\delta} \left[\frac{p-1}{p} \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}} \right)^{1/\delta} + \frac{1}{p} \left| \sqrt{C_1} - \sqrt{(p-1)C_2} \right|^{1/\delta} \right] . \end{aligned} \qquad (14)$$

6. Comparison of multi-task and single-task

In the two latter section we obtained precise approximations of the multi-task oracle risk, $\mathfrak{R}^\star_{\mathrm{MT}}$, and of the single-task oracle risk, $\mathfrak{R}^\star_{\mathrm{ST}}$, under either Assumption ($\mathbf{H_{2Points}}$) or ($\mathbf{H_{1Out}}$). We can now compare both risks in either setting,

by studying their ratio

$$\rho = \frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} \ .$$

We will express the quantity ρ as a factor of

$$r = \frac{C_2}{C_1} \ .$$

The parameter r controls the amount of the signal which is contained in the mean of the functions. When r is small, the mean of the tasks contains much more signal than the variance of the tasks, so that the tasks should be "similar". This is a case where the multi-task oracle is expected to perform better than the single-task oracle. On the contrary, when r is large, the variance of the tasks is more important than the mean of the tasks. This is a case where the tasks would be described as "non-similar". It is then harder to conjecture whether the single-task oracle performs better than the multi-task oracle and, as we will see later, the answer to this greatly depends on the setting.

6.1. *Analysis of the oracle multi-task improvement for the "2 points" case* (2Points)

We now express ρ as a function of r when the tasks are split in two groups.

Corollary 3. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > p$ and that Assumptions ($\mathbf{H_{2Points}}$) and ($\mathbf{H_K}(\beta)$) hold, then*

$$\rho \asymp \frac{p^{1/2\delta-1} + \left(\frac{p-1}{p}\right)^{1-(1/2\delta)} r^{1/2\delta}}{\left(1+\sqrt{r}\right)^{1/\delta} + \left|1-\sqrt{r}\right|^{1/\delta}} \ . \tag{15}$$

Remark 19. *The right-hand side of Equation (15) is always smaller than $\frac{1}{2}$. Thus, under the assumptions of Corollary 3, the multi-task oracle risk can never be arbitrarily worse than the single-task oracle risk.*

We can first see that, under the assumptions of Corollary 3, $\rho = \Theta\left(p^{1/2\delta-1}\right)$ as r goes to 0. This is the same improvement that we get we multiplying the sample-size by p. We also have $\rho = \Theta\left(\left(\frac{p-1}{p}\right)^{1-(1/2\delta)}\right)$ as r goes to $+\infty$, so that the multi-task oracle and the single-task oracle behave similarly. Finally, $\rho = \Theta\left(\frac{r^{1/2\delta}}{\left(1+\sqrt{r}\right)^{1/\delta}+\left|1-\sqrt{r}\right|^{1/\delta}}\right)$ as p goes to $+\infty$, so that the behaviours we just discussed are still valid with a large number of tasks.

6.2. *Analysis of the oracle multi-task improvement for the "1 outlier" case* (1Out)

We now express ρ as a function of r when the tasks are grouped in one group, with one outlier.

Corollary 4. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > p$ and that Assumptions ($\mathbf{H_{1Out}}$) and ($\mathbf{H_K}(\beta)$) hold, then*

$$\rho \asymp \frac{p^{1/2\delta-1} + \left(\frac{p-1}{p}\right)^{1-(1/2\delta)} r^{1/2\delta}}{\frac{p-1}{p}\left(1+\sqrt{\frac{r}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|1-\sqrt{r(p-1)}\right|^{1/\delta}} \ . \tag{16}$$

We can see that, under the assumptions of Corollary 4, $\rho = \Theta\left(p^{1/2\delta-1}\right)$ as r goes to 0. As in the latter section, this is the same improvement that we get we multiplying the sample-size by p. However, $\rho = \Theta\left(\left(\frac{p-1}{p}\right)^{1-1/2\delta} \times \frac{p(p-1)^{-1/2\delta}}{1+(p-1)^{1-1/\delta}}\right)$ as r goes to $+\infty$. This quantity goes to $+\infty$ as $p \longrightarrow +\infty$, so that the multi-task oracle performs arbitrarily worse than the single-task one in this asymptotic setting. Finally, $\rho = \Theta\left(r^{1/2\delta}\right)$ as p goes to $+\infty$. This quantity goes to $+\infty$ as r goes to $+\infty$, so that the behaviours we just mentioned stay valid with a large number of tasks.

6.3. Discussion

When r is small, either under Assumption (2Points) or (1Out), the mean of the signal is much stronger than the variance. Thus, the multi-task procedure performs better than the single-task one.

Example 2. *If $r = 0$, then all the tasks are equal. The improvement of the multi-task procedure over the single-task one then is $p^{1/2\delta-1}$. This was expected: it corresponds to the risk of a ridge regression with a np-sample.*

As r goes to 0, the multi-task oracle outperforms its single-task counterpart by a factor $p^{1/2\delta-1}$. When p is large (but, remember, this only holds when $p/\sigma^2 \leq n$, so n also has to be large), this leads to a substantial improvement. It is easily seen that, for any constant $C > 1$, if $r \leq (C-1)^{2\delta}(p-1)^{1-2\delta}$, then the right-hand side of Equation (15) becomes smaller than $Cp^{1/2\delta-1}$. Thus, if the tasks are similar enough, the multi-task oracle performs as well as the oracle for a np-sample, up to a constant.

On the contrary, when r is large, the variance carries most of the signal, so that the tasks differ one from another. As r goes to $+\infty$, the two settings have different behaviours:

- under Assumption (2Points) (that is, when we are faced to two equally-sized groups), the oracle risks of the multi-task and of the single-task estimators are of the same order: they can only differ by a multiplicative constant;
- under Assumption (1Out) (that is, when we are faced to one cluster and one outlier), the single-task oracle outperforms the multi-task one, by a factor which is approximatly $p^{1/\delta}$.

Finally, Assumption (2Points) presents no drawback for the multi-task oracle, since under those hypotheses its performance cannot be worse than the single-task oracle's one. On the contrary, Assumption (1Out) presents a case where the use of a multi-task technique greatly increases the oracle risk, when the variance between the tasks is important, while it gives an advantage to the multi-task oracle when this variance is small. The location where the multi-task improvement stops corresponds to the barrier $\rho = 1$. Studying this object seems difficult, since we only know ρ up to a multiplicative constant. Also, finding the contour lines of the righ-hand side of Equation (16) does not seem to be an easy task. In Section 8, we will run simulations in situations where the oracle risk can no longer be explicitly derived. We will show that the behaviours found in these two examples still appear in the simulated examples.

7. Risk of a multi-task estimator

Solnon et al. [27] introduced an entirely data-driven estimator to calibrate $M_{\mathrm{AV}}(\lambda,\mu)$ over $\mathbb{R}_+^2$. One of their main results is an oracle inequality, that compares the risk of this estimator to the oracle risk. Thus, $\mathfrak{R}^\star_{\mathrm{MT}}$ is attainable by a fully data-driven estimator. We now show that our estimation of the multi-task oracle risk is precise enough so that we can use it in the mentionned oracle inequality and still have a lower risk than the single-task oracle one.

The following assumption will be used, with $\mathrm{df}(\lambda) = \mathrm{tr}(A_\lambda)$ and $A_\lambda = K(K + n\lambda I_n)^{-1}$:

$$\left.\begin{array}{l} \forall j \in \{1,\dots,p\}\,,\ \exists \lambda_{0,j} \in (0,+\infty)\,, \\ \qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{and} \quad \dfrac{1}{n}\left\|(A_{\lambda_{0,j}} - I_n) f^j\right\|_2^2 \leq \sigma^2 \sqrt{\dfrac{\ln n}{n}} \end{array}\right\} \tag{Hdf}$$

We will also denote $\mathcal{M} = \left\{M_{\mathrm{AV}}(\lambda,\mu),\ (\lambda,\mu) \in \mathbb{R}_+^2\right\}$ and $\widehat{M}_{\mathrm{HM}}$ the estimator introduced in Solnon et al. [27], which belongs to $\mathcal{M}$. Theorem 29 of Solnon et al. [27] thus states:

Theorem 2. *Let $\alpha = 2$, $\theta \geq 2$, $p \in \mathbb{N}^\star$ and assume (Hdf) holds true. An absolute constant $L > 0$ and a constant $n_1(\theta)$ exist such that the following holds as soon as $n \geq n_1(\theta)$.*

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}_{HM}} - f\right\|_2^2\right] \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \mathbb{E}\left[\inf_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2\right\}\right] \\ + L\sigma^2(2+\theta)^2 p\frac{\ln(n)^3}{n} + \frac{p}{n^{\theta/2}}\frac{\|f\|_2^2}{np}\ . \end{aligned} \tag{17}$$

We first remark that

$$\mathbb{E}\left[\inf_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2\right\}\right] \leq \mathfrak{R}^\star_{\mathrm{MT}}\ .$$

We can now plug the oracle risk in the oracle inequality (17). Then, if we suppose that, for $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$, $(h_i^j)^2 = nC^j i^{-2\delta}$, we have that

$$\|f\|_2^2 = \sum_{j=1}^{p}\sum_{i=1}^{n}(h_i^j)^2 = n\sum_{j=1}^{p} C^j \sum_{i=1}^{n} i^{-2\delta} \leq n\zeta(2\delta)\sum_{j=1}^{p} C^j \quad .$$

Remark 20. *Assumption* (2Points) *means that for every* $i \in \{1,\dots,n\}$, *if* $1 \leq j \leq p/2$,

$$C^j = \left(\sqrt{C_1} + \sqrt{C_2}\right)^2$$

and if $p/2+1 \leq j \leq p$,

$$C^j = \left(\sqrt{C_1} - \sqrt{C_2}\right)^2 \quad .$$

Assumption (1Out) *means that for every* $i \in \{1,\dots,n\}$, *if* $1 \leq j \leq p-1$,

$$C^j = \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^2$$

while

$$C^p = \left(\sqrt{C_1} - \sqrt{(p-1)C_2}\right)^2 \quad .$$

Property 5. *Under Assumptions* ($\mathbf{H_K}(\beta)$) *and* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$) *with* $2\delta > 2$, *there exists a constant* N_1 *such that for every* $n \geq N_1$, *Assumption* (Hdf) *holds.*

Proof. We can see that Assumption (Hdf) is made independently on every task. Thus we can suppose that $p = 1$. Let us denote $b(\lambda) = n^{-1}\|(A_\lambda - I_n)f\|_2^2$. We can see that if there exists constants $c > 0$ and $d > 1$ such that for every $\lambda \in \mathbb{R}_+$ $b(\lambda) \leq c\sigma^2 \operatorname{df}(\lambda)^{-d}$, then Assumption (Hdf) holds for n large enough. Indeed, let $\lambda \in \mathbb{R}_+$ such that $\operatorname{df}(\lambda) \leq \sqrt{n}$. Then, if $b(\lambda) \leq c\sigma^2 \operatorname{df}(\lambda)^{-d}$, $b(\lambda) \leq \sigma^2 c(\sqrt{n})^{-d} \leq \sigma^2 c\frac{n^{(-d+1)/2}}{\sqrt{n}}$. It just suffices to see that, for n large enough, $cn^{-d+1} \leq \ln(n)$.

Using Lemmas 6 and 5 we can see that, for every $\lambda \in \mathbb{R}_+$,

$$b(\lambda) \leq \frac{\lambda^{\frac{2\delta-1}{2\beta}}}{\beta} I_1(\beta,\delta)$$

and, for n large enough, there exists a constant α such that, for every $\lambda \in \mathbb{R}_+$,

$$\operatorname{df}(\lambda) = \operatorname{tr} A_\lambda \geq \alpha\frac{\lambda^{\frac{-1}{2\beta}}}{2\beta} I_2(\beta)$$

Thus, for n large enough, there exists a constant c (depending on σ^2, β and δ) such that, for every $\lambda \in \mathbb{R}_+$,

$$b(\lambda) \leq c\sigma^2 \operatorname{tr}(A_\lambda)^{-(2\delta-1)} \quad .$$

Hence, if $2\delta > 2$, there exists a constant N_1 such that for every $n \geq N_1$, Assumption (Hdf) holds. □

Thus, we can apply Theorem 2 to the estimator $\widehat{f}_{\widehat{M}_{\mathrm{HM}}}$ under either Assumption (2Points) or (1Out) (and we denote by ρ either $\rho_{2Points}$ or ρ_{1Out}).

Property 6. *For every positive numbers $(\beta, \delta, \theta, C_1, C_2)$ verifying $4\beta > 2\delta > 2$ and $\theta > 1$, there exists positive constants $(N(\beta,\delta,\theta), L)$ such that, for every (n, p, σ^2) verifying $n \geq N$ and $\frac{p}{\sigma^2} \leq n$, if Assumption ($\mathbf{H_K}(\beta)$) and if either Assumption ($\mathbf{H_{2Points}}$) or Assumption ($\mathbf{H_{1Out}}$) hold, the ratio between the risk of the estimator $\widehat{f}_{\widehat{M}_{HM}}$ and the single-task oracle risk verifies*

$$\frac{\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}_{HM}} - f\right\|_2^2\right]}{\mathfrak{R}^{\star}_{\mathrm{ST}}} \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \rho + Cst \times \frac{L\sigma^2(2+\theta)^2 p\frac{\ln(n)^3}{n} + \frac{p\zeta(2\delta)}{n^{\theta/2}}\frac{1}{p}\sum_{j=1}^{p} C^j}{\left(\frac{n}{\sigma^2}\right)^{1/2\delta-1}\kappa(\beta,\delta) \times \frac{1}{p}\sum_{j=1}^{p}(C^j)^{1/2\delta}} \ .$$

Proof. This is a straightforward application of the preceding results. □

We now show that the latter fully data-driven multi-task ridge estimator achieves a lower risk than the single-task ridge oracle, in both settings (2Points) and (1Out).

Corollary 5. *For every positive numbers $(\beta, \delta, \theta, \sigma^2, \varepsilon)$ verifying $4\beta > 2\delta > 2$ and $\theta > 2$, there exists positive constants (N, r) such that, for every (n, p, C_1, C_2) verifying $n \geq N$, $\frac{p}{\sigma^2} \leq n^{1/4\delta}$ and $\frac{C_2}{C_1} \leq r$, if Assumptions ($\mathbf{H_K}(\beta)$) holds and if either Assumption ($\mathbf{H_{2Points}}$) or Assumption ($\mathbf{H_{1Out}}$) hold, the ratio between the risk of the estimator $\widehat{f}_{\widehat{M}_{HM}}$ and the single-task oracle risk verifies*

$$\frac{\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}_{HM}} - f\right\|_2^2\right]}{\mathfrak{R}^{\star}_{\mathrm{ST}}} < \varepsilon \ .$$

Proof. First, we can see that under either Assumption (2Points) or Assumption (1Out), both $\frac{1}{p}\sum_{j=1}^{p} C^j$ and $\frac{1}{p}\sum_{j=1}^{p}(C^j)^{1/2\delta}$ converge, as p goes to $+\infty$, to quantities only depending on C_1, C_2 and δ and are thus bounded with respect to p. Then, as it was shown in the previous section, both $\rho_{2Points}$ and ρ_{1Out} go to 0 as $\frac{C_1}{C_2}$ goes to 0. Finally, we can see that $\frac{p}{\sigma^2} \leq n^{1/4\delta}$ implies that $\frac{p}{\sigma^2} \leq n$ and that

$$\frac{\sigma^2 p\frac{\ln(n)^3}{n}}{\left(\frac{n}{\sigma^2}\right)^{1/2\delta-1}} = \left(\sigma^2\right)^{1/2\delta} \times p \times \frac{\ln(n)^3}{n^{1/2\delta}} \leq \left(\sigma^2\right)^{1+1/2\delta} \times \frac{\ln(n)^3}{n^{1/4\delta}} \underset{n\to+\infty}{\longrightarrow} 0$$

together with

$$\frac{\frac{p}{n^{\theta/2}}}{\left(\frac{n}{\sigma^2}\right)^{1/2\delta-1}} \leq \left(\sigma^2\right)^{1/2\delta} \times n^{1-\theta/2-1/4\delta} \underset{n\to+\infty}{\longrightarrow} 0 \ .$$

□

Remark 21. *The result shown in Corollary (5) establishes that a fully data-driven multi-task estimator outperforms an oracle single-task estimator, which is minimax optimal .*

8. Numerical experiments

The hypotheses we used in the former sections, although sufficient to precisely derive the risk of the estimator, do not reflect realistic situations. In this section we study less restrictive settings. However, we are no longer able to obtain simple formulas for the oracle risk as we did before. Thus, we resort to numerical simulations to illustrate the behaviour of both single-task and multi-task oracles.

8.1. Setting A: relaxation of Assumptions ($\mathbf{H_{AV}(\delta, C_1, C_2)}$) and (2Points) in order to get one general group of tasks

In the latter sections we modeled the fact that the p target functions are close. However, due to technical constraints we were only able to deal with cases where the functions are split into two groups and are then equal inside each group, thus introducing Assumptions (2Points) and (1Out). We propose here to extend this setting by simulating a more general group of tasks. Those tasks should all be at a comparable distance from a centroid function.

We suppose that $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$ is a sequence of i.i.d. random variables, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution (that is, such that $\mathbb{P}(\varepsilon_i^j = 1) = \mathbb{P}(\varepsilon_i^j = -1) = 1/2$). The target functions are then defined by

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\},\ h_i^j = \sqrt{n} i^{-\delta}\left(\sqrt{C_1} + \varepsilon_i^j \sqrt{C_2}\right) \quad . \tag{18}$$

Thus, all the p target functions are "close" to a centroid function, whose coordinates on the eigenvectors of the kernel matrix are $\sqrt{n} i^{-\delta}\sqrt{C_1}$, with a "dispersion factor" $\sqrt{C_2}$. In this setting, we can easily express the key elements for the analysis of this risk :

$$\frac{\mu_i^2}{p} = n i^{-2\delta}\left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p}\sqrt{C_2}\right)^2$$

and

$$\varsigma_i^2 = n i^{-2\delta}\left(\frac{1}{p}\sum_{j=1}^p \left(\sqrt{C_1} + \varepsilon_i^j\sqrt{C_2}\right)^2 - \left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p}\sqrt{C_2}\right)^2\right) \quad .$$

Remark 22. *The theoretical analysis developed previously cannot be applied here, due to the presence of random terms, which depend on i, in front of the decay term $n i^{-2\delta}$.*

8.2. Setting B: random drawing of the input points and functions

Assumptions ($\mathbf{H_K}(\beta)$) and ($\mathbf{H_{AV}}(\delta, C_1, C_2)$) model the behaviour of the spectral elements of f and K as if they exactly follow the spectral elements of the kernel operator and the input function. Although convenient for the analysis, this setting is unlikely to hold in practice and we propose here to draw the input points $(X_i)_{i=1}^n$ and compute the risk using the eigenvalues of the kernel matrix.

We suppose here that $(X_i)_{i=1}^n$ is a sequence of i.i.d. random variables uniformly drawn on $[-\pi, \pi]$. As in the latter section, we also suppose that we have an i.i.d. sequence of random variables $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution. The target functions are then defined by

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\},\ f^j(X_i) = \left(\sqrt{C_1} + \varepsilon_i^j \sqrt{C_2}\right) |X_i| \quad . \tag{19}$$

As stated in Wahba [30] and in Gu [15], a natural kernel to use is, with $m \in \mathbb{N}^\star$,

$$R(x,y) = 2\sum_{i=1}^{+\infty} \frac{\cos\left(i(x-y)\right)}{i^{2m}} \quad .$$

In this setting, the coefficients of the decomposition of $f : x \mapsto |x|$ on the Fourier basis are known to be asymptotically equivalent to i^{-2}. Thus, this setting is a natural extension of Assumptions ($\mathbf{H_K}(\beta)$) and ($\mathbf{H_{AV}}(\delta, C_1, C_2)$), with $\beta = m$—since the eigenvalues of the kernel R are i^{-2m}—and $\delta = 2$.

8.3. Setting C: further relaxation of Assumptions ($\mathbf{H_{AV}}(\boldsymbol{\delta, C_1, C_2})$) and (2Points) in one group of tasks

We consider the same tasks than in Setting A, but also allow the regularity of the variance to vary. This gives the following model, supposing that $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$ is a sequence of i.i.d. random variables, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution:

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\},\ h_i^j = \sqrt{n}\left(\sqrt{C_1} i^{-\delta_1} + \varepsilon_i^j \sqrt{C_2} i^{-\delta_2}\right) \quad .$$

We allow the variance to have a varying regularity and intensity by changing C_2 and δ_2. This gives us the following quantities of interest: for every $i \in \{1,\dots,n\}$,

$$\frac{\mu_i^2}{p} = n i^{-2\delta_1} \left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p} \sqrt{C_2} i^{-(\delta_2-\delta_1)}\right)^2$$

and

$$\varsigma_i^2 = n i^{-2\delta_1} \left(\frac{1}{p}\sum_{j=1}^p \left(\sqrt{C_1} + \varepsilon_i^j \sqrt{C_2} i^{-(\delta_2-\delta_1)}\right)^2 - \left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p} \sqrt{C_2} i^{-(\delta_2-\delta_1)}\right)^2\right) \quad .$$

8.4. Setting D: relaxation of Assumptions (1Out) and ($\mathbf{H_{AV}(\delta, C_1, C_2)}$)

Assumption (1Out) states that we have one of $p-1$ identical tasks and one outlier. We now simulate a slightly more general setting by having one cluster of $p-1$ around 0 and an outlier. This gives the following model, supposing that $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$ is a sequence of i.i.d. random variables, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution:

$$\forall i \in \{1,\dots,n\}\,,\ \forall j \in \{1,\dots,p-1\}\,,\ h_i^j = \sqrt{n}\varepsilon_i^j i^{-2}$$

and

$$\forall i \in \{1,\dots,n\}\,,\quad h_i^p = \sqrt{nC_2}\varepsilon_i^p i^{-\delta_2}\ .$$

We allow the outlier to have a varying regularity and intensity by changing C_2 and δ_2. This gives us the following quantities of interest: for every $i \in \{1,\dots,n\}$,

$$\frac{\mu_i^2}{p} = ni^{-\delta_1}\left(\frac{\sqrt{C_1}}{p}\sum_{j=1}^{p-1}\varepsilon_i^j + \frac{\varepsilon_i^p}{p}\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2$$

and

$$\varsigma_i^2 = ni^{-\delta_1}\left(\frac{p-1}{p}C_1 + \frac{1}{p}C_2 i^{-2(\delta_2-\delta_1)} - \left(\frac{1}{p}\sum_{j=1}^{p-1}\varepsilon_i^j + \frac{\varepsilon_i^p}{p}\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2\right)\ .$$

8.5. Methodology

In every setting, we computed the oracle risks of both the multi-task estimator and the single-task one. As shown before, for instance in Equation (5), both the multi-task risk (which has two hyper-parameters, λ and μ) and the single-task risk (which has p hyper-parameters, λ^1 to λ^p) can be decomposed as a sum of several functions, each depending on a unique hyper-parameter. We used Newton's method to optimize each of those $p+2$ functions over, respectively, λ, μ, $\lambda^1,\dots,\lambda^p$. Our stopping criterion was that the derivative of the function being optimized was inferior to 10^{-5}, in absolute value. We replicated each experiment $N = 100$ times. This gives us N independent realisations of $(\mathfrak{R}^\star_{\mathrm{MT}}, \mathfrak{R}^\star_{\mathrm{ST}})$, the randomness coming from the repartition of the tasks and, in Setting B, from the drawing of the input points $(X_i)_{i=1}^n$.

In Settings A and B, we first test the hypothesis $\mathbb{H}_0 = \{\mathbb{P}\left(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}\right) < 0.5\}$ against $\mathbb{H}_1 = \{\mathbb{P}\left(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}\right) \geq 0.5\}$. This amounts to testing whether the median of $\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}$ is larger than one. For every iteration $i \in \{1,\dots,N\}$, we observe $B_i = \mathbf{1}_{\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}}$. Since the random variables $(B_i)_{i\in\{1,\dots,N\}}$ follow a Bernouilli distribution of parameter $\mathbb{P}\left(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}\right)$, we can apply Hoeffding's inequality

[24] and see that, for every $\varepsilon > 0$, $[\bar{B}_N - \varepsilon, 1]$ is a confidence interval of level $1 - e^{-2N\varepsilon^2}$ for $\mathbb{P}\left(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}\right)$. This leads to the following p-value:

$$\pi_1 = \begin{cases} e^{-2N\left(\bar{B}_N - 0.5\right)^2} & \text{if } \bar{B}_N \geq 0.5 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

In those two settings, we also test the hypothesis $\mathbb{H}_0 = \left\{\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] > 1\right\}$ against $\mathbb{H}_1 = \left\{\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] \leq 1\right\}$. Let us denote by $\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$ the empirical mean of the random variables $\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}$, $\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$ the resulting standard deviation and Φ the cumulative distribution function of a standard gaussian distribution. Then, a classical use of the central limit theorem and of Slutsky's Lemma gives that

$$\left[0, \widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] + \frac{\varepsilon}{\sqrt{n}}\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]\right]$$

is an asymptotic confidence interval of level $\Phi(\varepsilon)$ for $\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$. This leads to the following asymptotic p-value:

$$\pi_2 = \Phi\left[\sqrt{n}\left(\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] - 1\right)\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]^{-1}\right] \ .$$

The results of those tests are shown in Table 1 for Setting A and in Table 2 for Setting B.

In Settings C and D, we use the same asymptotic framework and show error bars corresponding to the asymptotic confidence interval

$$\left[\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] - \frac{z_{0.975}}{\sqrt{n}}\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right], \widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] + \frac{z_{0.975}}{\sqrt{n}}\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]\right]$$

of level 95%, where z_α denotes the quantile of order α of the standard gaussian distribution. The results of those simulations are shown in Figure 1 for Setting C and in Figure 2 for Setting D.

We used the following values for the parameters: $n = 50$, $p = 5$, $\sigma^2 = 1$ and $C_1 = 1$. We finally settled $\delta = 2$ in Settings A and B and $\delta_1 = 2$ in Settings C and D.

8.6. Interpretation

When all the tasks are grouped in one cluster (Settings A, B and C), the same phenomenon as under Assumption (2Points) appears. In situations where the mean component of the signal has more weight than the variance component (in Settings A and B, that is when r is small, in Setting C, this occurs when

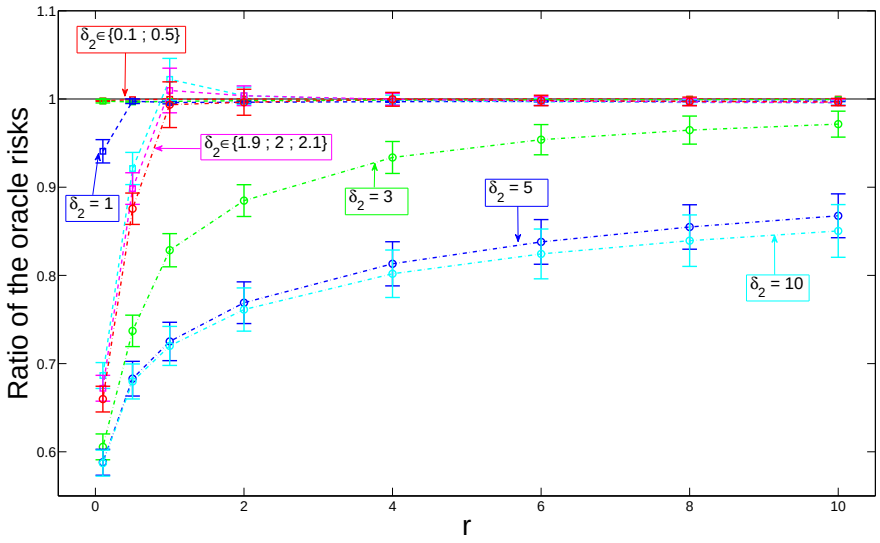

FIG 1. *Further relaxation of Assumption* (2Points) *(Experiment C), improvement of multi-task compared to single-task:* $\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$. *Best seen in colour.*

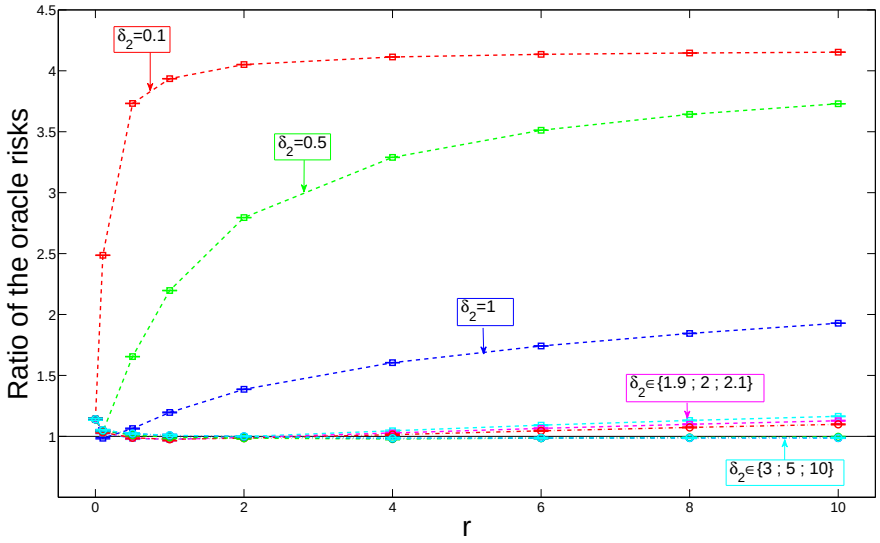

FIG 2. *Relaxation of Assumption* (1Out) *(Experiment D), improvement of multi-task compared to single-task:* $\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$. *Best seen in colour.*

C_2	$r = \frac{C_2}{C_1}$	β	$\bar{B}_{100}$	π_1	$\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	$\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	π_2
0.01	0.01	2	1	$< 10^{-15}$	0.434	0.0324	$< 10^{-15}$
and 0.1	0.1	2	1	$< 10^{-15}$	0.672	0.0747	$< 10^{-15}$
0.5	0.5	2	0.94	$< 10^{-15}$	0.898	0.0913	$< 10^{-15}$
1	1	2	0.51	9.80×10^{-1}	1.01	0.129	0.773
5	5	2	0.38	1	0.998	0.0292	0.302
10	10	2	0.42	1	0.996	0.0172	9.90×10^{-3}
100	100	2	0.76	1.35×10^{-6}	0.997	5.44×10^{-3}	5.97×10^{-10}
0.01	0.01	4	1	$< 10^{-15}$	0.426	0.0310	$< 10^{-15}$
0.1	0.1	4	1	$< 10^{-15}$	0.703	0.0737	$< 10^{-15}$
0.5	0.5	4	0.75	3.73×10^{-6}	0.934	0.113	1.80×10^{-9}
1	1	4	0.31	1	1.08	0.163	1.00
5	5	4	0.38	1	1.01	0.0439	0.965
10	10	4	0.43	1	0.993	0.0304	0.0113
100	100	4	0.83	3.48×10^{-10}	0.992	0.0103	1.22×10^{-14}

TABLE 1

Comparison of the multi-task oracle risk to the single-task oracle risk in Setting A.

C_2	$r = \frac{C_2}{C_1}$	m	$\bar{B}_{100}$	π_1	$\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	$\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	π_2
0.01	0.01	2	1	$< 10^{-15}$	0.570	0.0409	$< 10^{-15}$
0.1	0.1	2	1	$< 10^{-15}$	0.745	0.0333	$< 10^{-15}$
0.5	0.5	2	0.99	$< 10^{-15}$	0.907	0.0406	$< 10^{-15}$
1	1	2	0.80	1.52×10^{-8}	0.961	0.0459	$< 10^{-15}$
5	5	2	0.55	0.607	0.995	0.205	2.59×10^{-3}
10	10	2	0.53	0.835	0.996	0.114	6.23×10^{-4}
100	100	2	0.81	4.50×10^{-9}	0.996	6.35×10^{-3}	1.03×10^{-11}
0.01	0.01	4	1	$< 10^{-15}$	0.527	0.0409	$< 10^{-15}$
0.1	0.1	4	1	$< 10^{-15}$	0.756	0.0534	$< 10^{-15}$
0.5	0.5	4	0.93	$< 10^{-15}$	0.917	0.0650	$< 10^{-15}$
1	1	4	0.49	1	1.01	0.0896	0.855
5	5	4	0.40	1	0.997	0.0295	0.170
10	10	4	0.41	1	0.998	0.0179	0.114
100	100	4	0.84	9.10×10^{-11}	0.994	8.71×10^{-3}	7.36×10^{-14}

TABLE 2

Comparison of the multi-task oracle risk to the single-task oracle risk in Setting B.

δ_2 is large and C_2 is small) then the multi-task oracle seems to outperform the single-task one. On the contrary, when the mean component of the signal is negligible compared to the variance component (likewise, this occurs in Settings A and B when r is large and in Setting C when δ_2 is small or when C_2 large), then both oracles seem to perform similarly.

Adversary settings to the multi-task oracle appear when one task is added outside of a cluster (Setting D). When this outlier is less regular than the tasks belonging to the cluster (that is, when δ_2 is large), the single-task oracle performs better than the multi-task one, which confirms the theoretical analysis performed in Section 6.2.

9. Conclusion

This paper shows the existence of situations where the multi-task kernel ridge regression, with a perfect parameter calibration, can perform better than the single-task one. This happens when the tasks are distributed given simple specifications, which are studied both theoretically and on simulated examples.

The analysis performed here allows us to have a precise estimation of the risk of the multi-task oracle (Theorem 1), this result holding under a few hypotheses on the regularity of the kernel, of the mean of the tasks and of its resulting variance. Several simple single-task settings are then investigated, with the constraint that they respect the latter assumptions. This theoretical grounding, backed-up by our simulated examples, allows us to understand better when and where the multi-task procedure outperforms the single-task one.

- The situation where all the regression functions are close in the RKHS (that is, their differences are extremely regular) is favorable to the multi-task procedure, when using the matrices $\mathcal{M} = \left\{ M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}_+^2 \right\}$. In this setting, the multi-task procedure can do much better than the single-task one (as if it had p times more input points). It is also shown to never do worse (up to a multiplicative constant) !
- On the contrary, when one outlier lies far apart from this cluster, this multi-task procedure suddenly performs badly, that is, arbitrarily worse than the single-task one. This comes as no surprise, since the addition of a far less regular task naturally destroys the joint learning of a group of tasks. In this case, the use of a multi-task procedure which clusters the tasks together (because of the choice of $\mathcal{M}$) is inadapted to the situation.

Our analysis can easily be adapted to a slightly wider set of assumptions on the tasks than the one presented here (all the tasks are grouped together, in one cluster). It is for instance possible to treat the case where the tasks are grouped in two (or more) clusters—when the allocation of each task to its cluster is known to the statistician, at the price of introducing more hyperparameters. We are still limited, though, to certain cases of hypotheses, reflected on the set of matricial hyperparameters $\mathcal{M}$. The failure of the multi-task oracle on the case where one outlier stays outside of one group of tasks can be seen, not as the impossibility to use multi-task techniques in this situation, but rather as the fact the set of matrices used here, $\mathcal{M} = \left\{ M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}_+^2 \right\}$, is inadapted to the situation. We can at least see two different solutions to this kind of inadaptation. First, the use of prior knowledge can help the statistician to craft an *ad hoc* set $\mathcal{M}$. Second, we could seek to automatically adapt to the situation in order to learn a good set $\mathcal{M}$ from data.

Learning more complex sets $\mathcal{M}$ is an important—but complex—challenge, that we want to address in the future. This question can at least be split into three (not necessarily independent) problems, that call for the elaboration of new tools:

- a careful study of the risk, to find a set $\mathcal{M}^\star \subset \mathcal{S}_p^{++}(\mathbb{R})$ of candidate matrices;
- optimization tools, to derive an algorithm able to select a matrix in this set $\mathcal{M}^\star$;
- new concentration of measure results, to be able to show oracle inequalities that control the risk of the output of the algorithm.

Our estimation of the multi-task oracle risk is also shown to be precise enough so that we can plug it in an oracle inequality, hereby showing the existence of a multi-task estimator that has a lower risk than the single-task oracle (under the same favorable circumstances as before).

Finally, it would be intereting to extend the analysis developped here to the random-design setting. This could be done, for instance, by using the tools brought by Hsu et al. [17], that link random-design convergence rates to fixed-design ones.

Acknowledgments: The author thanks Sylvain Arlot and Francis Bach for inspireful discussions and their precious comments, which greatly helped to improve the quality of this paper.

References

[1] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005.

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

[3] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, July 2011. arXiv:0909.1884v2.

[4] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, May 1950.

[5] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. *International Conference on Learning Theory*, (26), 2013.

[6] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, December 2003.

[7] Jonathan Baxter. A model of inductive bias learning. *Journal Of Artificial Intelligence Research*, 12:149–198, 2000.

[8] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

[9] Philip J. Brown and James V. Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1):pp. 64–74, 1980.

[10] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
[11] Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, July 1997.
[12] Bradley Efron and Carl N. Morris. Stein's paradox in statistics. *Scientific American*, 236:119–127, 1977.
[13] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.
[14] Sergey Feldman, Maya R. Gupta, and Bela A. Frigyik. Multi-task averaging. *Advances in Neural Information Processing Systems 25*, pages 1178–1186, 2012.
[15] Chong Gu. *Smoothing spline ANOVA models*. Springer, 2002.
[16] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
[17] Daniel Hsu, Sham M Kakade, and Tong Zhang. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 2011.
[18] Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation. *Computing Research Repository*, pages – 1–1, 2008.
[19] William James and Charles Stein. Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, 1(1961):361–379, 1961.
[20] Iain M. Johnstone. Minimax bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics V*, pages 303–326. Springer New York, 1994.
[21] Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.
[22] Percy Liang, Francis Bach, Guillaume Bouchard, and Michael I. Jordan. Asymptotically optimal regularization in smooth parametric models. In *Advances in Neural Information Processing Systems*, 2010.
[23] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.
[24] Pascal Massart. *Concentration Inequalities and Model Selection*. École d'Été de Probabilités de Saint Flour XXXIII - 2003. Springer, 2007.
[25] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–17, 2011.
[26] Angelika Rohde and Alexandre Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 2011.
[27] Matthieu Solnon, Sylvain Arlot, and Francis Bach. Multi-task Regression using Minimal Penalties. *Journal of Machine Learning Research*, 13:2773–2812, September 2012.
[28] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley symposium*

on mathematical statistics and probability, 1(399):197–206, 1956.

[29] Sebastian Thrun and Joseph O'Sullivan. Discovering structure in multiple learning tasks: The TC algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 1996.

[30] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

Appendices

Appendix A: Decomposition of the matrices $M_{\mathbf{SD}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $M_{\mathbf{AV}}(\boldsymbol{\lambda}, \boldsymbol{\mu})$

We now give a few technical results that were used in the former sections.

Lemma 1. *The penalty used in Equation* (3) *can be obtained by using in Equation* (2) *the matrix* $M_{\mathrm{SD}}(\alpha, \beta)$*, such that*

$$M_{\mathrm{SD}}(\alpha, \beta) = \frac{\alpha}{p} \frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\alpha + p\beta}{p} \left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right) \ . \tag{20}$$

The penalty used in Equation (4) *can be obtained by using in Equation* (2) *the matrix* $M_{\mathrm{AV}}(\alpha, \beta)$*, such that*

$$M_{\mathrm{AV}}(\lambda, \mu) = \frac{\lambda}{p} \frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p} \left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right) \ . \tag{21}$$

Proof. For the first part, since

$$\begin{aligned} \sum_{j=1}^{p} \sum_{k=1}^{p} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 &= \sum_{j,k} \langle g^j, g^j \rangle_{\mathcal{F}} - 2\langle g^j, g^k \rangle_{\mathcal{F}} + \langle g^k, g^k \rangle_{\mathcal{F}} \\ &= 2p \sum_{j=1}^{p} \langle g^j, g^j \rangle_{\mathcal{F}} - 2 \sum_{j,k} \langle g^j, g^k \rangle_{\mathcal{F}} \ , \end{aligned}$$

the penalty term of Equation (3) can be written as

$$\frac{\alpha}{p} \sum_{j=1}^{p} \langle g^j, g^j \rangle_{\mathcal{F}} + \beta \sum_{j=1}^{p} \langle g^j, g^j \rangle_{\mathcal{F}} - \frac{\beta}{p} \sum_{j,k} \langle g^j, g^k \rangle_{\mathcal{F}} \ ,$$

leading to the matrix

$$\frac{\alpha + p\beta}{p} I_p - \frac{\beta}{p} \mathbf{1}\mathbf{1}^\top = \frac{\alpha}{p} \frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\alpha + p\beta}{p} \left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right) = M_{\mathrm{SD}}(\alpha, \beta) \ .$$

For the second part, since

$$\left\| \sum_{j=1}^{p} g^j \right\|_{\mathcal{F}}^2 = \sum_{j,k} \langle g^j, g^k \rangle_{\mathcal{F}} \ ,$$

the penalty term of Equation (4) can be written as

$$\frac{\lambda}{p^2} \sum_{j,k} \langle g^j, g^k \rangle_{\mathcal{F}} + \frac{\mu}{p} \sum_{j=1}^{p} \langle g^j, g^j \rangle_{\mathcal{F}} - \frac{\mu}{p^2} \sum_{j,k} \langle g^j, g^k \rangle_{\mathcal{F}} \ ,$$

leading to the matrix

$$\frac{\lambda - \mu}{p^2} \mathbf{1}\mathbf{1}^\top + \frac{\mu}{p} I_p = \frac{\lambda}{p} \frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p} \left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right) = M_{\mathrm{AV}}(\lambda, \mu) \ .$$

□

Appendix B: Useful control of some sums

Let us introduce the following integrals :

$$I_1 = I_1(\beta, \delta) = \int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du \ ,$$
$$I_2 = I_2(\beta) = \int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du = I_1(\beta, 0) \ .$$

Under Assumption ($\mathbf{H_M}(\beta, \delta)$), both integrals converge. We also introduce their discrete counterparts. For every $n \in \mathbb{N}^\star$ and every $\lambda \in \mathbb{R}_+$:

$$S_1(n, \lambda) = \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \ ,$$
$$S_2(n, \lambda) = \sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} \ .$$

We here give a first elementary technical result.

Lemma 2. *The map defined on $\mathbb{R}_+$ by*

$$t \mapsto \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}$$

is positive, increasing on $[0, t^\star]$ and decreasing on $[t^\star, +\infty)$ to 0, with

$$t^\star = \left(\frac{4\beta - 2\delta}{2\delta\lambda} \right)^{1/2\beta}$$

Proof. This map is nonnegative and converges to 0 in 0 and $+\infty$. Furthermore

$$\begin{aligned}\frac{d}{dt}\left(\frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}\right) &= (4\beta-2\delta)\frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^2} - 4\beta\lambda t^{2\beta-1}\frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^3} \\ &= \frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^3}\left[(4\beta-2\delta)(1+\lambda t^{2\beta}) - 4\beta\lambda t^{2\beta}\right] \\ &= \frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^3}\left[4\beta+4\beta\lambda t^{2\beta}-2\delta-2\delta\lambda t^{2\beta}-4\beta\lambda t^{2\beta}\right] \\ &= \frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^3}\left[(4\beta-2\delta)-2\delta\lambda t^{2\beta}\right] \quad .\end{aligned}$$

The only parameter $t^\star$ that cancels out this equation is

$$t^\star = \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \quad .$$

□

We now give a serie of technical results to control I_1, I_2, S_1 and S_2, which will be useful in the following sections.

Lemma 3.

$$\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt = \frac{\lambda^{(2\delta-1)/2\beta}}{2\beta\lambda^2}\int_0^{+\infty}\frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du = \frac{\lambda^{(2\delta-1)/2\beta}}{2\beta\lambda^2}I_1 \quad .$$

Proof. Apply the change of variables $u = \lambda t^{2\beta}$ see Bach (5) for more details. □

Lemma 4.

$$\int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2}dt = \frac{\lambda^{-1/2\beta}}{2\beta}\int_0^{+\infty}\frac{u^{\frac{1-2\beta}{2\beta}}}{(1+u)^2}du = \frac{\lambda^{-1/2\beta}}{2\beta}I_2 \quad .$$

Proof. Apply the change of variables $u = \lambda t^{2\beta}$ see Bach (5) for more details. □

Lemma 5. *We have the following bounds S_2. For every $n \in \mathbb{N}^\star$ and every $\lambda \in \mathbb{R}_+^\star$,*

- $$S_2(n,\lambda) \leq \frac{\lambda^{-1/2\beta}}{2\beta}I_2 \quad .$$
- $$S_2(n,\lambda) \geq \int_1^{n+1}\frac{1}{(1+\lambda t^{2\beta})^2}dt \quad .$$

Proof. To show the first point we just remark that

$$S_2(n,\lambda) = \sum_{i=1}^n \frac{1}{\left(1+\lambda i^{2\beta}\right)^2} \leq \int_0^n \frac{1}{(1+\lambda t^{2\beta})^2}dt \leq \int_0^{+\infty}\frac{1}{(1+\lambda t^{2\beta})^2}dt \quad .$$

The second point is likewise straightforward. □

Lemma 6. *We have the following bounds on S_1 : for every $n \in \mathbb{N}^\star$, every $(\beta, \delta) \in \mathbb{R}_+^2$ such that $4\beta > 2\delta$ and every $\lambda \in \mathbb{R}_+^\star$,*

$$S_1(n,\lambda) \leq \frac{\lambda^{(2\delta-1)/2\beta}}{\beta\lambda^2} I_1 \ ,$$

Furthermore, let

$$t^\star = \left(\frac{4\beta - 2\delta}{2\delta\lambda}\right)^{1/2\beta}$$

and $n^\star = \lfloor t^\star \rfloor$.

- *If $n^\star < n-1$*

$$S_1(n,\lambda) \geq \int_0^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - \int_{n^\star}^{n^\star+2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ ;$$

- *while if $n^\star \geq n$*

$$S_1(n,\lambda) \geq \int_0^{n} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

Proof. Lemma 2 shows that $t \mapsto \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}$ is increasing on $[0, t^\star]$ and decreasing on $[t^\star, +\infty($. Thus we have the following comparisons :

$$\int_0^{n^\star} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \leq \sum_{i=1}^{n^\star} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \leq \int_1^{n^\star+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt$$

and

$$\int_{n^\star+2}^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \leq \sum_{i=n^\star+1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \leq \int_{n^\star}^{n} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

By adding those two lines we get

$$\begin{aligned} S_1(n,\lambda) = \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} &\leq \int_1^{n^\star+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt + \int_{n^\star}^{n} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \\ &\leq 2\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ , \end{aligned}$$

which shows the first point. We also get, if $n^\star < n-1$

$$\begin{aligned} S_1(n,\lambda) &\geq \int_0^{n^\star} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt + \int_{n^\star+2}^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \\ &\geq \int_0^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - \int_{n^\star}^{n^\star+2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ . \end{aligned}$$

The last point is evident, since if $n^\star \geq n$ the integrand is increasing on $[0, n]$. □

Appendix C: Proof of Property 1

Let n and p be integers, σ, β and δ real numbers such that ($\mathbf{H_M}(\beta,\delta)$) hold. We want to study the value and the location of the infimum on $\mathbb{R}_+$ of

$$\lambda \longmapsto R(n,p,\sigma^2,\lambda,\beta,\delta,C) = C\lambda^2 \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} \quad (22)$$

Property 7. *For every λ in $\mathbb{R}_+$, we have*

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \leq \frac{CI_1}{\beta}\lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np}\lambda^{-1/2\beta} \ . \quad (23)$$

Proof. This is a straightforward application of the majorations of the finite sums by integrals given in Lemmas 5 and 6, together with the change of variables done in Lemmas 3 and 4. □

Lemma 7. *Let $A \in \mathbb{R}_+$, the minimum over $\mathbb{R}_+^\star$ of $\lambda \mapsto \lambda^{(2\delta-1)/2\beta} + A\lambda^{-1/2\beta}$ is attained for*

$$\lambda^\star = \left(\frac{A}{2\delta-1}\right)^{\beta/\delta}$$

and has for value

$$A^{1-(1/2\delta)} \frac{2\delta}{(2\delta-1)^{1-(1/2\delta)}} \ .$$

Proof. This mapping is differentiable and has $+\infty$ for limit in 0 and in $+\infty$. Then

$$\frac{d}{d\lambda}\left(\lambda^{2\delta/(2\delta-1)} + A\lambda^{-1/2\beta}\right) = \frac{1}{\lambda}\left(\frac{2\delta-1}{2\beta}\lambda^{(2\delta-1)/2\beta} - \frac{A}{2\beta}\lambda^{-1/2\beta}\right) \ .$$

We see there is only one minimizer $\lambda^\star$ verifying

$$\begin{array}{rrcl} & \frac{2\delta-1}{2\beta}(\lambda^\star)^{(2\delta-1)/2\beta} & = & \frac{A}{2\beta}(\lambda^\star)^{-1/2\beta} \\ \Leftrightarrow & (2\delta-1)^{2\beta}(\lambda^\star)^{2\delta-1} & = & A^{2\beta}(\lambda^\star)^{-1} \\ \Leftrightarrow & (\lambda^\star)^{2\delta} & = & \frac{A^{2\beta}}{(2\delta-1)^{2\beta}} \\ \Leftrightarrow & \lambda^\star & = & \left(\frac{A}{2\delta-1}\right)^{\beta/\delta} \ . \end{array}$$

Pluging-in the value of $\lambda^\star$ leads to the optimal value

$$\begin{aligned} \left(\frac{A}{2\delta-1}\right)^{(2\delta-1)/2\delta} + A\left(\frac{A}{2\delta-1}\right)^{-1/2\delta} &= A^{(2\delta-1)/2\delta}\left((2\delta-1)^{(1/2\delta)-1} + (2\delta-1)^{1/2\delta}\right) \\ &= A^{(2\delta-1)/2\delta}(2\delta-1)^{1/2\delta}\left(\frac{1}{2\delta-1}+1\right) \\ &= A^{(2\delta-1)/2\delta}(2\delta-1)^{1/2\delta}\left(\frac{2\delta}{2\delta-1}\right) \\ &= A^{(2\delta-1)/2\delta}\frac{2\delta}{(2\delta-1)^{1-(1/2\delta)}} \ . \end{aligned}$$

□

Definition 3. To simplify notations, since this quantity depends only on β and δ and appears throughout the paper, we will use the following notation :

$$\kappa(\beta,\delta) = I_1(\beta,\delta)^{1/2\delta} I_2(\beta)^{1-(1/2\delta)} (2\delta-1)^{1/2\delta} \frac{\delta}{\beta(2\delta-1)} \ . \tag{24}$$

We now prove Property 1

Proof. First $R(n,p,\sigma^2,0,\beta,\delta,C) = \frac{\sigma^2}{p}$, so that $R^\star(n,p,\sigma^2,\beta,\delta,C) \leq \frac{\sigma^2}{p}$. Then, the right-hand side of Equation (23) can be written as

$$\frac{CI_1}{\beta}\left[\lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2npCI_1}\lambda^{-1/2\beta}\right] \ .$$

Consequently, Lemma 7 implies that the optimal value of this upper bound with respect to λ is

$$\frac{CI_1}{\beta}\left(\frac{\sigma^2 I_2}{2npCI_1}\right)^{1-(1/2\delta)} \frac{2\delta}{(2\delta-1)^{1-(1/2\delta)}} \ ,$$

which is exactly the right-hand side of Equation (9). □

Appendix D: Proof of Property 2

In order to perform this analysis we observe that R is composed of two factors :

- a bias factor $C\lambda^2 \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2}$, which is an increasing function of λ;
- a variance factor $\frac{\sigma^2}{np}\sum_{i=1}^{n}\frac{1}{(1+\lambda i^{2\beta})^2}$, which is a convex, decreasing function of λ.

We show that, if λ is too large, then the bias term exceeds the upper bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$ given in Equation (9).

Proof. We see that, using Equation (22), for every $\lambda \in \mathbb{R}_+$,

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \geq C\frac{\lambda^2}{(1+\lambda)^2} \ .$$

The right-hand side of this equation is increasing. Thus, if a real number ε matches this bound with the upper bound of $R^\star$, that is,

$$C\frac{\varepsilon^2}{(1+\varepsilon)^2} = \frac{1}{np} \times (np)^{1/2\delta} C^{(1/2\delta)} 2^{1/2\delta} \kappa(\beta,\delta) \ ,$$

we can state that the infimum of R is attained by a parameter $\lambda^\star \in [0,\varepsilon]$. The latter equation is equivalent to

$$\varepsilon^2 = A\left(\frac{np}{\sigma^2}\right)^{(1/2\delta)-1}(1+\varepsilon)^2 \ ,$$

with

$$A = C^{(1/2\delta)-1} 2^{1/2\delta} \kappa(\beta, \delta) \ .$$

This leads to

$$\varepsilon \left(1 - \sqrt{A} \left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}\right) = \sqrt{A} \left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-2} \ ,$$

so that if $\sqrt{A} \left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2} < 1$ that is, if

$$\frac{np}{\sigma^2} > \frac{1}{C} \times 2^{\frac{1}{2\delta-1}} \times \kappa(\beta, \delta)^{\frac{2\delta}{2\delta-1}} \ ,$$

then

$$\varepsilon = \frac{\sqrt{A} \left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}}{1 - \sqrt{A} \left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}} = \sqrt{A} \left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2} \left(1 + \eta\left(\frac{np}{\sigma^2}\right)\right) \ , \qquad (25)$$

where $\eta(x)$ goes to 0 as x goes to $+\infty$. □

Appendix E: On the way to showing Property 3

The proof of Property 3 uses two results that we give here.

E.1. Control of the risk on $\left[0, n^{-2\beta}\right]$

Property 8. *For every n, p, σ^2, C, δ and β, we have*

$$\inf_{\lambda \in [0, n^{-2\beta}]} \left\{R(n, p, \sigma^2, \lambda, \beta, \delta, C)\right\} \geq \frac{\sigma^2}{4p} \ .$$

Proof. For every $\lambda \in \left[0, n^{-2\beta}\right]$ we have

$$\begin{aligned} R(n, p, \sigma^2, \lambda, \beta, \delta, C) &\geq \frac{\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{\left(1 + \lambda i^{2\beta}\right)^2} \\ &\geq \frac{\sigma^2}{p} \times \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\left(1 + \left(\frac{i}{n}\right)^{2\beta}\right)^2} \\ &\geq \frac{\sigma^2}{4p} \ . \end{aligned}$$

□

E.2. Control of the risk on $\left[n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)\right]$

Property 9. *There exists an integer N and a constant $\alpha \in (0,1)$ such that for every (n,p,σ^2) such that $np/\sigma^2 \geq N$, every $(\beta,\delta) \in \mathbb{R}_+^2$ such that $4\beta > 2\delta > 1$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$ we have*

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \geq \alpha\left(\frac{CI_1}{\beta}\lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np}\lambda^{-1/2\beta}\right) \quad . \tag{26}$$

Proof. We seek to minor the two sums composing R, which was definded in Equation (22), by their integral counterparts, uniformly on $[n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$. The technical details are exposed in Lemmas 5 and 6.

For the first sum, using Lemma 5, we have that

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} &\geq \int_0^{n+1} \frac{1}{(1+\lambda t^{2\beta})^2} dt - \int_0^1 \frac{1}{(1+\lambda t^{2\beta})^2} dt \\ &\geq \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt - \int_{n+1}^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt - \int_0^1 \frac{1}{(1+\lambda t^{2\beta})^2} dt \quad .\end{aligned}$$

First, with the change of variables $u = \lambda t^{2\beta}$ (as in 5),

$$\begin{aligned}\int_{n+1}^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{n+1}^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt}{\int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt} \\ &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{\lambda(n+1)^{2\beta}}^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \\ &\leq \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_1^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \quad ,\end{aligned}$$

since $\lambda \geq n^{-2\beta}$.

We also have , with the change of variables $u = \lambda t^{2\beta}$ (as in 5),

$$\begin{aligned}\int_0^{1} \frac{1}{(1+\lambda t^{2\beta})^2} dt &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_0^{1} \frac{1}{(1+\lambda t^{2\beta})^2} dt}{\int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt} \\ &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_0^{\lambda} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \\ &\leq \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_0^{\varepsilon} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \quad .\end{aligned}$$

Since ε, which was defined in Equation (25), verifies $\varepsilon(x) \longrightarrow 0$ as $x \longrightarrow +\infty$, we get

$$\frac{\int_0^{\varepsilon(x)} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \underset{x\to+\infty}{\longrightarrow} 0 \ .$$

All those arguments imply that there exists an integer n_1 and real number $c_1 \in (0,1)$ such that, for every (n,p,σ^2) such that $np/\sigma^2 \geq n_3$ and for every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$,

$$\sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} \geq c_1 \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \ .$$

For the second sum we carry a similar analysis, using Lemma 6 instead of Lemma 5. First, supposing that $4\beta > 2\delta$, we know that

$$\frac{\left\lfloor \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \right\rfloor}{\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}} \underset{\lambda\to 0}{\longrightarrow} 1 \ .$$

Since $\varepsilon(np/\sigma^2)$ goes to 0 as np/σ^2 goes to $+\infty$. Consequently, let $\zeta > 0$ and n_3 be an integer such that for every (n,p,σ^2) such that $np/\sigma^2 \geq n_3$, and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$, we have

$$\left| \frac{\left\lfloor \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \right\rfloor}{\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}} - 1 \right| < \zeta \quad \text{and} \quad \left| \frac{\left\lfloor \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \right\rfloor + 2}{\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}} - 1 \right| < \zeta \ .$$

Consequently, for every (n,p,σ^2) such that $np/\sigma^2 \geq n_3$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$, we have (with $t^\star = \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}$ and $n^\star = \lfloor t^\star \rfloor$) :

$$n^\star \geq (1-\zeta)\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} = z_1 \quad \text{and} \quad n^\star + 2 \leq (1+\zeta)\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} = z_2 \ .$$

We can remark that $\lambda z_1^{2\beta}$ and $\lambda z_2^{2\beta}$ do not depend on λ. Consequently, for every (n,p,σ^2) such that $np/\sigma^2 \geq n_3$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$, we get

$$\int_{n^\star}^{n^\star+2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \leq \int_{z_1}^{z_2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

We finally see that

$$\begin{aligned}\int_{z_1}^{z_2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \frac{\int_{z_1}^{z_2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt}{\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt} \\ &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \frac{\int_{\lambda z_1^{2\beta}}^{\lambda z_2^{2\beta}} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du} \\ &= c_3 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \ ,\end{aligned}$$

with

$$c_3 = \frac{\int_{\lambda z_1^{2\beta}}^{\lambda z_2^{2\beta}} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du} \in (0,1)$$

being independent of λ and arbitrarily close to 0. Thus, we have that, using Lemma 6,

- if $n^\star \geq n-1$:

$$\begin{aligned}\sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} &\geq \int_0^n \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \\ &\geq \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt - \int_n^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \ ;\end{aligned}$$

- if $n^\star < n-1$ and $np/\sigma^2 \geq n_3$:

$$\begin{aligned}&\sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \\ \geq\ & \int_0^n \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt - c_3 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \\ \geq\ & \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt - \int_n^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt - c_3 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \ .\end{aligned}$$

With the change of variables $u = \lambda t^{2\beta}$ (as in 5),

$$\begin{aligned}\int_n^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2}dt &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \frac{\int_n^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt}{\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt} \\ &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \frac{\int_{\lambda n^{2\beta}}^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du} \\ &\leq \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt \frac{\int_1^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du} \ ,\end{aligned}$$

since $\lambda \geq n^{-2\beta}$. This implies that there exists an integer n_2 and real number $c_2 \in (0,1)$ such that, for every (n,p,σ^2) such that $np/\sigma^2 \geq n_2$ and for every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$,

$$\sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \geq c_2 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

By taking $N = \max(n_1, n_2)$ and $\alpha = \min(c_1, c_2)$, we have that for every (n,p,σ^2) such that $np/\sigma^2 \geq N$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \geq \alpha \left(\frac{CI_1}{2\beta} \lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np} \lambda^{-1/2\beta} \right) \ .$$

□

E.3. Proof of Property 3

Proof. This proof uses two results proved in Sections E.1 and E.2 of the appendix. Property 2 shows that R attains its minimum on $\left[0, \varepsilon\left(\frac{np}{\sigma^2}\right)\right]$, where $\varepsilon(x)$ goes to 0 as x goes to 0. First, Property 8 shows that

$$\inf_{\lambda \in [0, n^{-2\beta}]} \left\{ R(n,p,\sigma^2,\lambda,\beta,\delta,C) \right\} \geq \frac{\sigma^2}{4p} \ .$$

Then, using Property 9 shows that there exists an integer N and a constant $\alpha \in (0,1)$ such that for every (n,p,σ^2) such that $\frac{np}{\sigma^2} \geq N$, every $(\beta,\delta) \in \mathbb{R}_+^2$ such that $4\beta > 2\delta > 1$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$ we have

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \geq \alpha \left(\frac{CI_1}{\beta} \lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np} \lambda^{-1/2\beta} \right) \ . \tag{27}$$

Thus, using the same analysis than for Property 1, we get

$$\inf_{\lambda \in \left[n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)\right]} \left\{ R(n,p,\sigma^2,\lambda,\beta,\delta,C) \right\} \geq \alpha \left(\frac{np}{\sigma^2} \right)^{1/2\delta-1} C^{1/2\delta} \kappa(\beta,\delta) \ .$$

□

E.4. Proof of Property 4

The proof of Property 1 clearly shows two regimes :

- when $\lambda_R^\star \leq n^{-2\beta}$, the multi-task risk is $\asymp \frac{\sigma^2}{p}$;
- when $\lambda_R^\star \geq n^{-2\beta}$, the multi-task risk is $\asymp \left(\frac{\sigma^2}{np}\right)^{1-1/2\delta}$.

We now show that if λ is too close to zero then the variance term exceeds the upper bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$ given in Equation (9).

Proof. Let us denote

$$m_1 = \inf_{\lambda\in[0,n^{-2\beta}]} \left\{R(n,p,\sigma^2,\lambda,\beta,\delta,C)\right\}$$

and

$$m_2 = \inf_{\lambda\in\left[n^{-2\beta},\varepsilon\left(\frac{np}{\sigma^2}\right)\right]} \left\{R(n,p,\sigma^2,\lambda,\beta,\delta,C)\right\} \ .$$

If $m_1 < m_2$ then $\lambda_R^\star \leq \frac{1}{n^{2\beta}}$, else $\lambda_R^\star \geq \frac{1}{n^{2\beta}}$. Under the present assumptions, we can use the proof Property 3 and state that there exists an integer N_1 and a constant $\alpha \in (0,1)$ such that

$$\frac{\sigma^2}{p} \geq m_1 \geq \frac{\sigma^2}{4p} \ ,$$

and

$$2^{1/2\delta}\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} C^{1/2\delta}\kappa(\beta,\delta) \geq m_2 \geq \alpha\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} C^{1/2\delta}\kappa(\beta,\delta) \ .$$

Both assumptions $n^{2\delta-1}\times\frac{\sigma^2}{p} \longrightarrow 0$ and $n^{2\delta-1}\times\frac{\sigma^2}{p} \longrightarrow +\infty$ ensure that either $m_1 > m_2$ or $m_2 < m_1$ asymptotically hold. □

Appendix F: Study of the different multi-task hypotheses

Lemma 8. *Under Assupmtion* ($\mathbf{H_{AV}}(\delta,C_1,C_2)$)*, Assumption* (2Points) *is equivalent to*

$$\exists(\varepsilon_i)_{i\in\mathbb{N}} \in \{-1,1\}^{\mathbb{N}}, \ \forall i \in \{1,\dots,n\},$$
$$\left\{\begin{array}{rcl} \forall j \in \left\{1,\dots,\frac{p}{2}\right\}, \ h_i^j &=& \sqrt{n}i^{-\delta}(\sqrt{C_1}+\varepsilon_i\sqrt{C_2}) \\ \forall j \in \left\{\frac{p}{2}+1,\dots,p\right\}, \ h_i^j &=& \sqrt{n}i^{-\delta}(\sqrt{C_1}-\varepsilon_i\sqrt{C_2}) \end{array}\right. .$$

The risk of the estimator $\widehat{f}_\lambda^j = A_\lambda y^j$ for the jth task, which we denote by $R^j(\lambda)$, verifies

$$R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}-\sqrt{C_2}\right)^2) \leq R^j(\lambda)$$

and

$$R^j(\lambda) \leq R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}+\sqrt{C_2}\right)^2) \ .$$

Proof. We have that, for every $i \in \{1,\dots,n\}$

$$\left\{\begin{array}{rcl} \frac{\mu_i}{\sqrt{p}} &=& \frac{1}{2}h_i^1+\frac{1}{2}h_i^p \\ \varsigma_i^2 &=& \frac{1}{2}\left(h_i^1-\frac{\mu_i}{\sqrt{p}}\right)^2+\frac{1}{2}\left(h_i^p-\frac{\mu_i}{\sqrt{p}}\right)^2 \end{array}\right.$$
$$\Leftrightarrow\left\{\begin{array}{rcl} h_i^p &=& 2\frac{\mu_i}{\sqrt{p}}-h_i^1 \\ \varsigma_i^2 &=& \frac{1}{2}\left(h_i^1-\frac{\mu_i}{\sqrt{p}}\right)^2+\frac{1}{2}\left(2\frac{\mu_i}{\sqrt{p}}-h_i^1-\frac{\mu_i}{\sqrt{p}}\right)^2 \end{array}\right.$$
$$\Leftrightarrow\left\{\begin{array}{rcl} h_i^p &=& 2\mu_i-h_i^1 \\ \varsigma_i^2 &=& \left(h_i^1-\mu_i\right)^2 \end{array}\right.$$

This is equivalent to $h_i^1 = \frac{\mu_i}{\sqrt{p}} + \varsigma_i$ and $h_i^p = \frac{\mu_i}{\sqrt{p}} - \varsigma_i$. Thus, the first point is proved. For the second point, let $j \in \{1, \dots, p\}$. There exists $(\varepsilon_i)_{i\in\mathbb{N}} \in \{-1,1\}^{\mathbb{N}}$ such that $(h_i^j)^2 = ni^{-2\delta}\left(\sqrt{C_1} + \varepsilon_i\sqrt{C_2}\right)^2$. The risk of $\widehat{f}_\lambda^j$ then is

$$\lambda^2 \sum_{i=1}^n \frac{i^{4\beta-2\delta}\left(\sqrt{C_1} + \varepsilon_i\sqrt{C_2}\right)^2}{(1+\lambda i^{2\beta})^2} + \frac{\sigma^2}{n}\sum_{i=1}^n \frac{1}{(1+\lambda i^{2\beta})^2} \ .$$

We conclude by seeing that, for every $\varepsilon \in \{-1,1\}$, we have $\left(\sqrt{C_1} - \sqrt{C_2}\right)^2 \leq \left(\sqrt{C_1} + \varepsilon\sqrt{C_2}\right)^2 \leq \left(\sqrt{C_1} + \sqrt{C_2}\right)^2$ □

Lemma 9. *Under Assupmtion* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$)*, Assumption* (1Out) *is equivalent to*

$$\exists(\varepsilon_i) i \in \mathbb{N} \in \{-1,1\}^{\mathbb{N}}\,,\ \forall i \in \{1,\dots,n\},$$
$$\begin{cases} \forall j \in \{1,\dots,p-1\}\,, & h_i^j = \sqrt{n} i^{-\delta}\left(\sqrt{C_1} + \varepsilon_i\sqrt{\frac{C_2}{p-1}}\right) \\ & h_i^p = \sqrt{n} i^{-\delta}(\sqrt{C_1} - \varepsilon_i\sqrt{(p-1)C_2}) \end{cases} .$$

If $j \in \{1,\dots,p-1\}$, the risk of the estimator $\widehat{f}_\lambda^j = A_\lambda y^j$ for the jth task, which we denote by $R^j(\lambda)$, verifies

$$R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} - \sqrt{\frac{C_2}{p-1}}\right)^2) \leq R^j(\lambda)$$

and

$$R^j(\lambda) \leq R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^2) \ ,$$

while the risk of the estimator $\widehat{f}_\lambda^p = A_\lambda y^p$ for the pth task, which is denoted by $R^p(\lambda)$, verifies

$$R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} - \sqrt{(p-1)C_2}\right)^2) \leq R^p(\lambda)$$

and

$$R^p(\lambda) \leq R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} + \sqrt{(p-1)C_2}\right)^2) \ .$$

Proof. For the first part, we have that, for every $i \in \{1, \dots, n\}$

$$\begin{array}{rl}
& \left\{ \begin{array}{rcl} \frac{\mu_i}{\sqrt{p}} & = & \frac{p-1}{p} h_i^1 + \frac{1}{p} h_i^p \\ \varsigma_i^2 & = & \frac{p-1}{p} \left(h_i^1 - \frac{\mu_i}{\sqrt{p}} \right)^2 + \frac{1}{p} \left(h_i^p - \frac{\mu_i}{\sqrt{p}} \right)^2 \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{rcl} h_i^p & = & p \frac{\mu_i}{\sqrt{p}} - (p-1) h_i^1 \\ \varsigma_i^2 & = & \frac{p-1}{p} \left(h_i^1 - \frac{\mu_i}{\sqrt{p}} \right)^2 + \frac{1}{p} \left(p \frac{\mu_i}{\sqrt{p}} - (p-1) h_i^1 - \frac{\mu_i}{\sqrt{p}} \right)^2 \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{rcl} h_i^p & = & p \frac{\mu_i}{\sqrt{p}} - (p-1) h_i^1 \\ \varsigma_i^2 & = & \frac{p-1}{p} \left(h_i^1 - \frac{\mu_i}{\sqrt{p}} \right)^2 + \frac{(p-1)^2}{p} \left(h_i^1 - \frac{\mu_i}{\sqrt{p}} \right)^2 \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{rcl} h_i^p & = & p \frac{\mu_i}{\sqrt{p}} - (p-1) h_i^1 \\ \varsigma_i^2 & = & (p-1) \left(h_i^1 - \frac{\mu_i}{\sqrt{p}} \right)^2 \end{array} \right.
\end{array}$$

This is equivalent to saying that there exists $(\varepsilon_i)_{i \in \mathbb{N}} \in \{-1, 1\}^{\mathbb{N}}$ such that

$$\begin{array}{rl}
& \left\{ \begin{array}{rcl} h_i^p & = & p \frac{\mu_i}{\sqrt{p}} - (p-1) h_i^1 \\ h_i^1 & = & \frac{\mu_i}{\sqrt{p}} + \frac{\varepsilon_i}{\sqrt{p-1}} \varsigma_i \end{array} \right. \\
\Leftrightarrow & \left\{ \begin{array}{rcl} h_i^1 & = & \frac{\mu_i}{\sqrt{p}} + \frac{\varepsilon_i}{\sqrt{p-1}} \varsigma_i \\ h_i^p & = & \frac{\mu_i}{\sqrt{p}} - \varepsilon_i \sqrt{p-1} \varsigma_i \end{array} \right.
\end{array}$$

□

Lemma 10. *Assumption* ($\mathbf{H_{2Points}}$) *implies Assumption* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$).

Proof. For every $i \in \{1, \dots, n\}$, we suppose we have

$$\left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta} (\sqrt{C_1} + \sqrt{C_2}) \\ h_i^p & = & \sqrt{n} i^{-\delta} (\sqrt{C_1} - \sqrt{C_2}) \end{array} \right. .$$

Thus,

$$\mu_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} h_i^j = \frac{\sqrt{p}}{2} \left(h_i^1 + h_i^p \right) = \sqrt{p} \times \sqrt{n} i^{-\delta} \sqrt{C_1} \ ,$$

so that $\mu_i^2 = p C_1 n i^{-2\delta}$. Furthermore,

$$\varsigma_i^2 = \frac{1}{p} \sum_{j=1}^{p} \left(h_i^j - \frac{\mu_i}{\sqrt{p}} \right)^2 = \frac{1}{p} \sum_{j=1}^{p} \left(\sqrt{n} i^{-\delta} \sqrt{C_2} \right)^2 = C_2 n i^{-2\delta} \ .$$

□

Lemma 11. *Assumption* ($\mathbf{H_{1Out}}$) *implies Assumption* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$).

Proof. For every $i \in \{1, \ldots, n\}$, we suppose we have

$$\left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1} + \frac{1}{\sqrt{p-1}} \sqrt{C_2} \right) \\ h_i^p & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1} - \sqrt{p-1} \sqrt{C_2} \right) \end{array} \right. .$$

Thus,

$$\mu_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} h_i^j = \frac{1}{\sqrt{p}} \left((p-1) h_i^1 + h_i^p \right) = \sqrt{p} \times \sqrt{n} i^{-\delta} \sqrt{C_1} \ ,$$

so that $\mu_i^2 = p C_1 n i^{-2\delta}$. Furthermore,

$$\begin{aligned} \varsigma_i^2 &= \frac{1}{p} \sum_{j=1}^{p} \left(h_i^j - \frac{\mu_i}{\sqrt{p}} \right)^2 \\ &= \frac{1}{p} \left[(p-1) \left(\sqrt{n} i^{-\delta} \frac{\sqrt{C_2}}{\sqrt{p-1}} \right)^2 + \left(\sqrt{p-1} \sqrt{n} i^{-\delta} \sqrt{C_2} \right)^2 \right] = C_2 n i^{-2\delta} \ . \end{aligned}$$

□