SUPPLEMENTARY MATERIAL

Malik Tiomoko,
Laboratoire des Signaux et Systèmes Université Paris-Sud
Orsay, France
{malik.tiomoko}@u-psud.fr

Hafiz Tiomoko Ali
Huawei Technologies Research and Development (UK)
London, UK
{hafiz.tiomoko.ali@}@huawei.com

Romain Couillet[*]
Gipsa Lab
Université Grenoble-Alpes
Saint Martin d'Hères, France
{romain.couillet}@gipsa-lab.grenoble-inp.fr

ABSTRACT

This document contains the main technical arguments omitted in the core of the article due to space limitation and is organized as follows. Section 1 details the derivation of the solution of MTL LS-SVM optimization problem. Section 2 derives the asymptotic classification score of MTL LS-SVM. To this end, the necessary deterministic equivalents are given and proved in Lemma 1 which is further used to prove Theorem 1 of the main article (which in turn provides the asymptotic performance of MTL LS-SVM in the most general case of concentrated random vectors with arbitrary means and covariances). Practical remarks used in the main article as well as the algorithm for multi class extension are provided in Section 3. Supplementary experiments are provided in Section 4.

1 SOLUTION OF MTL LS SVM OPTIMIZATION

Let us recall the optimization problem, stated as:

$$\min_{(\omega_0,V,b)\in\mathbb{R}^p\times\mathbb{R}^{p\times k}\times\mathbb{R}^k} \mathcal{J}(\omega_0,V,b) \tag{1}$$

where

$$\mathcal{J}(\omega_0,V,b)\equiv\frac{1}{2\lambda}\|\omega_0\|^2+\frac{1}{2}\sum_{i=1}^k\frac{\|v_i\|^2}{\gamma_i}+\frac{1}{2}\sum_{i=1}^k\|\xi_i\|^2$$
$$\xi_i=y_i-(\mathring{X}_i^\mathsf{T}\omega_i+b_i\mathbb{1}_{n_i}),\quad\forall i\in\{1,\ldots,k\}.$$

The Lagrangian of the constrained optimization problem using the relatedness assumption ($\omega_i=\omega_0+v_i$) reads:

$$\mathcal{L}(\omega_0,v_i,\xi_i,\alpha_i,b_i)=\frac{1}{2\lambda}\|\omega_0\|^2+\frac{1}{2}\sum_{i=1}^k\frac{\|v_i\|^2}{\gamma_i}+\frac{1}{2}\sum_{i=1}^k\|\xi_i\|^2+\sum_{i=1}^k\alpha_i^\mathsf{T}\left(y_i-\mathring{X}_i^\mathsf{T}\omega_0-\mathring{X}_i^\mathsf{T}v_i-b_i\mathbb{1}_{n_i}-\xi_i\right)$$

with $\alpha_i\in\mathbb{R}^{n_i}$ the Lagrangian parameter attached to task i.

---

[*] Couillet's work is partially supported by MIAI at University Grenoble-Alpes (ANR-19-P3IA-0003) and the HUAWEI LarDist project.

Differentiating with respect to the unknowns ω_0, v_i, ξ_i, α_i, and b_i leads to the following system of equations:

$$\begin{aligned}
\frac{1}{\lambda}\omega_0 - \sum_{i=1}^{k} \mathring{X}_i \alpha_i &= 0 && (2)\\
\frac{1}{\gamma_i} v_i - \mathring{X}_i \alpha_i &= 0 && (3)\\
\xi_i - \alpha_i &= 0 && (4)\\
y_i - \mathring{X}_i^{\mathsf{T}} \omega_0 - \mathring{X}_i^{\mathsf{T}} v_i - b_i \mathbb{1}_{n_i} - \xi_i &= 0 && (5)\\
\alpha_i^{\mathsf{T}} \mathbb{1}_{n_i} &= 0. && (6)
\end{aligned}$$

Plugging the expression of ω_0 (equation 2), v_i (equation 3) and ξ_i (equation 4) into equation 5 leads to:

$$\begin{aligned}
y_i &= (\lambda + \gamma_i)\, X_i^{\mathsf{T}} X_i \alpha_i + \lambda \sum_{j \neq i} X_i^{\mathsf{T}} X_j \alpha_j + b_i \mathbb{1}_{n_i} + \alpha_i\\
\mathbb{1}_{n_i}^{\mathsf{T}} \alpha_i &= 0.
\end{aligned}$$

With $y = [y_1^{\mathsf{T}}, \ldots, y_k^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^n$, $\alpha = [\alpha_1^{\mathsf{T}}, \ldots, \alpha_k^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^n$, $Z = \sum_{i=1}^{k} e_i^{[k]} e_i^{[k]^{\mathsf{T}}} \otimes \mathring{X}_i \in \mathbb{R}^{kp \times n}$ and $P \in \mathbb{R}^{n \times k}$ such that the j-th column is $P_{\cdot j} = [\mathbf{0}_{n_1+\ldots+n_{j-1}}^{\mathsf{T}}, \mathbb{1}_{n_j}^{\mathsf{T}}, \mathbf{0}_{n_{j+1}+\ldots+n_k}^{\mathsf{T}}]^{\mathsf{T}}$, this system of equations can be written under the compact form:

$$\begin{aligned}
Pb + Q^{-1}\alpha &= y\\
P^{\mathsf{T}}\alpha &= \mathbf{0}_k
\end{aligned}$$

with $Q = \left(\frac{Z^{\mathsf{T}} A Z}{kp} + I_n\right)^{-1} \in \mathbb{R}^{n \times n}$, and $A = \left(\mathcal{D}_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k^{\mathsf{T}}\right) \otimes I_p \in \mathbb{R}^{kp \times kp}$.

Solving for α and b then gives:

$$\begin{aligned}
\alpha &= Q(y - Pb)\\
b &= (P^{\mathsf{T}} Q P)^{-1} P^{\mathsf{T}} Q y.
\end{aligned}$$

Moreover, using $\omega_i = \omega_0 + v_i$, equation 2 and equation 3, the expression of ω_i becomes:

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

With this formulation for the solution pair (ω_i, b), the prediction of the class of any new data point $\mathbf{x} \in \mathbb{R}^p$ for Task i is then obtained from the classification score $g_i(\mathbf{x})$ given by

$$g_i(\mathbf{x}) = \frac{1}{kp}\left(e_i^{[k]} \otimes \mathring{\mathbf{x}}_i\right)^{\mathsf{T}} \omega_i + b_i = \frac{1}{kp}\left(e_i^{[k]} \otimes \mathring{\mathbf{x}}_i\right)^{\mathsf{T}} A Z \alpha + b_i \tag{7}$$

where $\mathring{\mathbf{x}} = (\mathbf{x} - \frac{1}{n_i} X_i \mathbb{1}_{n_i})$ is a centered version of $\mathbf{x}$ with respect to the training dataset for Task i and $\left(e_i^{[k]} \otimes \mathring{\mathbf{x}}_i\right)^{\mathsf{T}}$ specifies that the classification score of $\mathbf{x}$ is computed for Task i.

2 ASYMPTOTIC CLASSIFICATION SCORE STATISTICS

2.1 PRELIMINARIES

Using the definition of α, $g_i(x)$ can be expanded as

$$g_i(\mathbf{x}) = \frac{1}{kp}\left(e_i^{[k]} \otimes \mathbf{x}\right)^{\mathsf{T}} A Z Q (y - Pb) + b_i.$$

By the identity $(I + DB)^{-1} D = D (I + BD)^{-1}$ for matrices D and B, one can further conveniently write $g_i(\mathbf{x})$ as:

$$g_i(\mathbf{x}) = \frac{1}{kp} \left(e_i^{[k]} \otimes \mathbf{x}\right)^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z(y - Pb) + b_i \tag{8}$$

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

The mean m_{ij} and the variance σ_{ij} of $g_i(\mathbf{x})$ for any data vector $\mathbf{x}$ such that $\mathbb{E}[\mathbf{x}] = \mu_{ij}$ and $\mathrm{Cov}[\mathbf{x}] = \Sigma_{ij}$ are respectively given by:

$$\begin{aligned} m_{ij} &= \mathbb{E}\left[\frac{1}{kp} \left(e_i^{[k]} \otimes \mu_{ij}\right)^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z(y - Pb) + b_i\right] \\ \sigma_{ij}^2 &= \mathbb{E}\left[\frac{1}{(kp)^2} (y - Pb)^\mathsf{T} Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z(y - Pb)\right]. \end{aligned}$$

with $S_{ij} = e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes \Sigma_{ij}$.

To compute the statistics of $g_i(\mathbf{x})$, 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 m_{ij} and σ_{ij}^2.

Definition 1 (Deterministic equivalents)**.** *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 the statistics of $g_i(\mathbf{x})$ are bilinear forms involving $\tilde{Q} A^{\frac{1}{2}} Z$ and $Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z$, Lemma 1 will provide the deterministic equivalent of the first and second order statistics.

To this end, we first specify in the following assumptions the distribution of the data matrix X and the test data $\mathbf{x}$ and the growth rate of n and p.

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 $\mathrm{Cov}[x_{ij}] = \Sigma_{ij}$, which only depend on (i, j).

Assumption 2 (Distribution of $\mathbf{x}$)**.** There exist two constants $C, c > 0$ (independent of n, p) such that, for any 1-Lipschitz function $f : \mathbb{R}^p \to \mathbb{R}$,

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

and $\mathbf{x}$ is independent from the columns of data matrix X.

Assumptions 1 and 2 are more general than the Gaussian assumption considered in the core article. This wide class of random vectors known as concentrated random vectors (see Ledoux (2001) for details) 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 of radius $\sqrt{p}$, 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, issued from generative models, as was recently demonstrated in Seddik et al. (2019) with the specific example of concentrated random vectors arising from generative adversarial networks (GANs).

Assumption 3 (Growth Rate)**.** As $n \to \infty$, $n/p \to c_0 > 0$ and, for $1 \leq i \leq k$, $1 \leq j \leq m$, $\frac{n_{ij}}{n} \to c_{ij} > 0$. We further denote $c_i = c_{i1} + c_{i2}$ and $c = [c_1, \ldots, c_k]^\mathsf{T} \in \mathbb{R}^k$. Besides, for each i, letting $\Delta\mu_i \equiv \mu_{i1} - \mu_{i2}$.

Under the two aforementioned assumptions, deterministic equivalents needed to compute the statistics of $g_i(\mathbf{x})$ are derived in the following lemma.

2.2 DETERMINISTIC EQUIVALENTS

Lemma 1 (Deterministic equivalents)**.** *Define, for class j in Task i, the data deterministic 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

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

and of second order

$$\begin{aligned} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} &\leftrightarrow B_{ij} \\ Z^{\mathsf{T}} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z &\leftrightarrow J M_{\Delta}^{\mathsf{T}} A^{\frac{1}{2}} (B_{ij} A^{\frac{1}{2}} M_{\Delta} J^{\mathsf{T}} - \bar{\tilde{Q}} A^{\frac{1}{2}} M_{\Delta} W_{ij}) + E_{ij} \end{aligned}$$

in which we defined

$$\begin{aligned} W_{ij} &= [w_{11}, \ldots, w_{k2}]^{\mathsf{T}}, \quad w_{sl} = \left[\boldsymbol{0}^{\mathsf{T}}_{n_{11}+\ldots+n_{(s-1)l}}, \frac{2\mathrm{tr}\,(B_{ij} C_{sl})}{kp(1+\Delta_{sl})} \mathbb{1}^{\mathsf{T}}_{n_{sl}}, \boldsymbol{0}^{\mathsf{T}}_{n_{(s+1)l}+\ldots+n_{k2}}\right]^{\mathsf{T}} \\ E_{ij} &= \sum_{l,m} \frac{\mathrm{tr}(C_{lm} B_{ij})}{(1+\Delta_{lm})^2} e_{lm}^{[2k]} e_{lm}^{[2k]\mathsf{T}} \\ B_{ij} &= \bar{\tilde{Q}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \bar{\tilde{Q}} + \sum_{a=1}^{k} \sum_{b=1}^{2} d_{ab} T_{ab}^{(ij)} [\bar{\tilde{Q}} C_{ab} \bar{\tilde{Q}}] \\ D &= \sum_{i,j} d_{ij} e_{ij}^{[2k]} e_{ij}^{[2k]\mathsf{T}}, \; d_{ij} = \frac{n_{ij}}{kp(1+\Delta_{ij})^2} \\ J &= [j_{11}, \ldots, j_{k2}], \\ j_{lm} &= \left(0^{\mathsf{T}}_{n_{11}+\ldots+n_{(i-1)2}}, \mathbb{1}^{\mathsf{T}}_{n_{ij}}, 0^{\mathsf{T}}_{n_{(i+1)1}+\ldots+n_{k2}}\right)^{\mathsf{T}}, \\ M_{\Delta} &= M \sum_{ij} \frac{1}{1+\Delta_{ij}} e_{ij}^{[2k]} e_{ij}^{[2k]\mathsf{T}} \\ S_{ij} &= e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes \Sigma_{ij} \\ T &= \bar{T}(I_k - D\mathcal{C})^{-1}, \; \mathcal{C}_{(il)}^{(jm)} = \frac{1}{kp} \mathrm{tr}(C_{ij} \bar{\tilde{Q}} C_{lm} \bar{\tilde{Q}}), \bar{T}_{ab}^{(ij)} = \frac{1}{kp} tr \left(C_{ab} \bar{\tilde{Q}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \bar{\tilde{Q}} \right) \end{aligned}$$

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

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

2.3 PROOF OF LEMMA 1

First order deterministic equivalent A deterministic equivalent for $\tilde{Q}$ is provided in Louart & Couillet (2018). Our objective is then to find, based on this result, a deterministic equivalent for the random matrix $A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z$. To this end, consistently with Definition 1, we will evaluate the scalar quantity $\mathbb{E}\left[u^{\mathsf{T}} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z v\right]$ for any deterministic vector $u \in \mathbb{R}^{kp}$ and $v \in \mathbb{R}^{n}$ such that $\|u\| = 1$ and $\|v\| = 1$, which we can write

$$\mathbb{E}\left[u^{\mathsf{T}} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z v\right] = \sum_{i=1}^{n} v_i \mathbb{E}\left[u^{\mathsf{T}} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} z_i\right]. \tag{9}$$

Furthermore, let us define for convenience the matrix Z_{-i}, which is the matrix Z with a vector of zeros on its i-th column such that $ZZ^\intercal = Z_{-i}Z_{-i}^\intercal + z_i z_i^\intercal$. Using the Sherman-Morrison matrix inversion lemma (i.e., $\left(A + uv^\intercal\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\intercal A^{-1}}{1+v^\intercal A^{-1}u}$), we find:

$$\tilde{Q} = \left(\frac{A^{\frac{1}{2}}ZZ^\intercal A^{\frac{1}{2}}}{kp} + I_{kp}\right)^{-1} = \tilde{Q}_{-i} - \frac{1}{kp}\frac{\tilde{Q}_{-i}A^{\frac{1}{2}}z_i z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}}{1 + \frac{1}{kp}z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i} \tag{10}$$

with $\tilde{Q}_{-i} = (\frac{A^{\frac{1}{2}}Z_{-i}Z_{-i}^\intercal A^{\frac{1}{2}}}{kp} + I_{kp})^{-1}$. Furthermore,

$$\tilde{Q}A^{\frac{1}{2}}z_i = \frac{\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1 + \frac{1}{kp}z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}. \tag{11}$$

Plugging equation 11 into equation 9 leads to

$$\mathbb{E}\left[u^\intercal A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv\right] = \sum_{i=1}^{n} v_i \mathbb{E}\left[u^\intercal \frac{A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1 + \frac{1}{kp}z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}\right]. \tag{12}$$

Moreover, following the same line of reasoning as in (Seddik et al., 2020, Proposition A.3), based on Assumption 1 and tools from concentration of measure theory (see also (Ledoux, 2001; Louart et al., 2018)), one can show that:

$$\sum_{i=1}^{n} v_i \mathbb{E}\left[u^\intercal \frac{A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1 + \frac{1}{kp}z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}\right] = \sum_{i=1}^{n} v_i \mathbb{E}\left[u^\intercal \frac{A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1 + \Delta_{ij}}\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \tag{13}$$

with $\Delta_{ij} \equiv \mathbb{E}\left[\frac{1}{kp}z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i\right]$. Note that Δ_{ij} can be estimated in the large n, p limit as the solution of the fixed point equation (with $\mathcal{O}(\cdot)$ part discarded)

$$\Delta_{ij} = \frac{1}{kp}\mathbb{E}\left[\operatorname{tr}\left(A^{\frac{1}{2}}z_i z_i^\intercal A^{\frac{1}{2}}\tilde{Q}_{-i}\right)\right] = \frac{1}{kp}\operatorname{tr}\left(C_{ij}\bar{\tilde{Q}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$$

where we used the fact that z_i is independent from $\tilde{Q}_{-i}$.

We then conclude that:

$$\mathbb{E}\left[u^\intercal A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv\right] = \sum_{i=1}^{n} v_i u^\intercal \frac{\mathbb{E}\left[A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i\right]}{1 + \Delta_{ij}} + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) = u^\intercal A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\Delta v + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)$$

where in the last equality, we again used the fact that $\tilde{Q}_{-i}$ is independent from z_i. This concludes the proof.

Second order deterministic equivalent We aim in the following section to prove that $Z^\intercal A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z \leftrightarrow JM_\Delta^\intercal A^{\frac{1}{2}}(B_{ij}A^{\frac{1}{2}}M_\Delta J^\intercal - \bar{\tilde{Q}}A^{\frac{1}{2}}M_\Delta W_{ij}) + E_{ij}$.

Let us define for convenience $\mathcal{C}(i)$ the class of the i-th sample (i.e, $\mathcal{C}(i) \in \{1, \ldots, 2k\}$).

Similarly as done for the first order deterministic equivalents, the focus will be on $\mathbb{E}[u^\intercal Z^\intercal A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv]$.

Using successively equation 10 and equation 13 on i and j , we have for $i \neq j$,

$$
\begin{aligned}
&\sum_{\substack{i,j=1\\ i\neq j}}^{n} u_i v_i \mathbb{E}\left[z_i^\top A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}z_j\right] \\
&= \sum_{\substack{i,j=1\\ i\neq j}}^{n} u_i v_i \mathbb{E}\left[\frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-j}A^{\frac{1}{2}}z_j}{(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})}\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \\
&= \sum_{\substack{i,j=1\\ i\neq j}}^{n} u_i v_i \mathbb{E}\left[\frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})} - \frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_i z_i^\top A^{\frac{1}{2}}\tilde{Q}_{-j}A^{\frac{1}{2}}z_j}{kp(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})}\right. \\
&\left. - \frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}z_j z_j^\top A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-j}A^{\frac{1}{2}}z_j}{kp(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})}\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \\
&= \sum_{\substack{i,j=1\\ i\neq j}}^{n} u_i v_i \mathbb{E}\left[\frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})} - \frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_i z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{kp(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})(1+\Delta_{\mathcal{C}(i)})}\right. \\
&- \frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}z_j z_j^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{kp(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})(1+\Delta_{\mathcal{C}(j)})} \\
&\left. + \frac{1}{(kp)^2}\frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}z_j z_j^\top A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_i z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})(1+\Delta_{\mathcal{C}(i)})}\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \\
&= \sum_{\substack{i,j=1\\ i\neq j}}^{n} u_i v_i \mathbb{E}\left[\frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})} - \frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_i z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{kp(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})(1+\Delta_{\mathcal{C}(i)})}\right. \\
&\left. - \frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}z_j z_j^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{kp(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})(1+\Delta_{\mathcal{C}(j)})}\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right).
\end{aligned}
$$

where the term $\frac{1}{(kp)^2}\frac{z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-i\\-j}}A^{\frac{1}{2}}z_j z_j^\top A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_i z_i^\top A^{\frac{1}{2}}\tilde{Q}_{\substack{-j\\-i}}A^{\frac{1}{2}}z_j}{(1+\Delta_{\mathcal{C}(i)})(1+\Delta_{\mathcal{C}(j)})(1+\Delta_{\mathcal{C}(i)})}$ is proved to be order $\mathcal{O}(\frac{1}{\sqrt{p}})$ using (Seddik et al., 2020, Lemma A.2).

The deterministic equivalent for the off-diagonal entries of $Z^\top A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z$ is then :

$$Eq = JM_\Delta^\top A^{\frac{1}{2}}B_{ij}A^{\frac{1}{2}}M_\Delta J^\top - \left(JM_\Delta^\top A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\Delta W_{ij}\right)$$

with $A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}} \leftrightarrow B_{ij}$ and $W_{ij} = [w_{11}, \ldots, w_{k2}]^\top$, $w_{sl} = \left[\mathbf{0}^\top_{n_{11}+\ldots+n_{(s-1)l}}, \frac{2\operatorname{tr}(B_{ij}C_{sl})}{kp(1+\Delta_{sl})}\mathbb{1}^\top_{n_{sl}}, \mathbf{0}^\top_{n_{(s+1)l}+\ldots+n_{k2}}\right]^\top$

Similarly as for the non diagonal element, the diagonal element is proved to be

$$E_{ij} = \sum_{l,m}\frac{\operatorname{tr}(C_{lm}B_{ij})}{(1+\Delta_{lm})^2}e_{lm}^{[2k]}e_{lm}^{[2k]^\top}.$$

Put together, the complete deterministic equivalent is then:

$$JM_\Delta^\top A^{\frac{1}{2}}B_{ij}A^{\frac{1}{2}}M_\Delta J^\top - JM_\Delta^\top A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\Delta W_{ij} + \sum_{l,m}\frac{\operatorname{tr}(C_{lm}B_{ij})}{(1+\Delta_{lm})^2}e_{lm}^{[2k]}e_{lm}^{[2k]^\top}.$$

This proves that $Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z \leftrightarrow J M_\Delta^\mathsf{T} A^{\frac{1}{2}} (B_{ij} A^{\frac{1}{2}} M_\Delta J^\mathsf{T} - \bar{\tilde{Q}} A^{\frac{1}{2}} M_\Delta W_{ij}) + E_{ij}$. It then remains to retrieve the deterministic equivalent B_{ij} of $\tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}$.

Calculus of B_{ij} Similar derivations and results are provided in detail in Louart et al. (2018). For conciseness, we sketch the most important elements of the proof. The interested reader can refer to (Louart et al., 2018, Section 5.2.3). Let us evaluate $\mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} (\tilde{Q} - \bar{\tilde{Q}}) v\right]$ for any deterministic vector $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$ such that $\|u\| = 1$ and $\|v\| = 1$ by using successively equation 13 and equation 10,

$$\begin{aligned}
\mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} (\tilde{Q} - \bar{\tilde{Q}}) v\right] &= \mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} \Big(-\frac{A^{\frac{1}{2}} Z Z^\mathsf{T} A^{\frac{1}{2}}}{kp} + C_\Delta\Big) \bar{\tilde{Q}} v\right] \\
&= -\frac{1}{kp} \sum_i \mathbb{E}\left[\frac{u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} A^{\frac{1}{2}} z_i z_i^\mathsf{T} A^{\frac{1}{2}} \bar{\tilde{Q}} v}{1 + \Delta_{ij}}\right] + \mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} C_\Delta \bar{\tilde{Q}} v\right] \\
&\quad - \frac{1}{kp} \mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} A^{\frac{1}{2}} z_i z_i^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} C_\Delta \bar{\tilde{Q}} v\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right).
\end{aligned}$$

Using Assumption 1 and following the work of Louart & Couillet (2018), $\frac{1}{kp}\mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} A^{\frac{1}{2}} z_i z_i^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} C_\Delta \bar{\tilde{Q}} v\right] = \mathcal{O}(\frac{1}{p})$ and where $C_\Delta = \sum\limits_{ij} \frac{c_{ij}}{c_0} \frac{C_{ij}}{1+\Delta_{ij}}$. Furthermore,

$$\begin{aligned}
\mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} (\tilde{Q} - \bar{\tilde{Q}}) v\right] &= -\frac{1}{kp} \sum_i \mathbb{E}\left[\frac{u^\mathsf{T} \tilde{Q}_{-i} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} A^{\frac{1}{2}} z_i z_i^\mathsf{T} A^{\frac{1}{2}} \bar{\tilde{Q}} v}{1 + \Delta_{ij}}\right] \\
&\quad + \frac{1}{kp} \sum_i \mathbb{E}\left[\frac{u^\mathsf{T} \tilde{Q}_{-i} A^{\frac{1}{2}} z_i z_i^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} A^{\frac{1}{2}} z_i z_i^\mathsf{T} A^{\frac{1}{2}} \bar{\tilde{Q}} v}{kp(1 + \Delta_{ij})^2}\right] \\
&\quad + \mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} Q_{-i} C_\Delta \bar{\tilde{Q}} v\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \\
&= \frac{1}{kp} \sum_i \mathbb{E} \frac{\operatorname{tr}\left(C_{\mathcal{C}(i)} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}\right)}{(1 + \Delta_{\mathcal{C}(i)})^2} \mathbb{E}\left[u^\mathsf{T} \bar{\tilde{Q}} C_{\mathcal{C}(i)} \bar{\tilde{Q}} v\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)
\end{aligned}$$

where $-\frac{1}{kp} \sum\limits_i \mathbb{E}\left[\frac{u^\mathsf{T} \tilde{Q}_{-i} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-i} z_i z_i^\mathsf{T} \bar{\tilde{Q}} v}{1+\Delta_{ij}}\right] + \mathbb{E}\left[u^\mathsf{T} \tilde{Q}_{-i} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} Q C_\Delta \bar{\tilde{Q}} v\right] = \mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$, following again Louart & Couillet (2018).

Let us then denote $d_{ab} = \frac{n_{ab}}{kp(1+\Delta_{ab})^2}$. We have the following identity involving $\mathbb{E}\left[\tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}\right]$

$$\mathbb{E}[\tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}] = \bar{\tilde{Q}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \bar{\tilde{Q}} + \sum_{a=1}^{k} \sum_{b=1}^{2} \frac{d_{ab}}{kp} \mathbb{E}\left[\operatorname{tr}\left(C_{ab} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}\right)\right] \bar{\tilde{Q}} C_{ab} \bar{\tilde{Q}} + \mathcal{O}_{\|\cdot\|}\Big(\frac{\sqrt{\log p}}{\sqrt{p}}\Big), \tag{14}$$

where $A = B + \mathcal{O}_{\|\cdot\|}(\alpha(p))$ means that $\|A - B\| = \mathcal{O}(\alpha(p))$, and where the norm $\|\cdot\|$ is understood as the euclidean norm for vectors and the operator norm for matrices.

Further introduce two matrices $\bar{T}$ and T defined as: $\bar{T}_{ab}^{(ij)} = \frac{1}{kp} tr\left(C_{ab} \bar{\tilde{Q}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \bar{\tilde{Q}}\right)$ and $T_{ab}^{(ij)} = \frac{1}{kp}\mathbb{E}\left[\operatorname{tr}\left(C_{ab} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}\right)\right]$. These satisfy the following equations (i.e., by right multiplying equation 14 by C_{ab} and taking the trace)

$$T_{ab}^{(ij)} = \bar{T}_{ab}^{(ij)} + \sum_{e=1}^{k} \sum_{f=1}^{2} d_{ef} T_{ef}^{(ij)} \mathcal{C}_{ef}^{(ab)},$$

so that $T = \bar{T}(I_k - D\mathcal{C})^{-1}$ by denoting $D = \mathcal{D}_{[d_{11},\ldots,d_{k2}]^\mathsf{T}}$ and $\mathcal{C}_{ef}^{(ab)} = \frac{1}{kp} \operatorname{tr}\left(C_{ef} \bar{\tilde{Q}} C_{ab} \bar{\tilde{Q}}\right)$.

Finally,

$$\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q} \leftrightarrow \bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}} + \sum_{a=1}^{k}\sum_{b=1}^{2} d_{ab}T_{ab}^{(ij)}\mathbb{E}[\bar{\tilde{Q}}C_{ab}\bar{\tilde{Q}}]$$

with $T = \bar{T}(I_k - D\mathcal{C})^{-1}$.

2.4 CLASSIFICATION SCORE ASYMPTOTICS

Theorem 2. *Under Assumptions 1–3 and the notations of Lemma 1,*

$$E[g_i(x)] - m_{ij} \to 0, \quad Var[g_i(x)] - \sigma_{ij}^2 \to 0$$

where, letting $m = [m_{11}, \ldots, m_{k2}]^{\mathsf{T}}$,

$$m = \tilde{y} - \mathcal{D}_{\tilde{\Delta}}^{-\frac{1}{2}}\Gamma\mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\mathring{\tilde{y}}$$

$$\sigma_{ij}^2 = \frac{1}{(kp)^2}\tilde{y}^{\mathsf{T}}\mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\left(\Gamma\mathcal{D}_{\kappa}\Gamma + \Gamma\mathcal{M}^{\mathsf{T}}\bar{\tilde{Q}}_0\mathbb{V}_{ij}\bar{\tilde{Q}}_0\mathcal{M}\Gamma\right)\mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\tilde{y}$$

with $\Gamma = \left(I_{2k} + \mathbb{M}^{\mathsf{T}}\bar{\tilde{Q}}_0\mathbb{M}\right)^{-1}$, $\bar{\tilde{Q}}_0 = \left[\sum_{i=1}^{k}\left(\mathcal{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k\right)^{\frac{1}{2}} e_i e_i^{\mathsf{T}}\left(\mathcal{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k\right)^{\frac{1}{2}} \otimes \left(\tilde{\Delta}_{i1}\Sigma_{i1} + \tilde{\Delta}_{i2}\Sigma_{i2}\right) + I_{kp}\right]^{-1}$, $\mathbb{V}_{ij} = A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}} + \sum_{a=1}^{k}\sum_{b=1}^{2}\kappa_{ab}A^{\frac{1}{2}}S_{ab}A^{\frac{1}{2}}$, $\mathbb{M} = [\mathbb{M}_{11}, \ldots, \mathbb{M}_{k2}]$ *with* $\mathbb{M}_{ij} = \left(\mathcal{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k\right)^{\frac{1}{2}} e_i \otimes \sqrt{\tilde{\Delta}_{ij}}\mu_{ij}$, $\tilde{\Delta}_{ij} = \frac{c_{ij}}{c_0(1+\Delta_{ij})}$ *and* $\kappa_{ab}^{(ij)} = d_{ab}T_{ab}^{(ij)}$.

Besides for independent Gaussian random variables $[X, x]$,

$$g_i(\mathbf{x}) - G_{ij} \to 0, \quad G_{ij} \sim \mathcal{N}(m_{ij}, \sigma_{ij}^2)$$

in law.

2.4.1 PROOF OF THEOREM 2

Proof of the convergence in distribution The convergence in distribution of the statistics of the classification score $g_i(\mathbf{x})$ is identical to the CLT derived in Appendix B of Liao & Couillet (2019) by writing the classification score $g_i(\mathbf{x})$ in polynomial form of a Gaussian vector and resorting to the Lyapounov CLT Billingsley (2008) under a Gaussian vector assumption.

Mean of the classification score Using the definition of the classification score, the mean is

$$m_{ij} = \mathbb{E}\left[\frac{1}{kp}\left(e_i^{[k]} \otimes \mu_{ij}\right)^{\mathsf{T}} A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z(y - Pb)\right] + b_i$$

Using Lemma 1, the mean then reads:

$$m_{ij} = \frac{1}{kp}\left(e_i^{[k]} \otimes \mu_{ij}\right)^{\mathsf{T}} A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_{\Delta}J^{\mathsf{T}}(y - P\bar{b}) + b_i \tag{15}$$

Since $C_{ij} = A^{\frac{1}{2}}\left(e_i^{[k]}e_i^{[k]\mathsf{T}} \otimes (\Sigma_{ij} + \mu_{ij}\mu_{ij}^{\mathsf{T}})\right)A^{\frac{1}{2}}$ is a rank-one update of Σ_{ij}, one can further use Woodbury's matrix identity (i.e $(A + UCV)^{-1} = A^{-1} + A^{-1}UC(I + VCU)VA^{-1}$ for matrices A, U, C, and V) to write $\bar{\tilde{Q}}$ as: $\bar{\tilde{Q}} = \bar{\tilde{Q}}_0 - \bar{\tilde{Q}}_0\mathbb{M}\left(I_{kp} + \mathbb{M}^{\mathsf{T}}\bar{\tilde{Q}}_0\mathbb{M}\right)^{-1}\mathbb{M}^{\mathsf{T}}\bar{\tilde{Q}}_0$, with $\bar{\tilde{Q}}_0 = \left[\sum_{i=1}^{k}\left(\mathcal{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k\right)^{\frac{1}{2}} e_i e_i^{\mathsf{T}}\left(\mathcal{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k\right)^{\frac{1}{2}} \otimes \left(\tilde{\Delta}_{i1}\Sigma_{i1} + \tilde{\Delta}_{i2}\Sigma_{i2}\right) + I_{kp}\right]^{-1}$, $\mathbb{M} = [\mathbb{M}_{11}, \ldots, \mathbb{M}_{k2}]$ with $\mathbb{M}_{ij} = \left(\mathcal{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k\right)^{\frac{1}{2}} e_i \otimes \sqrt{\tilde{\Delta}_{ij}}\mu_{ij}$ and $\tilde{\Delta}_{ij} = \frac{c_{ij}}{c_0(1+\Delta_{ij})}$

Plugging the expression of $\bar{\tilde{Q}}$ in equation 15, the mean reads:

$$m_{ij} = v^{\mathsf{T}}(I_{2k} - \Gamma)\frac{e_{ij}}{\sqrt{\tilde{\Delta}_{ij}}} + b_i$$

with $v = \mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\tilde{y}$, $\Gamma = \left(I_{2k} + \mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M}\right)^{-1}$ and e_{ij} is the canonical vector. Finally, let us conclude by remarking that one can show using deterministic equivalent for Q provided in Louart & Couillet (2018) that $b_i = \frac{\mathbb{1}_{n_i}^\mathsf{T} y_i}{n_i} + \mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$.

Letting $m = [m_{11}, \ldots, m_{k2}]^\mathsf{T}$, one can further show using notations of the main paper:

$$m = \tilde{y} - \mathcal{D}_{\tilde{\Delta}}^{-\frac{1}{2}}\Gamma\mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\mathring{\tilde{y}}.$$

In the particular case treated in the main article where $\Sigma_{ij} = I_p$, one can further simplify $\mathbb{M}^\mathsf{T}\tilde{Q}_0\mathbb{M}$ using elementary properties of the Kronecker matrix product (in particular, using $(A \otimes B)(C \otimes D) = AC \otimes BD$) so that $\mathbb{M}^\mathsf{T}\tilde{Q}_0\mathbb{M} = \left(\mathcal{A} \otimes \mathbb{1}_2^\mathsf{T}\mathbb{1}_2\right) \odot \mathcal{M}$ with the notations of the main paper. Thus, elementary algebraic manipulations provide the result of the mean of the main article. In the case of generic Σ_{ij}'s, the expression of $\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M}$ is somewhat less trivial and cannot further be simplified.

Variance of the classification score Using the expression of the classification score, the variance of the score is given by

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

Using Lemma 1, the expression further gives

$$\sigma_{ij}^2 = \frac{1}{(kp)^2}(y - P\bar{b})^\mathsf{T}\left(JM_\Delta^\mathsf{T}A^{\frac{1}{2}}B_{ij}A^{\frac{1}{2}}M_\Delta J + E_{ij}\right)(y - P\bar{b}) - \frac{1}{p^2}(y - P\bar{b})^\mathsf{T}JM_\Delta^\mathsf{T}A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\Delta W_{ij}(y - P\bar{b}).$$

Similarly as done for the mean, using again $\bar{\tilde{Q}} = \bar{\tilde{Q}}_0 - \bar{\tilde{Q}}_0\mathbb{M}\left(I_{kp} + \mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M}\right)^{-1}\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0$, one can show that the variance reads:

$$\sigma_{ij}^2 = \frac{1}{\tilde{\Delta}_i}\tilde{y}^\mathsf{T}\mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\left(\Gamma\mathcal{D}_\kappa\Gamma + \Gamma\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{V}_{ij}\bar{\tilde{Q}}_0\mathbb{M}\Gamma\right)\mathcal{D}_{\tilde{\Delta}}^{\frac{1}{2}}\tilde{y}$$

with $\mathbb{V}_{ij} = A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}} + \sum_{a=1}^{k}\sum_{b=1}^{2}\kappa_{ab}^{(ij)}A^{\frac{1}{2}}S_{ab}A^{\frac{1}{2}}$ and $\kappa_{ab}^{(ij)} = d_{ab}T_{ab}^{(ij)}$.

In the particular case of $\Sigma_{ij} = I_p$, one can show using again the Kronecker product formulas that $\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{V}_{ij}\bar{\tilde{Q}}_0\mathbb{M} = \frac{1}{c_0}\left(\mathcal{A}\mathcal{D}_{\frac{c_0\mathcal{K}_{i\cdot}}{c} + e_i^{[k]}}\mathcal{A} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}\right) \odot M$ with the notations of the main paper; this then leads to the result of the variance in the core of the paper.

3 PRACTICAL REMARKS AND ALGORITHM

In this section, the remark about the shift invariance of the scores is presented. Moreover Matlab and Julia codes are available in the supplementary files provided.

Remark 1 (Shift invariance of the scores). *If the score vectors $y_i \in \mathbb{R}^{n_i}$ are shifted by some constant vector $P\bar{y}$ for some matrix $\bar{y} \in \mathbb{R}^k$ i.e., if all data of the same task are affected by the same shift of their scores (or labels), then we find that the Lagrangian parameter α^{shift} after the shift is $\alpha^{\text{shift}} = Q\left(I_n - P(P^\mathsf{T}QP)^{-1}P^\mathsf{T}Q\right)(y + P\bar{y}) = \alpha$. As such, the normal vectors $\omega_i = (e_i^{[k]\mathsf{T}} \otimes I_p)AZ\alpha$, and as a consequence the performance of MTL LS-SVM, are insensitive to a constant shift in all the scores of each task.*

4 SUPPLEMENTARY EXPERIMENTS

4.1 BINARY DECISION

We here apply the results of MTL LS-SVM to a hypothesis test on a *target* task t based on training samples from both a source task s and the target task t. That is, instead of relying on the "averaged-mean" decision procedure (i.e $g_i(\mathbf{x}) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \frac{1}{2}(m_{i1} + m_{i2})$) , we instead consider the test

$$g_t(\mathbf{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \zeta$$

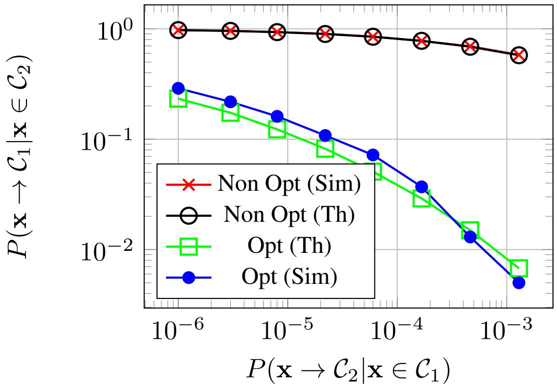

Figure 1: ROC curve for proposed versus non-optimized MTL LS-SVM on synthetic data with $p = 128$, $n_{11} = 384$, $n_{12} = 256$, $n_{21} = 64$, $n_{22} = 40$, $\mu_{11} = -\mu_{12} = [1, 0, \ldots, 0]^\mathsf{T}$, $\mu_{21} = -\mu_{22} = [.87, .5, 0, \ldots, 0]^\mathsf{T}$.

where $\mathcal{H}_0$ is the null hypothesis (say, Class 2) and $\mathcal{H}_1$ the alternative (say, Class 1) and $\zeta = \zeta(\eta)$ is a decision threshold selected in such a way to have the false alarm constraint rate $P(g_t(\mathbf{x}) \geq \zeta \mid \mathbf{x} \in \mathcal{H}_0) \leq \eta$, for some given η. The objective is then here to maximize the correct detection rate $P(g_t(\mathbf{x}) \geq \zeta \mid \mathbf{x} \in \mathcal{H}_1)$.

Figure 1 depicts the algorithm performance through a receiver-operating curve (ROC) for false alarm rates η on synthetic data. Both theoretical (Th) asymptotics (used to set the decision threshold ζ) and actual performances (Sim) are displayed for optimal (Opt) choices of $\tilde{y}$ (Opt) and for $\tilde{y} = [-1, 1, -1, 1]$ (Non-Opt).

The synthetic data is a two-task ($k = 2$) setting in which $x_{1j} \sim \mathcal{N}(\pm\mu_{11}, I_p)$ (i.e., $\mu_{12} = -\mu_{11}$) and $x_{2j} \sim \mathcal{N}(\pm\mu_{21}, I_p)$, where $\mu_{21} = \beta\mu_{11} + \sqrt{1 - \beta^2}\mu_{11}^{\perp}$, μ_{11} is a unit-norm vector and $\mu_{11}^{\perp}$ any unit-norm vector orthogonal to μ_{11}. We take here $\beta = 0.5$, so that both tasks are "slightly" correlated.

The graph confirms, here in the hypothesis testing problem, the large superiority of our proposed optimized MTL LS-SVM over the standard non optimized alternative. Besides, the theoretical classification error prediction is an accurate fit to the actual empirical performance, even for not so large values of p, n_{ij}.

4.2 EXPERIMENTS ON MULTI-CLASS IMAGE CLASSIFICATION

Similarly as in the main article, we now turn to the popular Office+Caltech256 multi-task image classification benchmark (Saenko et al., 2010; Griffin et al., 2007) often exploited for transfer learning. The overall database consists of 10 categories shared by both Office and Caltech256 datasets. As in Table 1 of the main article, we consider in sequence the transfer learning of one out of four possible source tasks, each of which consisting in classifying data from one sub-database (images issued from the Caltech set (c), Webcam images (w), Amazon pictures (a) or dslr images (d)), towards another task; this boils down to $4 \times (4 - 1) = 12$ source-target comparison pairs.)

The results in Table 1 of the main article using VGG features for the image representations are extremely close to 100%. For better discrimination, for the present experiment, we compare the more challenging (since less discriminating) $p = 800$ SURF-BoW features of the Office+Caltech256 images instead of their VGG features. Table 2 here also demonstrates that our proposed improved MTL-LSSVM, despite its simplicity, has stable performances and is highly competitive. The performance difference versus CDLS is nonetheless less accute than for VGG features, which is likely due to SUF-BoW features being less "concentrated" (i.e., not appropriately modelled by concentrated random vectors as per Assumption 1) than VGG features.[1]

---

[1] Those arising from deep neural network training exhibit better regularity Seddik et al. (2020).

Table 1: Classification accuracy for transfer learning on the Office+Caltech256 database, against state-of-the-art alternatives. Here with c(Caltech), w(Webcam), a(Amazon), d(dslr) based on SURF-BoW features. Our proposed approach is systematically best or second to best and best on average.

S/T	c → w	w → c	c → a	a → c	w → a	a → d	d → a	w → d	c → d	d → c	a → w	d → w	Mean score
LSSVM	79.47	47.70	68.10	49.65	68.13	57.50	70.00	73.75	67.50	46.45	74.83	84.11	65.60
MMDT	69.47	42.55	68.95	39.70	65.24	59.50	62.16	**86.06**	56.94	27.92	68.54	**87.88**	61.24
ILS	24.5	20.92	25.21	21.10	22.92	26.25	27.08	43.75	30.00	26.95	15.23	57.62	28.46
CDLS	*82.28*	**54.21**	*73.75*	**54.49**	**71.52**	*68.56*	*70.54*	69.44	*69.44*	**53.86**	**81.59**	82.78	*69.37*
Ours	**86.09**	*49.65*	**75.00**	*50.35*	*68.83*	**73.75**	**71.25**	*72.50*	**77.50**	*48.05*	*80.13*	*85.43*	**69.88**

4.3 EXPERIMENTS ON NATURAL LANGUAGE DATA

We next experiment our findings on the Multi Domain Sentiment Dataset (Blitzer et al., 2007) which consists in a benchmark made of reviews of Amazon products. This dataset contains Amazon product reviews from four different domains: Books, DVD, Electronics, and Kitchen appliances from Amazon.com. Each review is originally associated with a rating of 1 to 5 stars. For simplicity, we are only concerned with whether or not a review is positive (more than 3 stars) or negative (3 stars or lower). Reviews are encoded in 400-dimensional tf*idf feature vectors of bag-of-words unigrams and bigrams. From this data, we construct 12 cross-domain binary classification tasks. Each task consists of $2\,000$ labeled source examples and $1\,000$ labeled target examples. The test data contains $1\,000$ samples. To show the influence of label optimization, we improve the standard MTL LSSVM algorithm by choosing the optimal threshold $\zeta^\star$ obtained from theoretical analysis, rather than $\zeta = 0$ (leading to much worse results). Table 2 reports the accuracy of the classification for our proposed method versus CDLS and the non-optimized LS-SVM[2]. The table shows again that the proposed method is highly competitive. Furthermore, we represent in figure 2 the empirical and the theoretical score distribution for three scenarios of Transfer Learning (Book-DVD, Book-Electronics and Book-Kitchen). The figure illustrates the close fit between theoretical and empirical predictions even for data such as language data thereby suggesting that the Gaussian mixture model is sufficient to tackle a wide range of realistic data. Furthermore, a systematic improvement is remarked between the classical label (± 1) and the optimized one.

Table 2: Classification accuracy for transfer learning on the multi domain sentiment classification database, against state-of-the-art alternatives. Here with b(Book), d(Dvd), e(Elec), k(Kitchen). Our proposed approach is systematically best or second-to-best, and best on average.

S/T	b → d	b → e	b → k	d → b	d → e	d → k	e → b	e → d	e → k	k → b	k → d	k → e	Mean score
LSSVM	**79.70**	80.98	81.60	**79.30**	*82.98*	82.00	76.40	77.50	83.60	*78.30*	79.60	81.18	80.26
CDLS	*78.80*	**85.28**	*83.70*	77.60	82.88	**84.4**	*79.3*	**80.10**	*83.90*	78.00	*80.00*	**82.68**	*81.38*
Ours	78.70	*83.98*	**84.50**	*79.10*	**83.78**	*82.60*	**80.50**	*79.00*	**84.00**	**80.10**	**81.60**	*82.08*	**81.66**

4.4 IMPACT OF THE HYPERPARAMETER CHOICE

In this section, we investigate empirically the impact of the hyperparameter λ on the MTL performance. To that end, we consider the MNIST dataset (Deng, 2012). The setting is that of a binary classification for two tasks, mimicking a transfer learning setting: there, the "target" Task 2 aims to discriminate Class $\mathcal{C}_1$ and Class $\mathcal{C}_2$ respectively composed of images of digit 1 and digit 4. The "source" Task 1 is here used as a support for classification in the target task, and consists of the classification of other pairs of digits: either $(5, 9)$, $(9, 5)$, $(6, 2)$ or $(8, 3)$ (recall that the order of the set of digits (X, Y) is important for the non-optimized MTL-LSSVM since source and target tasks labels are "paired"; thus $(5, 9)$ or $(9, 5)$ digits for the source task will bring different results). We compare here again the non-optimized MTL-LSSVM with labels $y^{\rm bin} = [-1, 1, -1, 1]^{\sf T}$ to our proposed optimized scheme (as detailed in the core of the article). For both methods, the optimal

---

[2] MMDT and ILS methods are only designed for multi class classification.

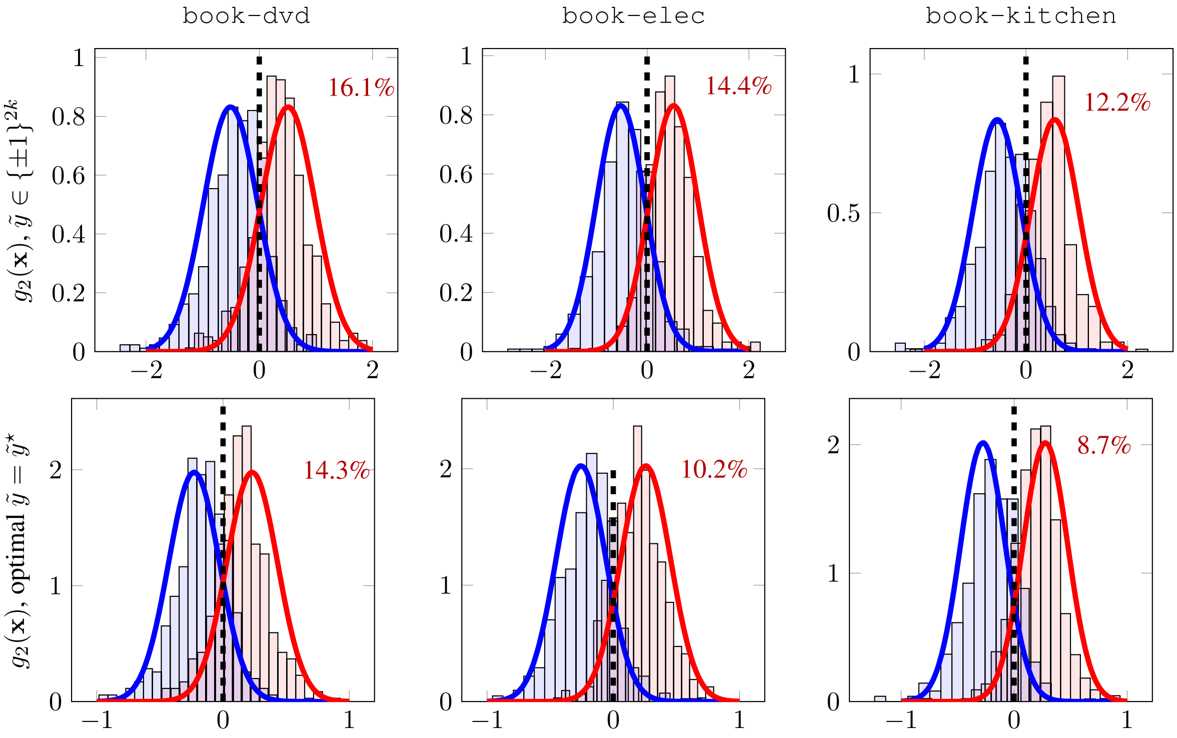

Figure 2: Scores $g_2(\mathbf{x})$ [empirical histogram vs. theory in solid lines] for $\mathbf{x}$ of Class $\mathcal{C}_1$ (red) or Class $\mathcal{C}_2$ (blue) for Task 2 in a 2-task ($k = 2$) setting for: **(top)** classical MTL-LSSVM with $y \in \{\pm 1\}$ and threshold $\zeta = 0$; **(bottom)** proposed optimized MTL-LSSVM with $\tilde{y}^\star$ and estimated threshold ζ; decision thresholds ζ represented in dashed vertical lines; red numbers are misclassification rates; chosen task between Book, DVD, Elec and Kitchen; $p = 400$, $[c_{11}, c_{12}, c_{21}, c_{22}] = [0.25, 0.25, 0.25, 0.25]$, $\gamma = \mathbb{1}_2$, $\lambda = 1$. Histograms drawn from $1\,000$ test samples of each class.

theoretical threshold decision ζ is used (rather than $\zeta = 0$ for the non-optimized setup) in order to emphasize the influence of input score (label) optimization. Figure 3 depicts the performance for both methods as a function of the hyperparameter λ. We recall that, as $\lambda \to 0$, the multi-task scheme becomes equivalent to independent single-task classifiers, while as $\lambda \to \infty$, both source and target tasks are considered together as one task. Figure 3 demonstrates the stability of optimal input labelling with respect to λ: this is explained by the fact that $y^\star$ is a *function of* λ and thus adapts to each value of λ, even if suboptimal. Besides, for appropriate values of λ, the proposed improved labelling can largely outperform the non-optimized setting, even here on real data.

4.5 ANALYSIS OF AN INCREASING NUMBER OF TASKS

The next experiment illustrates the effect of adding more tasks to the transfer learning setting on both synthetic and MNIST datasets. For synthetic data, Gaussian classes with mean $\mu_{ij} = \beta\mu_{i1} + \sqrt{1-\beta^2}\mu_{i1}^\perp$ and various values of β are successively added. For the MNIST dataset, different classifications of digits are added progressively to help classify the specific pair of digits $(1, 4)$. Figure 4 depicts the classification error after each new task addition, both for a classical binary (± 1) input label choice and for the proposed optimized input labels. The figure forcefully illustrates that our proposed framework avoids negative transfer, as the classification error of MTL never increases as the number of tasks grows. This is quite unlike the non-optimized scheme which severely suffers from negative transfer.

REFERENCES

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

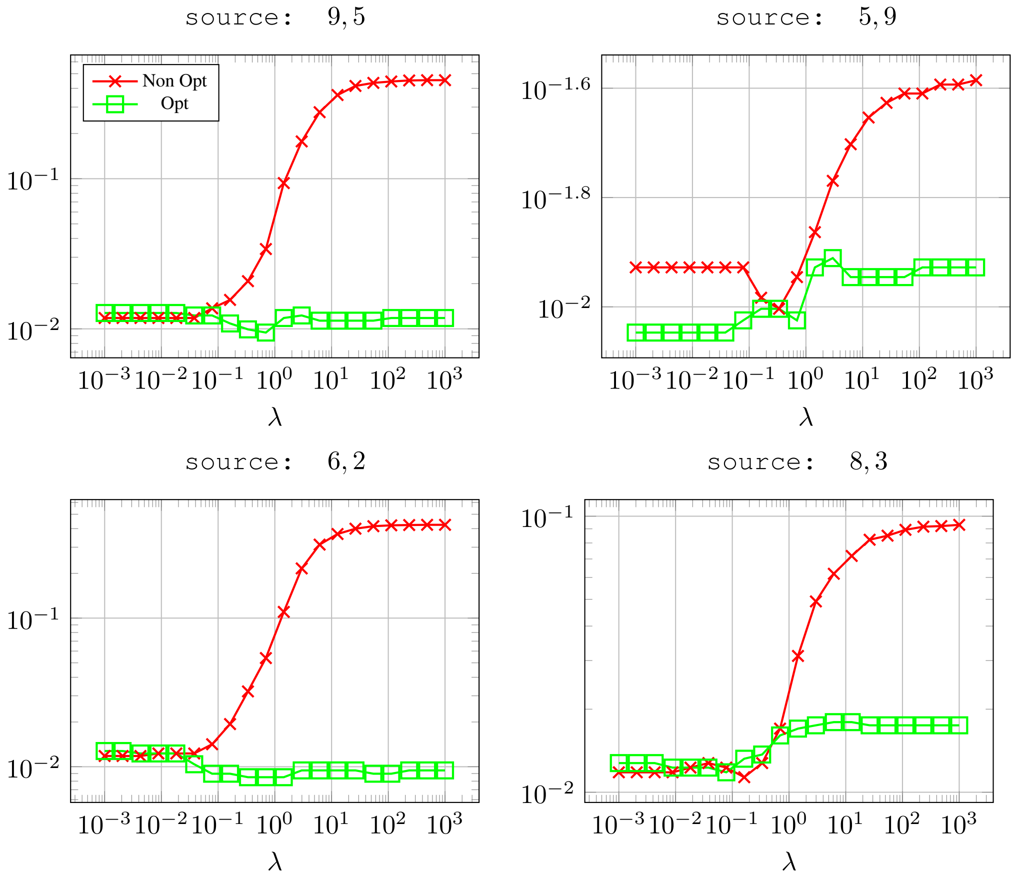

Figure 3: Classification error of digit pair $(1, 4)$ with different source training pairs for classical LSSVM and optimized LSSVM. $n_{11} = n_{12} = 100$, $n_{21} = n_{22} = 10$ and $\gamma = \mathbb{1}_2$. A PCA preprocessing is performed on each image to extract their $p = 100$ principal components; the accuracy is performed over $n_{\text{test}} = 1\,135$ test samples. The proposed method shows a low sensitivity to λ.

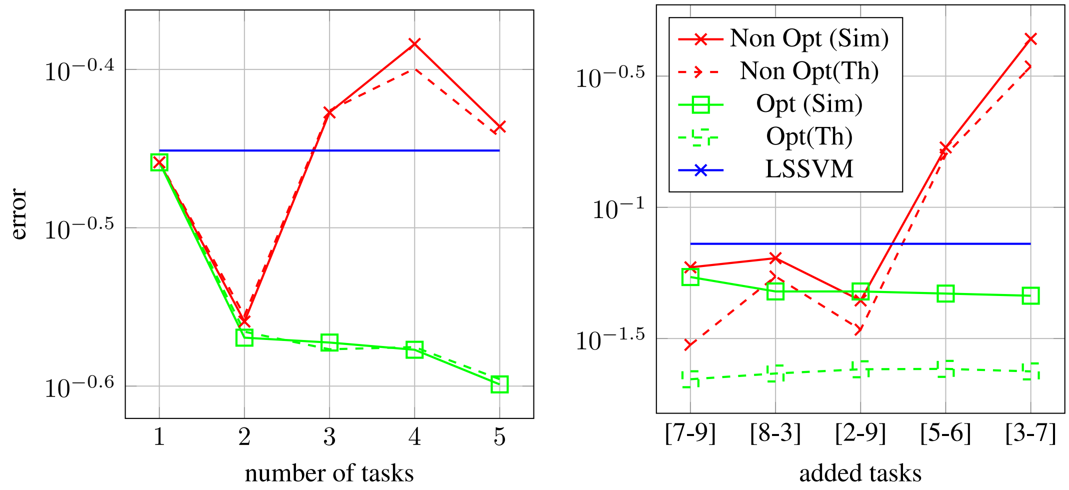

Figure 4: Classification accuracy for increasing number of tasks. **(Left)** Synthetic data with task correlations $\beta = 1, .9, .5, .2, .8$ in this order, $p = 100$ and $c = [.07, .11, .10, .10, .06, .08, .09, .12, .10, .11, .03, .03]^\mathsf{T}$; accuracy evaluated out of $10\,000$ test samples. **(Right)** MNIST dataset with digits $(1, 4)$ as target task, each added task being shown in x-axis; 100 training samples are used for each class of the source tasks and 10 training samples for each class of the target class; HOG features with $p = 144$ for each image digit; accuracy evaluated out of $n_\text{test} = 1\,135$ test samples. For both setting, $\gamma = \mathbb{1}_k$ and $\lambda = 10$. The optimized scheme avoids negative transfer by systematically benefiting from additional tasks.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. *technical report*, 2007.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89 in Mathematical surveys and monographs. American Mathematical Soc., 2001. ISBN 0821837923.

Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019.

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.

Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

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)*, pp. 7480–7484. IEEE, 2019.

Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. *arXiv preprint arXiv:2001.08370*, 2020.