A RANDOM MATRIX ANALYSIS AND OPTIMIZATION FRAMEWORK TO LARGE DIMENSIONAL TRANSFER LEARNING

Romain Couillet

GIPSA-lab, University Grenoble-Alpes
CentraleSupélec, University ParisSaclay.

ABSTRACT

This article proposes a first performance analysis and optimization of a simple transfer learning method, extending the standard least squares support vector machine. By means of a random matrix analysis, we prove that, for simultaneously large and numerous data, the correct classification rate of the learning task is asymptotically predictable and the hyperparameters in the problem are easily tuned so to maximize the output performance. Simulations confirm our findings. This preliminary work opens the path to a systematic exploration of transfer learning methods by means of large dimensional statistics.

Index Terms— Random matrix theory; transfer learning; large dimensional statistics.

1. INTRODUCTION

The advent of the powerful deep learning architectures has resurrected the interest for the field of transfer learning (or domain adaptation) [1]. The idea of transfer learning consists in possessing two databases: a *source* (usually large) and a *target* (usually small) database, and in designing an algorithm that learns to classify the target dataset by exploiting its similarity to the source dataset (the classification of which is *not* the sought-for objective). In a deep learning context, the existence of enormous source databases (such as ImageNet) has allowed for the design of powerful neural networks that can process numerous classes of images; these neural networks are now reused in order to classify data arising from different and much smaller (target) datasets, from the intuitive idea that these networks have essentially learned the ability to “see” and discriminate differing items from any given dataset.

Many algorithms exist that similarly extend popular machine learning architectures (such as support vector machines and Bayesian methods) to transfer learning machines. Yet, as is the case for most advanced (non-linear, optimization-based) machine learning methods, the theoretical understanding of their inner-workings, limitations, and expected performances is to date quite restricted. Yet, thanks to recent advances in random matrix theory and large dimensional statistics, first lights are being cast on the performance analysis of most of these machine learning methods [2, 3, 4]. In this article, we specifically focus on a simple generalization of a kernel regression method (also known as least-squares support vector machine), adapted to the transfer learning context. This is inspired, from a theoretical standpoint by the technical findings of [4] and from a practical aspect by the ideas of [5] simplified to finding a single hyperplane (or regression vector).[1] By proposing a simple and clear setting and, consequently, simple and readable results, the article serves as a preliminary exploration of a random matrix framework for the theoretical analysis of transfer learning asymptotics.

By means of a large dimensional analysis, whereby the size p and number n of data are large and of the same order of magnitude, we manage to successively (i) find the asymptotic probability of success of the transfer learning task for every given regression output vector $y \in \mathbb{R}^n$ (containing the outputs for every data point, from both source and target classes), (ii) from this result, we find the regression output vector y that maximizes this probability of success, and finally (iii) we determine the optimal kernel regularization value.

2. MODEL AND PRELIMINARIES

We consider the general setting of a k-class kernel linear regression problem. Let $X = [X_1, \ldots, X_k] \in \mathbb{R}^{p\times n}$ be the collection of independent data vectors from a k-class Gaussian mixture model defined, for $a \in \{1, \ldots, k\}$, by $X_a = [x_{a1}, \ldots, x_{an_a}] \in \mathbb{R}^{p\times n_a}$, $n_1 + \ldots + n_k = n$, and

$$x_{ai} \sim \mathcal{N}(\mu_a, I_p).$$

We denote $\mathcal{C}_a$ the a-th class. Our objective is to discriminate data arising from any *target* pair $(\mathcal{C}_{t_1}, \mathcal{C}_{t_2})$ $(t_1, t_2 \in \{1, \ldots, k\})$ of classes by benefiting from the existence of data from other (possibly correlated) classes.

To this end, we proceed by a standard *kernel linear regression* method, also referred to as *least-square support vector machine*. Specifically, from a support vector machine standpoint, we first aim to solve, for all classes, the problem

$$\underset{\beta\in\mathbb{R}^p, b\in\mathbb{R}}{\operatorname{argmin}} \|\beta\|^2 + \frac{1}{\gamma}\|e\|^2, \quad y = \frac{1}{\sqrt{p}}PX^\top\beta + b1_n + e \qquad (1)$$

where $y = [y_{11}, \ldots, y_{kn_k}]^\top \in \mathbb{R}^n$ is the regression output vector attached to data vectors $X = [x_{11}, \ldots, x_{kn_k}] \in \mathbb{R}^{p\times n}$, $P = I_n - \frac{1}{n}1_n1_n^\top$ is a centering projector and $\gamma > 0$ is a tuning parameter. In the large dimensional regime $n, p \to \infty$ studied in this article, the normalization by $1/\sqrt{p}$ is necessary for the problem to be non degenerate in the limit (here for $\|X/\sqrt{p}\|$ to be of order $O(1)$). Similarly, the projector P ensures, by centering them around zero, that the average of the means $\mu_1, \ldots, \mu_k$ does not have an arbitrary high amplitude (note of course that classification cannot be impaired by a common translation of all data).

Couillet’s work is supported by the IDEX DataScience Chair GSTATS at University Grenoble-Alpes and the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

[1] That is, we set $\lambda_1 = 0$ in Problem 2.1 of [5] and replace the soft constraints by a quadratic cost.

In essence, the objective of the optimization is to design a hyperplane in $\mathbb{R}^p$ (defined by the normal vector β and shifted by b from the origin) that best separates the n vectors in two groups: those associated with negative versus positive regression outputs y_{ai}. Choosing a small γ in particular forces all x_{ai} to fall on the correct side of the hyperplane (thus inducing a *hard* constraint) while a large γ ensures the correct classification of the largest proportion of the data but allows for more errors.

In the present transfer learning context, we thus expect the k data classes to naturally split in two groups: those data "resembling" data from the first target class $\mathcal{C}_{t_1}$ and those resembling date from class $\mathcal{C}_{t_2}$. Since data from the same class are statistically equal, the second part of the method consists in attributing scalars $\tilde{y}_a = y_{ai}$ to each data class, so as to minimize the probability of misclassification of data from class $\mathcal{C}_{t_1}$ and $\mathcal{C}_{t_2}$. We will conventionally demand that $\tilde{y}_{t_1} > \tilde{y}_{t_2}$.

The solution to (1) is explicit and given by $\beta = X\alpha$ and

$$\alpha = S(y - b1_n), \quad b = \frac{1_n^\top S y}{1_n^\top S 1_n},$$
$$S = \left(\frac{1}{p} P X^\top X P + \gamma I_n\right)^{-1}.$$

More importantly, given some decision threshold ξ, for a new datum $x \in \mathbb{R}^p$ from one of the two target classes, the ultimate classification of x is given by the output of the test

$$g(x) \equiv \frac{1}{p}\alpha^\top X^\top \left(x - \frac{1}{n}X1_n\right) \underset{\mathcal{C}_{t_2}}{\overset{\mathcal{C}_{t_1}}{\gtrless}} \xi. \tag{2}$$

As such, the important parameter to study here is α rather than β.

As previously mentioned, our objective is to determine the vector y for which the aforementioned test has minimal probability of error to allocate x to $\mathcal{C}_{t_a}$ whenever $x \sim \mathcal{N}(\mu_{t_a}, I_p)$, $a \in \{1, 2\}$. This boils down to studying the statistics of $g(x)$ for such x. However, the intricate expression of α prevents this study for all fixed n, p in general. Our approach is to instead investigate $g(x)$ in the large dimensional asymptotics where $n, p \to \infty$ with $n/p \to c_0 > 0$.

3. MAIN RESULTS

3.1. Technical Results

In order to avoid trivial results in the large n, p limit, we need to impose a controlled growth rate for the distances $\|\mu_a - \mu_b\|$ for all $a, b \in \{1, \ldots, k\}$. Precisely, we make the following assumption.

Assumption 1 (Growth Rate) *As $n \to \infty$, $n/p \to c_0 > 0$ and for all $a \in \{1, \ldots, k\}$, $n_a/n \to c_a > 0$. We shall denote $c = [c_1, \ldots, c_k]^\top$ and $P_c = \operatorname{diag}(c) - cc^\top$. Besides, for $M = [\mu_1, \ldots, \mu_k] \in \mathbb{R}^{p \times k}$, as $p \to \infty$,*

$$M^\top M \to \mathcal{M} \in \mathbb{R}^{k \times k}.$$

The last assumption implies in particular that $\|\mu_a - \mu_b\| = O(1)$ with respect to p.

In order to introduce our main results, a few reminders from the random matrix asymptotics of the *resolvent matrix*

$$S = \left(\frac{1}{p} P X^\top X P + \gamma I_n\right)^{-1}$$

of the kernel $\frac{1}{p} P X^\top X P$ are necessary. Denoting $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$ where $j_a = [0_{n_1 + \ldots + n_{a-1}}, 1_{n_a}, 0_{n_{a+1} + \ldots + n_k}]^\top$ is the indicator vector of class $\mathcal{C}_a$, we may write $X = MJ^\top + W$ where W has independent standard Gaussian entries. As such, the kernel $\frac{1}{p} P X^\top X P$ may be seen as a low rank perturbation of the random matrix $\frac{1}{p} P W^\top W P$, the asymptotic behavior of which has been extensively studied in the random matrix literature. In particular, we have the following identities: as $n, p \to \infty$ with $n/p \to c_0$,

$$Q \equiv \left(\frac{1}{p} P X^\top X P + \gamma I_n\right)^{-1} \leftrightarrow \left(\frac{P}{1 + c_0 m(-\gamma)} + \gamma I_p\right)^{-1}$$
$$\tilde{Q} \equiv \left(\frac{1}{p} X P X^\top + \gamma I_p\right)^{-1} \leftrightarrow \tilde{m}(-\gamma) I_p$$

where $A \leftrightarrow B$ stands for the fact that (i) for all deterministic vectors u, v of unit norm, $u^\top (A - B) v \xrightarrow{\text{a.s.}} 0$, and (ii) for all deterministic matrix D of bounded operator norm, $\frac{1}{n} \operatorname{tr} D(A - B) \xrightarrow{\text{a.s.}} 0$. There $m(z)$ and $\tilde{m}(z)$ are the so-called Stieltjes transforms of the popular Marčenko–Pastur distribution, characterized by their being the unique positive solution, for $z < 0$, to the equations

$$m(z) = (1 - c_0 - z - c_0 z m(z))^{-1}$$
$$\tilde{m}(z) = c_0 m(z) + (c_0 - 1) z^{-1}.$$

Alternatively, $m(z)$ can be expressed under the (numerically more convenient but theoretically less useful) explicit form

$$m(z) = \frac{1 - c_0 - z}{2 c_0 z} - \frac{1}{2 c_0 z}\sqrt{(1 - c_0 - z)^2 - 4 c_0 z}.$$

For further use, we also mention their respective derivatives

$$m'(z) = m(z)^2 \frac{1 + c_0 m(z)}{1 - c_0 z m(z)^2}$$
$$\tilde{m}'(z) = c_0 m'(z) + (1 - c_0) z^{-2}.$$

These results are fundamental here as our proof technique (deferred to an extended version) consists in "splitting" the matrix S (defining the solutions α, b of the present problem) into the well-known "pure noise" matrices Q, $\tilde{Q}$ and the low-dimensional statistical properties of the data: $M^\top M$ and c.

Equipped with these notations and preliminary results, we present our main technical theorem.

Theorem 1 (Asymptotics of $g(x)$) *Let Assumption 1 hold and set the class regression outputs to be $y = J\tilde{y}$ for some $\tilde{y} \in \mathbb{R}^k$. Then, for $x \sim \mathcal{N}(\mu_a, I_p)$ independent of X, and $g(x)$ defined as per (2),*

$$g(x) \to \mathcal{N}(m_a, \sigma^2)$$

in distribution where

$$m_a = (1 - \gamma m(-\gamma)) \tilde{y}^\top R c_0 P_c \mathcal{M} e_a$$
$$\sigma^2 = (m(-\gamma) - \gamma m'(-\gamma)) \tilde{y}^\top R (c_0 P_c + c_0^2 P_c \mathcal{M} P_c) R^\top \tilde{y}.$$

with $R = (I_k + (1 - \gamma m(-\gamma)) c_0 P_c \mathcal{M})^{-1}$ and $e_a \in \mathbb{R}^k$ the vector with $[e_a]_i = \delta_{ai}$.

A few important remarks are in order. The theorem claims that, irrespective of the class of x, the output $g(x)$ is asymptotically Gaussian with constant variance σ^2; only $\mathrm{E}_x[g(x)]$ differs across classes.

Also note, as expected, that the means m_a and variance σ^2 exclusively depend (in a possibly intricate manner at first sight) on M through $\mathcal{M}$, on c_0 and γ through $m(-\gamma)$, and on c through P_c.

Theorem 1 now implies that, in the large n, p limit, the optimal decision rule for the task of deciding between class $\mathcal{C}_{t_1}$ and $\mathcal{C}_{t_2}$ should be to set

$$g(x) \equiv \frac{1}{p}\alpha^\top X^\top \left(x - \frac{1}{n}X1_n\right) \underset{\mathcal{C}_{t_2}}{\overset{\mathcal{C}_{t_1}}{\gtrless}} \frac{1}{2}(m_{t_1} + m_{t_2})$$

resulting in the probability of correct classification to be

$$\begin{aligned}
\mathrm{P}_{x\sim\mathcal{N}(\mu_{t_1},I_p)}\left(g(x) > \frac{1}{2}(m_{t_1}+m_{t_2})\right) &\to \mathcal{Q}\left(\frac{m_{t_2}-m_{t_1}}{2\sigma}\right)\\
\mathrm{P}_{x\sim\mathcal{N}(\mu_{t_2},I_p)}\left(g(x) < \frac{1}{2}(m_{t_1}+m_{t_2})\right) &\to \mathcal{Q}\left(\frac{m_{t_2}-m_{t_1}}{2\sigma}\right)
\end{aligned} \tag{3}$$

for $\mathcal{Q}(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$.

3.2. Optimization of $\tilde{y}$

A further consequence of the classification rate asymptotics, and of the independence of σ^2 on the classes, is that the vector $\tilde{y}$ that optimizes the probability of correct classification must (asymptotically) be the maximizer of $(m_{t_2} - m_{t_1})^2/\sigma^2$. Thus we look for

$$\operatorname*{argmax}_{\tilde{y}\in\mathbb{R}^k} \frac{\|\tilde{y}^\top R c_0 P_c \mathcal{M}(e_{t_1} - e_{t_2})\|^2}{\tilde{y}^\top R(c_0 P_c + c_0^2 P_c \mathcal{M} P_c)R^\top \tilde{y}}$$

which is easily solved as

$$\tilde{y} = \left[\mathcal{M} - \gamma m(-\gamma) c_0 \mathcal{M} P_c (I_k + c_0 \mathcal{M} P_c)^{-1}\mathcal{M}\right](e_{t_1} - e_{t_2})$$

up to a positive multiplicative factor.

For this choice of $\tilde{y}$, the associated asymptotic correct classification rate (Equation 3) is then given by

$$\begin{aligned}
&\mathcal{Q}\left(\frac{(1-\gamma m(-\gamma))^2}{m(-\gamma) - \gamma m'(-\gamma)}\zeta\right) \text{ with}\\
&\zeta = c_0(e_{t_1} - e_{t_2})^\top \mathcal{M} P_c \left(I_k + c_0 \mathcal{M} P_c\right)^{-1} \mathcal{M}(e_{t_1} - e_{t_2}).
\end{aligned}$$

This result clearly isolates the impact of γ from the rest of the parameters. It is thus possible to optimize on γ. Recalling that the Stieltjes transform $m(z)$ of a probability measure ν is defined as $\int (t-z)^{-1}\nu(dt)$, we easily find that

$$\begin{aligned}
1 - \gamma m(-\gamma) &= \int \frac{t\nu(dt)}{t+\gamma}\\
m(-\gamma) - \gamma m'(-\gamma) &= \int \frac{t\nu(dt)}{(t+\gamma)^2}
\end{aligned}$$

where ν is here the Marčenko–Pastur distribution. It is not difficult to show that $\frac{(1-\gamma m(-\gamma))^2}{m(-\gamma)-\gamma m'(-\gamma)}$ is then a growing function of γ with limiting value 1 as $\gamma \to \infty$. As such, quite surprisingly, the regularizer γ must be chosen as large as possible. As a consequence, we have the limiting optimal correct classification rate given by

$$\mathcal{Q}\left(c_0(e_{t_1} - e_{t_2})^\top \mathcal{M} P_c \left(I_k + c_0 \mathcal{M} P_c\right)^{-1} \mathcal{M}(e_{t_1} - e_{t_2})\right).$$

In passing, remarking that

$$\begin{aligned}
S &= \left(\frac{1}{p}PX^\top XP + \gamma I_n\right)^{-1}\\
&\simeq \frac{1}{\gamma}I_n - \frac{1}{\gamma^2}\frac{1}{p}PX^\top XP + O(\gamma^{-3})
\end{aligned}$$

in the large γ limit, suggests that in the large dimensional limit the linear regression operates merely as a matched-filter.

Consider for a moment the setting where $k = 4$ with two *source* classes $\mathcal{C}_{s_1} = \mathcal{C}_1, \mathcal{C}_{s_2} = \mathcal{C}_2$ and two *target* classes $\mathcal{C}_{t_1} = \mathcal{C}_3, \mathcal{C}_{t_2} = \mathcal{C}_4$. Then, the expression of the optimal $\tilde{y}$ induces the following consequences:

- if $\mu_{s_1} = \mu_{t_1}$ and $\mu_{s_2} = \mu_{t_2}$, then $\tilde{y} = [1, -1, 1, -1]^\top$ is optimal. This is expected as this boils down to a two-class regression problem.
- if $\mu_1, \ldots, \mu_4$ are pairwise orthogonal and $c_{t_1} = c_{t_2}$, then $\tilde{y} = [0, 0, 1, -1]^\top$ is optimal. This is also expected as μ_{s_1}, μ_{s_2} bring no additional information. Yet, this result no longer holds true if $c_{t_1} \neq c_{t_2}$; this is an interesting, not immediate, consequence that suggests the need to create an artificial bias in order to compensate for the uneven cardinality of the training data.
- the most interesting (and practical) setup is of course when $\mu_{s_a}^\top \mu_{t_a}/(\|\mu_{s_a}\|\|\mu_{t_a}\|) \simeq 1$ for both $a \in \{1, 2\}$. There, (i) the strength of the alignment between source and target statistics, characterized by $M^\top M$, and (ii) the relative sizes of the training datasets, evaluated by c, impact the entries of the optimal $\tilde{y}$. In general, if $c_{s_a} \gg c_{t_a}$ while source and target means are quite aligned, much larger absolute values are attributed to the "source" entries of $\tilde{y}$. On the opposite, if $c_{s_a} \sim c_{t_a}$ while source and target means are rather misaligned, then much of the weigh is affected to the "target" entries of $\tilde{y}$.

3.3. Practical application

From a practical standpoint, note that the optimal value for $\tilde{y}$ only depends on the unknown $\mathcal{M} = \lim_p M^\top M$, the rest of the variables being known to the experimenter. To estimate $\tilde{y}$, it is thus sufficient to evaluate $M^\top M$. Under the growth rate conditions of Assumption 1, it is easy to show that the following estimators are consistent:

$$\begin{aligned}
&\frac{1}{n_a n_b} 1_{n_a} X_a^\top X_b 1_{n_b} \xrightarrow{\text{a.s.}} \mathcal{M}_{ab}, \quad a \neq b\\
&\frac{4}{n_a^2} 1_{\frac{n_a}{2}} X_{a,1}^\top X_{a,2} 1_{\frac{n_a}{2}} \xrightarrow{\text{a.s.}} \mathcal{M}_{aa}
\end{aligned}$$

where we denoted $X_a = [X_{a,1}, X_{a,2}]$ with $X_{a,1}, X_{a,2} \in \mathbb{R}^{p\times n_a/2}$.

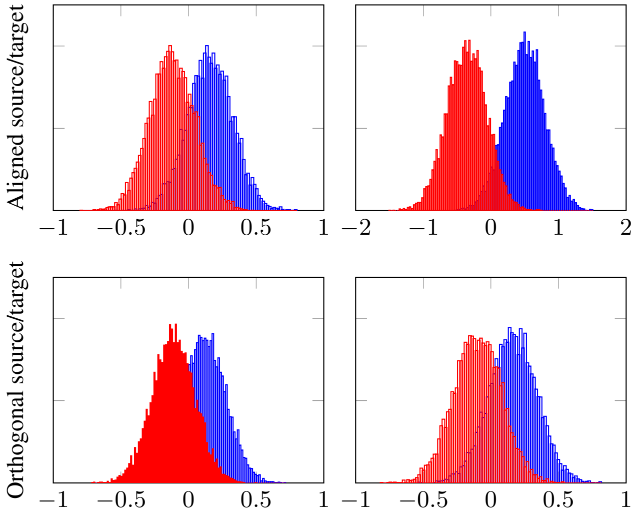
It is thus sufficient to plug these estimators into the formula for $\tilde{y}$ to reach, with probability one, optimal asymptotic empirical classification performance.

4. SIMULATION RESULTS

We consider here a transfer learning context with two source classes $\mathcal{C}_{s_1}$ and $\mathcal{C}_{s_2}$ and two target classes $\mathcal{C}_{t_1}$ and $\mathcal{C}_{t_2}$, as previously introduced. The considered settings are simple and quite symmetric in order to avoid confusing interpretations. Details are accessible in the various figure captions. In the whole section, we take $\gamma = 10^6$.

Figure 1 displays the histogram of $g(x)$ for $10\,000$ random draws of $x \in \mathcal{C}_{t_1}$ and $x \in \mathcal{C}_{t_2}$, in a setting where, for $a \in \{1,2\}$, μ_{s_a} is either strongly aligned or orthogonal to μ_{t_a}. In the figure are also compared the settings where $n_{s_1} = n_{s_2}$ is either small or large. The output regression vector $\tilde{y}$ is taken as the estimated optimal in either case (see Section 3.3). Significant classification performance gains are achieved in the setting of strongly correlated source and targets, while no gain is obtained otherwise.

Figure 2 carries on this analysis by now comparing various strategies for setting the regression output $\tilde{y}$. We compare here the standard SVM approach that treats $\mathcal{C}_{s_a}$ and $\mathcal{C}_{t_a}$ (for both $a \in \{1,2\}$) as a unique class, i.e., $\tilde{y} = [1,-1,1,-1]^\top$, to the optimum achieved by $\tilde{y}$ as presented in Section 3.2. A further comparison to the estimated optimal value of $\tilde{y}$, as per Section 3.3, is also depicted. It is interesting here to see, in this slightly extreme setting, that misjudging the proximity between source and target data can be severely detrimental to the transfer learning method, as observed with the SVM approach in the orthogonal source-target setting.



$(n_{s_1} = n_{s_2} = 0)$ $(n_{s_1} = n_{s_2} = 2000)$

Fig. 1. Histogram of the outputs $g(x)$ for $x \in \mathcal{C}_{t_1}$ (blue) and $x \in \mathcal{C}_{t_2}$ (red); $p = 512$, $n_{t_1} = 10$, $n_{t_2} = 20$, $\mu_{s_1} = \sqrt{2}[1,1,0,\ldots]^\top$, $\mu_{s_2} = \sqrt{2}[-1,-1,0,\ldots]^\top$. **(Top)** $\mu_{t_1} = [2,0,0,\ldots]$, $\mu_{t_2} = [-2,0,0,\ldots]^\top$; **(Bottom)** $\mu_{t_1} = [0,0,2,\ldots]$, $\mu_{t_2} = [0,0,-2,\ldots]^\top$. **(Left)** $n_{s_1} = n_{s_2} = 0$; **(Right)** $n_{s_1} = n_{s_2} = 2000$.

We conclude this section with a comparative study in Table 1 of the MNIST handwritten digit dataset classification under the present transfer learning setting. We consider here the task of learning to discriminate digits 8 and 9 with a source training on digits 3 and 4 (since 3 and 8, and 4 and 9 are respectively close in shape). Interestingly the naiv pure-SVM approach is counterproductive in the presence of numerous source data while an optimal weight distribution improves the classification by up to 2%.

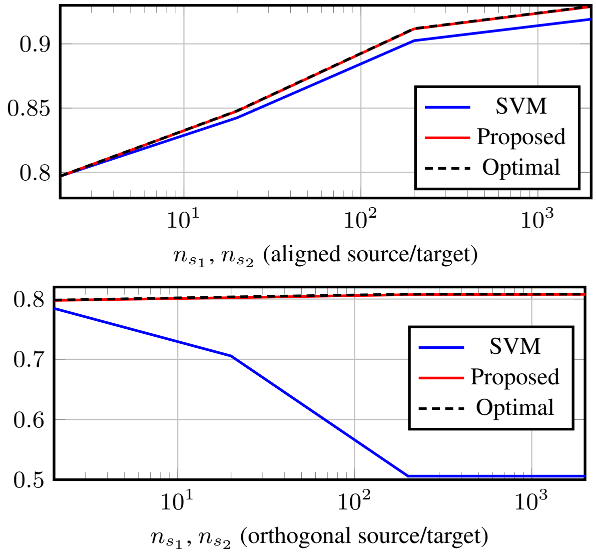


Fig. 2. Correct classification rate for $\tilde{y} = [1,-1,1,-1]^\top$ (SVM), $\tilde{y}$ obtained from the estimate of Section 3.3 (Proposed), and optimal $\tilde{y}$ (Optimal); setting of Figure 1 for varying values of $n_{s_1} = n_{s_2}$.

	SVM	Proposed	Optimal
$n_{s_1} = n_{s_2} = 0$	0.8941	0.8926	0.8966
$n_{s_1} = n_{s_2} = 200$	0.8656	0.9036	0.9171

Table 1. MNIST data correct classification rate; source: digits 3 & 4, target: digits 8 & 9; $n_{t_1} = n_{t_2} = 4$. Comparison between standard SVM, proposed approach and optimal (as if Gaussian).

5. CONCLUDING REMARKS

This article introduced first steps into a large dimensional analysis of transfer learning algorithms. The chosen illustrative example of a simple kernel regression approach is instrumental to already unveil multiple unexpected outcomes: (i) the possibility to largely optimize the problem in the large dimensional setting (by determining the optimal regression output coefficients y), and (ii) the observation that the kernel ridge regression parameter γ is optimal when taken arbitrarily large, thereby turning the regression into a mere matched-filtering operation. The present work is however extremely restricted in both its modelling and transfer learning design assumptions. Future works shall attempt to study state-of-the-art transfer learning algorithms under the same random matrix umbrella.

6. REFERENCES

[1] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[2] R. Couillet and F. Benaych-Georges, “Kernel spectral clustering of large dimensional data,” *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[3] X. Mai and R. Couillet, “Random matrix-inspired improved semi-supervised learning on graphs,” in *International Conference on Machine Learning*, 2018.

[4] Z. Liao and R. Couillet, “A large dimensional analysis of least squares support vector machines,” *(in Press) IEEE Transactions on Signal Processing, arXiv preprint arXiv:1701.02967*, 2018.

[5] T. Evgeniou and M. 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.