

Large Dimensional Asymptotics of Multi-Task Learning

Malik Tiomoko, Cosme Louart, Romain Couillet

▶ **To cite this version:**

Malik Tiomoko, Cosme Louart, Romain Couillet. Large Dimensional Asymptotics of Multi-Task Learning. ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, May 2020, Barcelone (virtual), Spain. pp.8787-8791, 10.1109/ICASSP40776.2020.9053557 . hal-02965810

HAL Id: hal-02965810
https://hal.science/hal-02965810

Submitted on 19 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LARGE DIMENSIONAL ASYMPTOTICS OF MULTI-TASK LEARNING

Malik Tiomoko[1,2]*, Cosme Louart*[2,3] *and Romain Couillet*[2,1*]

[1]CentraleSupélec, University of Paris–Saclay, [2]GIPSA-lab, University of Grenoble–Alpes, [3]CEA-LIST

ABSTRACT

Inspired by human learning, which transfers knowledge from learned tasks to solve new tasks, multitask learning aims at simultaneously solving multiple tasks by a smart exploitation of their similarities. How to relate the tasks so to optimize their performances is however a largely open problem.

Based on a random matrix approach, this article proposes an asymptotic analysis of a support vector machine-inspired multitask learning scheme. The asymptotic performance of the algorithm, validated on both synthetic and real data, sets forth the relation between the statistics of the data in each task and the hyperparameters relating the tasks together. The article, as such, provides first insights on an offline control of multitask learning, which finds natural connections to the currently popular transfer learning paradigm.

Index Terms— Multitask Learning, Transfer Learning, Random Matrix Theory.

1. INTRODUCTION

The performance of regression and classification tasks can be improved by making the tasks learn from each other. In machine learning, the subject is approached by multitask learning (MTL) where, by appropriately sharing the algorithm parameters, each task may benefit from the others. This has been widely used in face recognition [1], citywide passenger flow prediction [2], cancer survival analysis [3], market preferences [4, 5], to cite a few.

Modelling the relatedness between the tasks is one of the key aspects of MTL. Among the proposed approaches, hierarchical Bayes models [4, 5, 6, 7] are one of the most considered; these assume that the tasks share a common hypothesis class. Natural extensions of standard one-task methods (such as support vector machines) have also been devised [8]. In this precise case, to further cope with the possibly large complexity of training these algorithms, explicit least square alternatives have been proposed [9].

Following up on the asymptotic analysis performed in [10] for least-square support vector machines (in a single task setting), or in [11] for a semi-supervised extension, this article provides a large dimensional analysis of multitask least square support vector machines.

The analysis performed allows to predict the classification error, then providing some insights about the algorithm and an automatic way to tune the hyper parameters.

The remainder of the article is organized as follows. Section 2 introduces the multitask learning setting analyzed in Section 3.2, which provides the asymptotic classification error and discusses the results. Section 4 provides supporting experiments in both synthetic and real data settings, demonstrating a strong consistency with the theoretical findings.

Notation. $e_m^{[n]} \in \mathbb{R}^n$ is the canonical vector with $[e_m^{[n]}]_i = \delta_{mi}$. Moreover, $e_{ij}^{[2k]} = e_{2(i-1)+j}^{[2k]}$. The notation $A \otimes B$ for matrices or vectors A, B is the Kronecker product.

2. SETTING AND ASSUMPTIONS

Let $X = [X_1, \ldots, X_k] \in \mathbb{R}^{p \times n}$ be the collection of n independent data vectors drawn from k "tasks". Task i is a binary classification problem from the training samples $X_i = [X_i^{(1)}, X_i^{(2)}] \in \mathbb{R}^{p \times n_i}$ with $X_i^{(j)} = [x_{i1}, \ldots, x_{in_{ij}}] \in \mathbb{R}^{p \times n_{ij}}$ the n_{ij} vectors of class $j \in \{1, 2\}$ for the task i.

The multitask learning least square support vector machine (MTL-LSSVM) aims to predict an output $y \in \{-1, 1\}$ for any input vector $x \in \mathbb{R}^p$. To this end, MTL-LSSVM determines k hyperplanes $W = [\omega_1, \omega_2, \ldots, \omega_k]$ and a bias term $b = [b_1, b_2, \ldots, b_k]^\mathsf{T}$ that minimize the objective function with constraints

$$\min_{W \in \mathbb{R}^{kp}, b \in \mathbb{R}^k} \mathcal{J}(W, b) = \frac{1}{2}\mathrm{tr}(W^\mathsf{T} W) + \frac{\gamma}{2} \sum_{i=1}^{k} \|\xi_i\|^2,$$
$$\text{s.t.} \quad y_i = X_i^\mathsf{T} \omega_i + b_i \mathbb{1}_{n_i} + \xi_i \quad \forall i \in \{1, \ldots, k\}$$

with $y_i = [e_i \otimes \mathbb{1}_{n_{i1}}^\mathsf{T}, e_i \otimes -\mathbb{1}_{n_{i2}}^\mathsf{T}]^\mathsf{T}$ the label of the training data of task i, and $\xi_i \in \mathbb{R}^{n_i}$ the errors on the training data. The hyperparameter γ compromises the smallness of the errors $(\xi_i)_{1 \le i \le k}$ against the accuracy of the hyperplanes W.

In order to incorporate the relatedness between tasks, [9] further considers that each hyper plane ω_i can be written under the form $\omega_i = v_i + \omega_0$, where ω_0 carries the "common" information between tasks and v_i is specialized to each task. The cost function thus becomes

$$\mathcal{J}(\omega_0, V, b) = \frac{1}{2}\|\omega_0\|^2 + \frac{\lambda}{2k}\mathrm{tr}(V^\mathsf{T} V) + \frac{\gamma}{2} \sum_{i=1}^{k} \|\xi_i\|^2 \quad (1)$$

where the parameter λ enforces the relatedness of the tasks and $V = [v_1, \ldots, v_k]$. Figure 1 schematically depicts the multitask learning framework.

By introducing the Lagrangian parameter α and solving the dual formulation of the optimization problem, the solution of (1) is explicit and reads

$$\omega_i = \left(e_i^{[k]\mathsf{T}} \otimes I_p\right) A Z^\mathsf{T} \alpha$$

where

$$\alpha = Q(y - Pb)$$
$$b = (P^\mathsf{T} Q P)^{-1} P^\mathsf{T} Q y$$

*Couillet's work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006), the IDEX GSTATS Chair, and the MIAI LargeDATA Chair at University Grenoble Alpes.

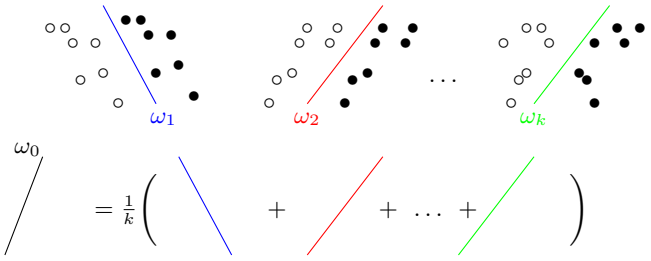


Fig. 1. Hyperplanes of the multitask learning setting.

with $y = [y_1^\mathsf{T}, y_2^\mathsf{T}, \ldots, y_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{kn}$, $Q = (\frac{Z^\mathsf{T} A Z}{kp} + \frac{1}{\gamma} I_{kp})^{-1}$, $A = (\frac{2}{\lambda} I_k + \mathbb{1}_k \mathbb{1}_k^\mathsf{T}) \otimes I_p$, $Z = \sum_{i=1}^k e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes X_i$, and $P = \sum_{i=1}^k e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes \mathbb{1}_{n_i}$.

The prediction of the label of any new data point $x \in \mathbb{R}^p$ for task i is then obtained from classification score $g_i(x)$ given by

$$g_i(x) = \frac{1}{kp} \left(e_i^{[k]} \otimes x \right)^\mathsf{T} AZ\alpha + b_i. \tag{2}$$

Our objective is to quantify the performance of multitask learning, and thus of the (a priori intricate) statistics of $g_i(x)$. To this end, we resort to a large dimensional analysis of $g_i(x)$ in the limit of large p and n_{ij}. In order to keep our results valid for very general (random) datasets, we work under the following concentration of measure assumption for X.

Assumption 1 (Distribution of X)**.** *There exist two constants $C, c > 0$ (independent of n, p) such that, for any 1-Lipschitz function $f : \mathbb{R}^{p \times n} \to \mathbb{R}$,*

$$\forall t > 0 : \mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq C e^{-(t/c)^2}.$$

We further define $\mathbb{E}[x_{ij}] = \mu_{ij}$ and $\mathrm{Cov}[x_{i\ell}] = \Sigma_{i\ell}$, which only depends on (i, j).

Assumption 1 notably encompasses the following scenarios: $x_{ij} \in \mathbb{R}^p$ are (i) independent Gaussian random vectors with covariance of bounded norm, (ii) independent random vectors uniformly distributed on the $\mathbb{R}^p$ sphere, and most importantly (iii) any Lipschitz transformation $\phi(x_{ij})$ of the above two cases, with bounded Lipschitz norm. Scenario (iii) is particularly relevant to model very realistic data settings as it was recently shown [12] that random data (in particular images) generated by generative adversarial networks (better known as GANs) are by definition concentrated random vectors.

For our analysis, we further place ourselves under the following large p, n_{ij} regime.

Assumption 2 (Growth Rate)**.** *As $n \to \infty$, $n/p \to c_0 > 0$ and for $1 \leq i \leq k$, $1 \leq j \leq 2$, $\frac{n_{ij}}{n} \to c_{ij} > 0$. Besides, $\|\mu_{ij}\| = \mathcal{O}(\sqrt{p})$.*

The assumption on the norm $\|\mu_{ij}\|$ is quite natural (typically, the norm of the vectors $x_{i\ell}$ is the same as that of their means). Yet, in order to reach non-trivial (asymptotically simple) results in the following, it will be necessary that $\|\mu_{ij} - \mu_{i'j'}\| = O(1)$ (i.e., tasks are sufficiently similar).

Under these assumptions, by a careful exploitation of tools from random matrix theory, we can derive the asymptotic distribution of the classification score $g_i(x)$, which is our main result, discussed next.

3. MAIN RESULT

3.1. Asymptotic classification error of MTL-LSSVM

Using algebraic identities and the definition of α, $g_i(x)$ can be conveniently rewritten as

$$\begin{aligned} g_i(x) &= \frac{1}{kp} \left(e_i^{[k]} \otimes x \right)^\mathsf{T} AZQ(y - Pb) + b_i \\ &= \frac{1}{kp} \left(e_i^{[k]} \otimes x \right)^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z(y - Pb) + b_i \end{aligned} \tag{3}$$

where $\tilde{Q} = (\frac{A^{\frac{1}{2}} Z Z^\mathsf{T} A^{\frac{1}{2}}}{kp} + \frac{1}{\gamma} I_{kp})^{-1}$.

To compute the statistics of $g_i(x)$, we shall resort to determining so-called *deterministic equivalents* for the matrices $\tilde{Q}$, $\tilde{Q} A^{\frac{1}{2}} Z$, etc., which appear at the core of the formulation of $g_i(x)$. A deterministic equivalent, say $\bar{F} \in \mathbb{R}^{n \times p}$, of a given random matrix $F \in \mathbb{R}^{n \times p}$, denoted $F \leftrightarrow \bar{F}$, is defined by the fact that, for any deterministic linear functional $f : \mathbb{R}^{n \times p} \to \mathbb{R}$, $f(F - \bar{F}) \to 0$ almost surely (for instance, for u, v of unit norm, $u^\mathsf{T}(F - \bar{F})v \xrightarrow{\text{a.s.}} 0$ and, for $A \in \mathbb{R}^{p \times n}$ deterministic of bounded operator norm, $\frac{1}{n} \mathrm{tr} A(F - \bar{F}) \xrightarrow{\text{a.s.}} 0$). Deterministic equivalents are thus particularly suitable to handle bilinear forms involving the random matrix F. Since $g_i(x)$ is precisely a bilinear form involving $\tilde{Q} A^{\frac{1}{2}} Z$, the following lemma is our key technical result.

Lemma 1 (Deterministic equivalents)**.** *Define, for class j in task i, the data statistics matrices*

$$\begin{aligned} M &= \left(e_1^{[k]} \otimes [\mu_{11}, \mu_{12}], \ldots, e_k^{[k]} \otimes [\mu_{k1}, \mu_{k2}] \right) \\ C_{ij} &= A^{\frac{1}{2}} \left(e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes (\Sigma_{ij} + \mu_{ij} \mu_{ij}^\mathsf{T}) \right) A^{\frac{1}{2}}. \end{aligned}$$

Then we have the deterministic equivalents of first order

$$\begin{aligned} \tilde{Q} \leftrightarrow \bar{\tilde{Q}} &\equiv \left(\sum_{i=1}^k \sum_{j=1}^2 \frac{c_{ij}}{c_0} \frac{C_{ij}}{1 + \Delta_{ij}} + \frac{1}{\gamma} I_{kp} \right)^{-1} \\ Q \leftrightarrow \bar{Q} &\equiv \sum_{i,j} \frac{\gamma}{1 + \Delta_{ij}} I_n - \frac{\gamma}{n} J M_\Delta^\mathsf{T} \bar{\tilde{Q}} M_\Delta J^\mathsf{T} \\ A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z &\leftrightarrow M^\mathsf{T} A^{\frac{1}{2}} \bar{\tilde{Q}} A^{\frac{1}{2}} M_\Delta J^\mathsf{T} \end{aligned}$$

and of second order

$$\begin{aligned} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} &\leftrightarrow \mathcal{A}_{ij} \\ Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z &\leftrightarrow M_\Delta \mathcal{A}_{ij} M_\Delta + \mathcal{E} - M_\Delta^\mathsf{T} \bar{\tilde{Q}} M_\Delta \mathcal{C} \end{aligned}$$

in which we defined

$$\begin{aligned} \mathcal{C} &= \mathrm{diag} \left(\frac{I_{n_{11}}}{1 + \Delta_{11}} \mathrm{tr}\,(C_{11}), \ldots, \frac{I_{n_{k2}}}{1 + \Delta_{k2}} \mathrm{tr}\,(C_{k2}) \right) \\ \mathcal{E} &= \sum_{i,j} \mathrm{tr}(C_{ij} \mathcal{A}_{ij}) e_{ij}^{[2k]} e_{ij}^{[2k]\mathsf{T}} \\ \mathcal{A}_{ij} &= \bar{\tilde{Q}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \bar{\tilde{Q}} + \sum_{i,j} \frac{d_{ij}}{n} \mathrm{tr}(A \bar{\tilde{Q}} C_{ij} \bar{\tilde{Q}}) \mathcal{B}_{ij} \\ \mathcal{B}_{ij} &= \bar{\tilde{Q}} C_{ij} \bar{\tilde{Q}} + \sum_{l=1}^k \sum_{m=1}^2 d_{lm} T_{(il)(jm)} \bar{\tilde{Q}} C_{lm} \bar{\tilde{Q}} \\ D &= \sum_{i,j} d_{ij} e_{ij}^{[2k]} e_{ij}^{[2k]\mathsf{T}}, \; d_{ij} = \frac{n_{ij}}{n(1 + \Delta_{ij})^2} \end{aligned}$$

$$J = [j_{11}, \ldots, j_{k2}],$$
$$j_{lm} = \left(0_{n_{11}+\ldots+n_{(i-1)2}}^\mathsf{T}, \mathbb{1}_{n_{ij}}^\mathsf{T}, 0_{n_{(i+1)1}+\ldots+n_{k2}}^\mathsf{T}\right)^\mathsf{T},$$
$$M_\Delta = M \sum_{ij} \frac{1}{1+\Delta_{ij}} e_{ij}^{[2k]} e_{ij}^{[2k]\mathsf{T}}$$
$$S_{ij} = e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes \Sigma_{ij}$$
$$T = \tilde{T}\left(I_{2k} - D\tilde{T}\right)^{-1},\ \tilde{T}_{(il)(jm)} = \frac{1}{n}\mathrm{tr}(C_{ij}\bar{\tilde{Q}}C_{lm}\bar{\tilde{Q}})$$

and the $(\Delta_{11}, \ldots, \Delta_{k2})$ *are the unique positive solution of*

$$\Delta_{ij} = \frac{1}{p}\mathrm{tr}(C_{ij}\bar{\tilde{Q}}),\ \forall i, j.$$

With the help of Lemma 1, we then have our main result.

Theorem 1. *Under Assumptions 1–2 and the notations of Lemma 1,*

$$g(x) \to \mathcal{N}(m_{ij}, \sigma_{ij}^2)$$

in distribution where

$$m_{ij} = \frac{1}{kp}\left(e_i^{[k]} \otimes \mu_{ij}\right)^\mathsf{T} A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}} M_\Delta J^\mathsf{T}(y - P\bar{b}) + \bar{b}_i$$
$$\sigma_{ij}^2 = \frac{1}{(kp)^2}(y - P\bar{b})^\mathsf{T}\left(M_\Delta^\mathsf{T}\mathcal{A}_{ij}M_\Delta + \mathcal{E}\right)(y - P\bar{b})$$
$$- \frac{1}{p^2}(y - P\bar{b})^\mathsf{T} M_\Delta^\mathsf{T}\bar{\tilde{Q}}M_\Delta\mathcal{C}(y - P\bar{b})$$

with $\bar{b} = \left(P^\mathsf{T}\bar{Q}P\right)^{-1}\left(P^\mathsf{T}\bar{Q}y\right)$.

Sketch of Proof. Using the definition of the score in (3), the mean m_{ij} and the variance σ_{ij} are respectively given by:

$$m_{ij} = \mathbb{E}\left[\frac{1}{kp}e_{ij}^{[2k]\mathsf{T}} M^\mathsf{T} A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}} Z(y - Pb)\right]$$
$$\sigma_{ij}^2 = \mathbb{E}\left[\frac{1}{(kp)^2}(y - Pb)^\mathsf{T} Z^\mathsf{T} A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}} Z(y - Pb)\right].$$

Both quantities involve expectation of bilinear forms involving the random matrices $A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z$ and $Z^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z$ for which Lemma 1 provides deterministic equivalents (these being obtained from random matrix tools from, e.g., [13], omitted here for conciseness). The result of the theorem then unfolds from the lemma. □

Since $g_i(x)$ has an asymptotic Gaussian limit centered about m_{ij}, the (asymptotic) standard decision for x to be in Class 1 ($x \in \mathcal{C}_1$) or Class 2 ($x \in \mathcal{C}_2$) for Task i is obtained by the "averaged-mean" test

$$g_i(x) \equiv \frac{1}{\sqrt{kp}}\left(e_i^{[k]} \otimes x\right)^\mathsf{T} AZ^\mathsf{T}\alpha + b_i \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \frac{1}{2}(m_{i1} + m_{i2})$$

the probability of classification error of which is, from Theorem 1,

$$P\left(g_i(x) \geq \frac{m_{i1}+m_{i2}}{2}\middle| x \in \mathcal{C}_1\right) = Q\left(\frac{m_{i1}-m_{i2}}{2\sigma_{i1}}\right) + o(1) \tag{4}$$

with m_{ij}, σ_{ij} defined in Theorem 1 and $Q(t) = \int_t^\infty e^{-\frac{u^2}{2}}du$.

In practice, the values of the means and variances can be consistently estimated. Indeed, they mostly involve up-to-$2k$ dimensional vectors of inner products of the means μ_{ij} or bilinear forms of the covariance matrices Σ_{ij}; random matrix methods have long been developed to obtain such estimates.

3.2. Discussion about MTL-LSSVM

Theorem 1 is quite involved and seemingly leaves little room to interpretation. We show here that, in simplified settings, interesting intuitions in fact naturally arise.

First, it is interesting to note that the matrix $A = (\frac{2}{\lambda}I_k + \mathbb{1}_k\mathbb{1}_k^\mathsf{T}) \otimes I_p$ weighs the constraint of common hyperplane through the term $\mathbb{1}_k\mathbb{1}_k^\mathsf{T}$ against the need for isolating tasks through the term $\frac{2}{\lambda}I_k$, with λ compromising the two terms.

In particular, letting $\gamma \ll 1$, the matrix $\bar{Q}$ and $\bar{\tilde{Q}}$ (which, as in conventional least-square methods, mostly control the variance of the algorithm) are essentially proportional to identity matrices. Further assuming equal sized data per class and per task, up to a leading constant κ_i, the averaged scores m_{ij} simplify as:

$$m_{ij} \approx \kappa_i \sum_{a=1}^{k}\sum_{b=1}^{2}\left(\frac{2}{\lambda}\delta_{ia} + 1\right)\mu_{ij}^\mathsf{T}\mu_{ab}(-1)^{b-1} + \bar{b}_i.$$

As such, $\lambda \gg 1$, m_{ij} is driven by $\sum_{a,b}\mu_{ij}^\mathsf{T}\mu_{ab}(-1)^{b-1}$ so that the distance between μ_{i1} and μ_{i2} depends on the difference in the projections $\mu_{i1}^\mathsf{T}(\sum_a \mu_{a1} - \sum_a \mu_{a2})$ and $\mu_{i2}^\mathsf{T}(\sum_a \mu_{a1} - \sum_a \mu_{a2})$. This is all the more convenient that the μ_{a1} (and μ_{a2}) are correctly aligned across a: in this case the tasks learn from each other. If instead $\lambda \ll 1$, μ_{i1} and μ_{i2} differ by their projections onto $\mu_{i1} - \mu_{i2}$ and the tasks become independent of each other.

Note also interestingly that, if there exists a task a for which $\|\mu_{a1} - \mu_{a2}\| \gg 1$ and that $\mu_{ij}^\mathsf{T}\mu_{aj}$ is positive and non-vanishing, then Task a will make the possibly non-trivial Task i much easier (and in fact asymptotically trivial). Exploiting the fact that $J^\mathsf{T}y$ leverages the size of each class in each task, the same reasoning holds for any Task a for which $n_{a1}, n_{a2} \gg 1$ and $\mu_{ij}^\mathsf{T}\mu_{aj}$ is positive and non-vanishing.

4. EXPERIMENTS

4.1. Application to synthetic data

The asymptotic classification error derived in Equation (4) opens the possibility to automatically tune the hyperparameters of the algorithm. We focus here our analysis on λ which weighs the relatedness between tasks. As previously mentioned, since the statistics m_{ij} and σ_{ij}^2 can be asymptotically estimated, the value of λ minimizing the probability of error can be estimated (i.e., the algorithm can "auto-tune" the relatedness of tasks) by solving

$$\min_{\lambda>0} Q\left(\frac{m_{i1}-m_{i2}}{2\sigma_{i1}}\right),\ m_{ij} = m_{ij}(\lambda),\ \sigma_{ij} = \sigma_{ij}(\lambda).$$

We experiment this approach on the following two-task setting ($k = 2$): $x_{1\ell} \sim \mathcal{N}(\pm\mu_1, I_p)$ and $x_{2\ell} \sim \mathcal{N}(\pm\mu_2, I_p)$, where $\mu_2 = \beta\mu_1 + \sqrt{1-\beta^2}\mu_1^\perp$, where $\mu_1^\perp$ is any vector orthogonal to μ_1 and $\beta \in [0, 1]$. This setting allows us to tune, through β, the similarity between tasks.

Figure 2 compares, for different values of λ, the theoretical and empirical classification errors, and emphasizes the error-minimizing value of λ. Despite the not-so-large values assumed by n and p, a very precise match is reported between the asymptotic theory and the practical experiment, with in particular an accurate estimation of the optimal value for λ.

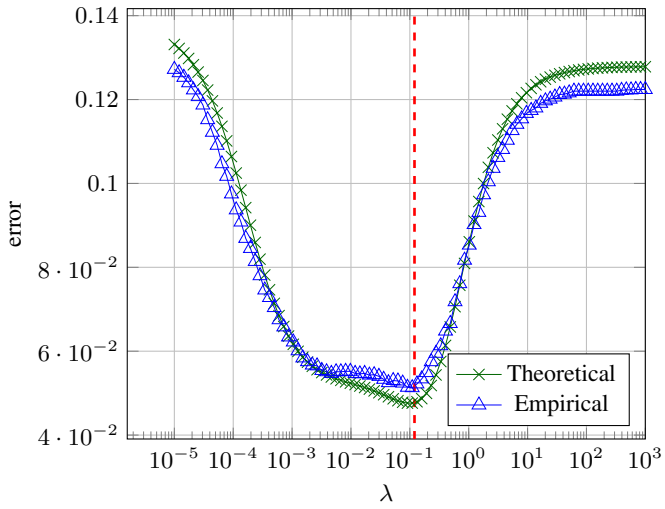


Fig. 2. Theoretical and empirical error of classification as function of λ and predicted value of λ in red dashed line. $p = 256$, $n_{11} = n_{12} = 512$, $n_{21} = n_{22} = 64$, $\mu_1 = \sqrt{2}(e_1^{[p]} + e_2^{[p]})$, $\mu_1^{\perp} = \sqrt{2}(e_3^{[p]} + e_4^{[p]})$, $\beta = 0.5$, $\gamma = 10^{-3}$.

4.2. Application to real dataset

To extend the theoretical analysis performed on concentrated random vectors (and tested on Gaussian mixtures) to real data, we apply the multitask classification asymptotic to the handwritten digits of the MNIST dataset [14]. Figure 3 depicts, as a function of λ, the theoretical versus empirical classification errors for three different "transfer learning" settings (that is, for $k = 2$, using Task 1 as a support, or source, to the targeted Task 2).

It is interestingly observed that theoretical and practical graphs are again a very close match and that, while for very similar tasks with large numbers of data in the source task significant gains are achieved at large λ (thereby corroborating the discussion of Section 3.2), non-trivial gains are still reached on resembling tasks with a non-trivial λ, but almost no gain is obtained for too dissimilar tasks.

5. CONCLUDING REMARKS

The present work provides a first theoretical analysis of multitask learning, and indirectly of transfer learning, in the simplified setting of a least square support vector machine adaptation, and for a linear kernel ($X^{\mathsf{T}}X$) model. Already in this setting, the asymptotic algorithm performance appears to be non trivial, yet carries several non-trivial insights and opens the possibility of an on-line tuning of the hyperparameters of the model. These insights are key to understanding the (likely more involved) behavior of large dimensional data in refined modern formulations of multitask and transfer learning.

A natural extension of the present analysis notably includes the possibility to involve more elaborate data representations as well as kernel models of data similarity as developed in the asymptotic analysis of [15] and [13]. Similarly, the tools provided in [16] to analyze the performance of implicit (rather than explicit, as in the present work) solutions of optimization problems (standard support vector machines, logistic regression, neural network formulations) will bring the present analysis closer to modern considerations in multitask learning.

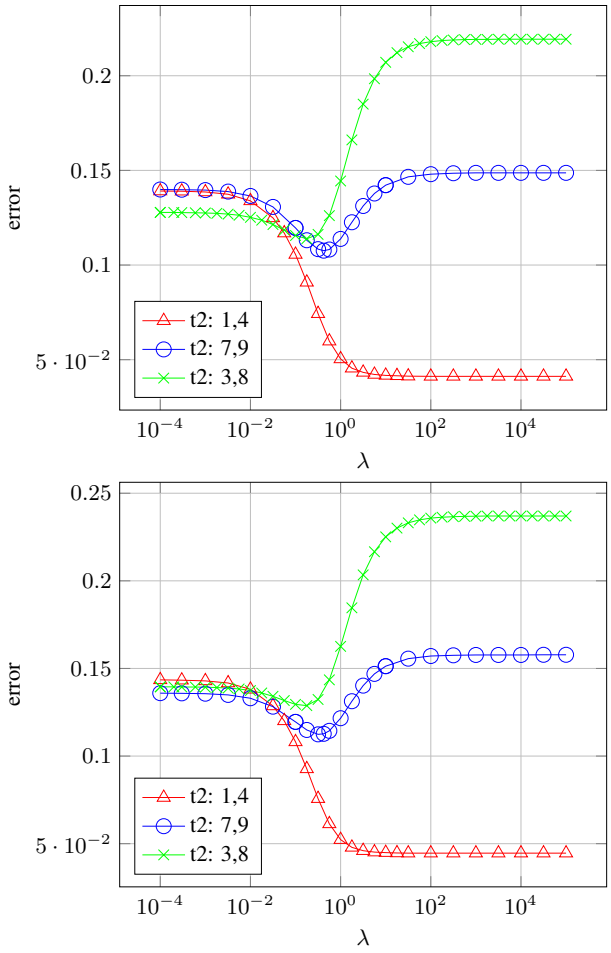


Fig. 3. Empirical **(Top)** versus theoretical **(Bottom)** classification error of Task 1 of a two-task problem for different λ. Training set of Task 1: noisy versions of digits 1 and 4 of MNIST dataset ($X_1^{(1)} = Z_1 + W_1$ and $X_1^{(2)} = Z_4 + W_2$ with Z_1 and Z_4 images of 1 and 4 from the MNIST database, columns of W_1, W_2 independent standard Gaussian), $n_{11} = n_{12} = 50$. Transfer learning performed onto three possible Task 2 (in order of dissemblance): digits $(1, 4)$, digits $(7, 9)$ and digits $(3, 8)$ with $n_{21} = n_{22} = 450$.

6. REFERENCES

[1] Ni Zhuang, Yan Yan, Si Chen, and Hanzi Wang, "Multi-task learning of cascaded cnn for facial attribute classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2069–2074.

[2] Runxing Zhong, Weifeng Lv, Bowen Du, Shuo Lei, and Runhe Huang, "Spatiotemporal multi-task learning for citywide passenger flow prediction," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2017, pp. 1–8.

[3] Wanrong Gu, Ziye Zhang, Xianfen Xie, and Yichen He, "An improved muti-task learning algorithm for analyzing cancer

survival data," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[4] Greg M Allenby and Peter E Rossi, "Marketing models of consumer heterogeneity," *Journal of econometrics*, vol. 89, no. 1-2, pp. 57–78, 1998.

[5] Neeraj Arora, Greg M Allenby, and James L Ginter, "A hierarchical bayes model of primary and secondary demand," *Marketing Science*, vol. 17, no. 1, pp. 29–44, 1998.

[6] Bart Bakker and Tom Heskes, "Task clustering and gating for bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, no. May, pp. 83–99, 2003.

[7] TM Heskes, "Empirical bayes for learning to learn," 2000.

[8] Theodoros Evgeniou and Massimiliano Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 109–117.

[9] Shuo Xu, Xin An, Xiaodong Qiao, Lijun Zhu, and Lin Li, "Multi-output least-squares support vector regression machines," *Pattern Recognition Letters*, vol. 34, pp. 1078–1084, 07 2013.

[10] Zhenyu Liao and Romain Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1065–1074, 2019.

[11] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3074–3100, 2018.

[12] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet, "Kernel random matrices of large concentrated data: the example of gan-generated images," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7480–7484.

[13] Cosme Louart and Romain Couillet, "Concentration of measure and large random matrices with an application to sample covariance matrices," *arXiv preprint arXiv:1805.08295*, 2018.

[14] Li Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[15] Romain Couillet, Florent Benaych-Georges, et al., "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[16] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet, "A large scale analysis of logistic regression: Asymptotic performance and new insights," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3357–3361.