Apprentissage semi-supervisé avec données partiellement étiquetées

Victor LÉGER Romain COUILLET

Laboratoire d'Informatique de Grenoble
700 avenue Centrale, 38401 St Martin d'Hères, France

Résumé – Cet article propose d'étendre le cadre de l'apprentissage semi-supervisé en y intégrant des données mal étiquetées, ou encore étiquetées avec des avis divergents. Nous montrons quel gain de telles étiquettes peuvent apporter à un problème de classification, à la fois d'un point de vue théorique et d'un point de vue pratique. Notre modèle s'accompagne de garanties théoriques de performance qui permettent un contrôle précis de l'algorithme et de ses paramètres. De plus, la théorie de l'information nous montre que notre algorithme atteint des performances quasi-optimales.

Abstract – This article extends the framework of semi-supervised learning by integrating poorly labeled data, or data labeled with different opinions. We show which benefits such labels can bring to a classification problem, from both theoretical and practical standpoints. Our model comes along with theoretical guarantees that allow for a precise control of the algorithm and its parameters. Moreover, information theory shows that our algorithm is close to optimal.

1 Introduction

La classification supervisée repose sur l'existence de grandes quantités de données *étiquetées*. Pour pallier l'étiquetage des données, processus généralement fastidieux, l'apprentissage semi-supervisé exploite des données non étiquetées qui augmentent "gratuitement" la base de données. Dans le cas de données gaussiennes, l'étude théorique de l'apprentissage semi-supervisé a permis d'une part de calculer les performances maximales atteignables par cette approche [1], et d'autre part de proposer un algorithme se rapprochant de ces performances maximales [2].

Pour nous rapprocher encore plus d'un contexte pratique où les données ne sont pas toujours correctement étiquetées par les experts, ou encore étiquetées différemment (biaisées) selon les avis, notre modèle intègre un niveau de confiance dans l'étiquetage. Nous verrons à l'aide de simulations comment de telles données incertaines permettent de faire la jonction entre les données étiquetées et non étiquetées.

Aux antipodes de ce que propose le *deep-learning*, mis en avant de manière indiscriminée dans le domaine de l'apprentissage, nous produisons ici un modèle qui prend en compte autant de paramètres réalistes que possible et pertinents pour l'usager, mais qui reste néanmoins simple, flexible et analysable. Nous démontrons alors, par le biais de la théorie des matrices aléatoires qu'un optimum de classification peut être atteint, qui étend ainsi *théoriquement* – et donc avec de réelles garanties de contrôle et de performance – les outils pratiquement étriqués d'apprentissage semi-supervisé.

La Section 2 présente le modèle et les résultats théoriques qui permettent à notre méthode d'atteindre des performances quasi-optimales dans le cas de données correctement étiquetées. La Section 3 confirme cette quasi-optimalité par la simulation, et analyse empiriquement comment notre méthode se comporte lorsque l'on utilise des données incertaines.

2 Modèle et résultats théoriques

2.1 Modèle et hypothèses

Nous étudions une tâche de classification semi-supervisée binaire (avec 2 classes de données), dont la base d'entraînement $\mathbf{X} \in \mathbb{R}^{p\times n}$ est un ensemble de n vecteurs de données indépendantes de dimension p, divisé en n_ℓ données étiquetées $\mathbf{X}_\ell = \{\mathbf{x}_i\}_{i=1}^{n_\ell}$ et n_u données non étiquetées $\mathbf{X}_u = \{\mathbf{x}_i\}_{i=n_\ell+1}^{n}$. À chaque vecteur de donnée étiquetée $\mathbf{x}_i$ est associé un couple (d_{i1}, d_{i2}) tel que $d_{i1} + d_{i2} = 1$, représentant les probabilités *pré-estimées* que le vecteur appartienne à une classe ou à l'autre. Notre objectif est de prédire les classes *effectives* d'appartenance des données non étiquetées $\mathbf{X}_u$. Cette prédiction passe par l'évaluation d'un score f_i pour chaque donnée. En fonction de leurs scores, les données non étiquetées sont attribuées à une classe ou à l'autre.

Hypothèse 1 (Sur la distribution des données) *Les colonnes de la matrice des données* $\mathbf{X}$ *sont des vecteurs aléatoires gaussiens indépendants. Plus précisément, les échantillons de données* $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *sont des observations i.i.d. telles que* $\mathbf{x}_i \in \mathcal{C}_j \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{I}_p)$ *où* $\mathcal{C}_j$ *désigne la Classe* j.

Par souci de simplicité, nous utilisons cette hypothèse plutôt restrictive, mais il est à noter que celle-ci peut être assouplie et conduire aux mêmes résultats, en n'utilisant non pas des vecteurs gaussiens mais des vecteurs concentrés, classe de vecteurs aléatoires plus large, qui contient entre autres les vecteurs produits par des GANs (Generative Adversarial Networks). Ces réseaux, capables de générer des images réalistes, produisent des représentations qui, ici, se comportent de manière similaire aux vecteurs gaussiens [3].

Nous faisons l'hypothèse que la dimension p des données n'est pas négligeable par rapport au nombre de données d'entraînement n. Cette hypothèse classique de la théorie des matrices aléatoires couvre de manière plus réaliste les applications modernes d'apprentissage (où la dimension des données est

grande), par opposition à l'hypothèse historique où p est fixé et $n \to \infty$. Plus précisément :

Hypothèse 2 (Ratio asymptotique) *Lorsque $n \to \infty$, alors $p/n \to c > 0$. De plus, $n_j/n \to \rho_j > 0$, avec n_j le nombre de données qui appartiennent à la Classe j.*

Dans la plupart des modèles de classification supervisée, à chaque donnée $\mathbf{x}_i$ est associée une étiquette y_i, et cette étiquette prend en général la valeur $+1$ ou -1, selon que la donnée appartient à une classe ou à l'autre. Cependant, nous souhaiterions ici que les étiquettes puissent prendre plus que deux valeurs. En effet, dans notre cas, les valeurs $+1$ et -1 correspondraient respectivement aux couples de probabilité $(0, 1)$ et $(1, 0)$, c'est-à-dire aux cas extrêmes où les données sont étiquetées dans une classe ou dans l'autre de manière certaine. Pour tous les autres cas, y_i devrait prendre une valeur intermédiaire, dans l'intervalle $]-1, 1[$. Nous choisissons donc de fixer $y_i = d_{i1}.(-1) + d_{i2}.(+1) \in [-1, 1]$

Allons plus loin : le choix de -1 et $+1$ comme valeurs extrêmes est arbitraire, et des travaux comme [4] ont montré qu'il n'était pas optimal. Nous remplaçons donc -1 et $+1$ par les constantes $\tilde{y}_1$ et $\tilde{y}_2$, dont le choix sera étudié plus en détail dans la section 2.3. On a donc l'étiquette $y_i = d_{i1}\tilde{y}_1 + d_{i2}\tilde{y}_2$ pour chaque donnée étiquetée. Nous fixons également $y_i = 0$ pour chaque donnée *non étiquetée*, et nous justifierons ce choix dans la section suivante.

Les approches par régularisation d'un graphe (comme par exemple la méthode de régularisation du Laplacien) sont couramment utilisées pour résoudre des problèmes de classification semi-supervisée [5, 6]. L'idée générale est de propager les étiquettes associées aux données étiquetées aux données non étiquetées, en se basant d'une part sur le fait que les données étiquetées doivent avoir des scores proches de leurs étiquettes, et d'autre part sur le fait que deux points proches dans l'espace des données doivent se voir attribuer des scores proches. Pour quantifier la proximité entre deux données, on utilise les poids $\omega_{ii'} = \frac{1}{p}\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle$, qui sont d'autant plus grands que les données associées sont semblables. Cette méthode peut s'exprimer sous la forme d'un problème d'optimisation :

$$\min_{f_1,\ldots,f_n} \sum_{i=1}^{n}\sum_{i'=1}^{n} \omega_{ii'} \left(f_i - f_{i'}\right)^2 + \alpha \sum_{i=1}^{n} (f_i - y_i)^2. \qquad (1)$$

Cette équation est constituée de deux termes :

- Le premier terme encourage les données semblables à avoir des scores semblables.
- Le second terme encourage le score de chaque donnée étiquetée à être proche de l'étiquette associée, et permet de tempérer l'amplitude des scores.

L'hyperparamètre $\alpha > 0$ quantifie l'importance du second terme. Plus α est grand, plus on tend vers un régime supervisé.

Comme suggéré dans [2], les données sont centrées pour s'affranchir de certains biais liés à la grande dimension, ce qui nous permet de réécrire le problème précédent sous forme matricielle, avec $\tilde{\mathbf{W}}$ la matrice des poids $\omega_{ii'}$ après centrage des données, $\mathbf{f}$ le vecteur des f_i et $\mathbf{y}$ le vecteur des y_i :

$$\min_{\mathbf{f}} \alpha\|\mathbf{f} - \mathbf{y}\|^2 - \mathbf{f}^\mathsf{T}\tilde{\mathbf{W}}\mathbf{f}. \qquad (2)$$

2.2 Solution du problème

Ce problème est convexe, à condition que $\alpha > \alpha_0 \equiv \|\tilde{\mathbf{W}}\|$. Nous supposerons désormais que c'est le cas. Ce problème d'optimisation convexe a alors une solution explicite (voir par exemple [7]), qui est :

$$\mathbf{f} = \mathbf{y} + \frac{1}{\alpha p}\mathbf{X}^\mathsf{T}\left(\mathbf{I}_p - \frac{\mathbf{X}\mathbf{X}^\mathsf{T}}{\alpha p}\right)^{-1}\mathbf{X}\mathbf{y}. \qquad (3)$$

Notons $\mathbf{y}_\ell$ et $\mathbf{f}_\ell$ les vecteurs des étiquettes et des scores restreints aux données étiquetées, et $\mathbf{y}_u$ et $\mathbf{f}_u$ les vecteurs des étiquettes et des scores restreints aux données non étiquetées. On peut ainsi voir la solution (3) du problème comme un a priori sur les scores (le premier terme) qui serait corrigé par les données (le second terme). Pour les données non étiquetées, on n'a pas de raison d'avoir d'a priori sur le score, ce qui justifie le choix de $\mathbf{y}_u = 0$. De plus, dans le problème d'optimisation, en ce qui concerne les données non étiquetées, le terme $\|\mathbf{f}_u - \mathbf{y}_u\|^2$ n'a pas pour but d'encourager $\mathbf{f}_u$ à être proche de $\mathbf{y}_u$, mais simplement de tempérer l'amplitude de $\|\mathbf{f}_u\|$, ce qui est le cas avec le choix $\mathbf{y}_u = 0$.

Les scores qui nous intéressent sont ceux des données non étiquetées, qui sont, en définitive :

$$\mathbf{f}_u = \frac{1}{\alpha p}\mathbf{X}_u^\mathsf{T}\left(\mathbf{I}_p - \frac{\mathbf{X}\mathbf{X}^\mathsf{T}}{\alpha p}\right)^{-1}\mathbf{X}_\ell\mathbf{y}_\ell. \qquad (4)$$

2.3 Statistiques de la fonction de décision

Grâce aux outils de la théorie des matrices aléatoires, on peut prédire, à partir des hypothèses 1 et 2, le comportement asymptotique des scores $\mathbf{f}_u$. Ce vecteur joue un rôle central dans le processus de classification, et prédire sa loi nous permet d'anticiper les performances de notre algorithme. On a alors les clés pour optimiser les constantes $\tilde{y}_1$ et $\tilde{y}_2$, ainsi que l'hyperparamètre α, afin de maximiser ces performances.

Théorème 1 *Sous les hypothèses 1 et 2, pour tout $\mathbf{x} \in \mathcal{C}_j$ non étiqueté, f étant le score associé,*

$$f \to \mathcal{N}\left(m_j, \sigma^2\right), \qquad \textit{avec}$$

$$m_j = \mathbf{a}_j^\mathsf{T}\tilde{\mathbf{y}} \qquad \textit{et} \qquad \sigma^2 = \tilde{\mathbf{y}}^\mathsf{T}\mathbf{B}\tilde{\mathbf{y}},$$

avec $\tilde{\mathbf{y}} = (\tilde{y}_1\ \tilde{y}_2)^\mathsf{T}$ et où $\mathbf{a}_j \in \mathbb{R}^2$ et $\mathbf{B} \in \mathbb{R}^{2\times 2}$ sont fonctions des paramètres déterministes du modèle.

Le code permettant de calculer les $\mathbf{a}_j$ et $\mathbf{B}$ est disponible ici. Sans expliquer plus en détail comment les $\mathbf{a}_j$ et $\mathbf{B}$ sont reliés aux paramètres du problème, nous pouvons tout de même regarder quelles sont les implications du Théorème 1 pour notre problème de classification.

Définition 1 *Pour chaque élément non étiqueté $\mathbf{x} \in \mathcal{C}$, la probabilité de mal le classifier est :*

$$\mathbb{P}(\mathbf{x} \to \bar{\mathcal{C}} | \mathbf{x} \in \mathcal{C})$$

où $\mathbf{x} \to \bar{\mathcal{C}}$ signifie que $\mathbf{x}$ a été classifié dans la classe opposée $\bar{\mathcal{C}} \neq \mathcal{C}$. On définit alors la probabilité d'erreur :

$$\epsilon = \frac{\epsilon_1 + \epsilon_2}{2} = \frac{1}{2}\mathbb{P}(\mathbf{x} \to \mathcal{C}_2 | \mathbf{x} \in \mathcal{C}_1) + \frac{1}{2}\mathbb{P}(\mathbf{x} \to \mathcal{C}_1 | \mathbf{x} \in \mathcal{C}_2)$$

Proposition 1 *Il existe un unique (à une constante multiplicative près) vecteur $\tilde{\mathbf{y}}^\star$ qui minimise la probabilité d'erreur ϵ, donné par :*

$$\tilde{\mathbf{y}}^\star = \mathbf{B}^{-1}(\mathbf{a}_2 - \mathbf{a}_1). \tag{5}$$

Proposition 2 *La probabilité d'erreur ϵ minimale (atteinte avec les étiquettes optimales $\tilde{\mathbf{y}}^\star$) est asymptotiquement égale à :*

$$\epsilon^\star = \mathcal{Q}\left(\frac{1}{2}\sqrt{(\mathbf{a}_2 - \mathbf{a}_1)^\mathsf{T}\mathbf{B}^{-1}(\mathbf{a}_2 - \mathbf{a}_1)}\right), \tag{6}$$

où $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{u^2}{2}}\,\mathrm{d}u$.

Si les étiquettes optimales $\tilde{\mathbf{y}}^\star$ sont données explicitement par l'équation (5), ce n'est pas le cas de l'hyperparamètre α. Cependant, comme nous disposons d'une formule explicite donnant la probabilité d'erreur en fonction des paramètres du problème (et en particulier en fonction de α), nous pouvons optimiser α en procédant à une recherche de minimum sur l'intervalle $]\alpha_0, +\infty[$, en utilisant l'expression de la probabilité d'erreur optimale donnée par l'équation (6). Ainsi, nous pouvons optimiser l'hyperparamètre sans recourir à une validation croisée, coûteuse en temps de calcul et en nombre de données.

3 Expériences numériques

Dans cette section, nous simulons des contextes d'apprentissage dans lesquels l'utilisation de données partiellement étiquetées permet d'étendre le modèle semi-supervisé standard. Nous nous restreignons au cas où les données sont générées en simulant des variables gaussiennes suivant les hypothèses 1 et 2. Le code pour l'ensemble des figures présentées est disponible ici.

Avant toute chose, nous allons présenter une simulation qui d'une part valide empiriquement les résultats théoriques du Théorème 1 et d'autre part suggère l'optimalité de notre modèle. La Figure 1 montre la probabilité d'erreur de notre algorithme (que l'on cherche à minimiser) en fonction du logarithme du nombre de données, avec 20% de données étiquetées et 80% de données non étiquetées.

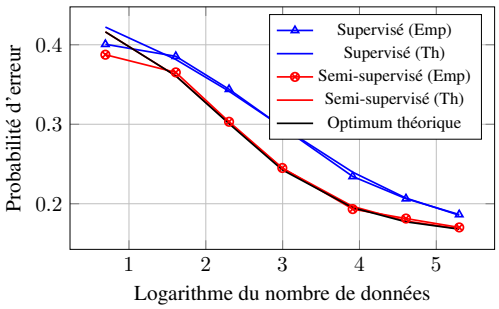

FIGURE 1 : Probabilité d'erreur empirique et théorique en fonction du logarithme du nombre de données. Les prédictions théoriques sont validées par la simulation. L'algorithme semi-supervisé tire pleinement profit des données disponibles, surpassant ainsi l'algorithme supervisé, et ses performances frôlent l'optimum donné par la théorie de l'information, suggérant l'optimalité de notre modèle.

Pour l'ensemble des expériences qui vont suivre, on considérera deux catégories de données étiquetées : les données *certaines*, étiquetées avec une probabilité de 1, et les données *incertaines*, étiquetées avec une probabilité différente de 1 que l'on appellera *fiabilité*. Par exemple, une donnée étiquetée dans la Classe 1 avec une fiabilité de 0.8 aura 80% de chances d'appartenir à la Classe 1 et 20% de chances d'appartenir à la Classe 2. À l'exception de la Figure 2, les figures qui vont suivre présenteront la probabilité d'erreur théorique calculée avec la Proposition 2.

3.1 Risques d'utiliser des données incertaines sans prendre en compte l'incertitude

Dans le cas où des données incertaines sont utilisées avec la même confiance que les données certaines, l'algorithme risque d'être induit en erreur par les données mal étiquetées. Pour mettre cela en évidence, on considère (pour le même jeu de données) trois versions d'étiquetage différentes :

- Un étiquetage *naïf*, où toutes les données étiquetées sont considérées comme certaines, y compris celles qui ne le sont pas. On s'expose alors à de nombreuses erreurs.
- Un étiquetage *adapté*, où les données incertaines sont étiquetées avec la bonne probabilité. C'est le meilleur étiquetage possible étant données les informations dont on dispose.
- Un étiquetage *oracle*, où toutes les données sont certaines. Il s'agit du cas fictif où l'on connaîtrait parfaitement toutes les étiquettes.

La Figure 2 montre la probabilité d'erreur empirique pour ces trois étiquetages, en fonction de la fiabilité des données incertaines. Dans le cas naïf, l'erreur explose à mesure que la fiabilité diminue. A contrario, l'étiquetage adapté permet d'éviter l'effondrement des performances lorsque la fiabilité est faible, et on se rapproche de l'oracle lorsque la fiabilité est grande.

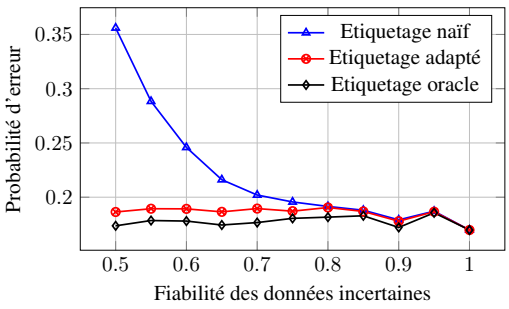

FIGURE 2 : Probabilité d'erreur empirique en fonction de la fiabilité des données. Si l'étiquetage ne prend pas en compte la fiabilité des données, les performances s'effondrent. Un étiquetage adapté permet de garder de bonnes performances, qui se rapprochent de l'oracle lorsque la fiabilité tend vers 1.

3.2 Intérêt d'utiliser des données incertaines plutôt que non étiquetées

Pour éviter toute confusion de l'algorithme, une solution serait simplement de ne pas utiliser les étiquettes des données incertaines, et de considérer ces données comme non étiquetées.

On appellera cette nouvelle version l'étiquetage *restreint*. On perd alors l'information apportée par les étiquettes incertaines. C'est ce que met en évidence la Figure 3, qui montre la probabilité d'erreur en fonction du nombre de données incertaines pour les étiquetages restreint, adapté et oracle. Les données incertaines ont ici une fiabilité de 0.8. Dans le cas de l'étiquetage restreint, cela revient à rajouter des données non étiquetées, ce qui augmente la performance, mais dans une moindre mesure par rapport à un étiquetage adapté.

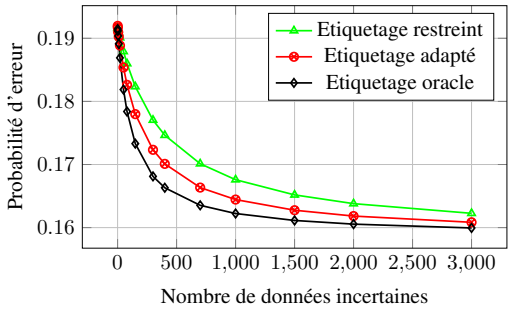

FIGURE 3 : Probabilité d'erreur théorique en fonction du nombre de données incertaines. L'erreur décroît plus vite lorsque la fiabilité des données incertaines est correctement estimée qu'avec un étiquetage qui ne prend pas en compte les données incertaines.

3.3 Jonction entre les données étiquetées et non étiquetées

On se concentre ici sur l'étiquetage adapté. L'un des intérêts d'utiliser des données incertaines est de se passer en partie des données certaines. On peut donc se demander combien de données incertaines sont nécessaires pour pallier l'absence de données certaines. On fixe la fiabilité des données, et on commence avec $n_\ell = 50$ données certaines par classe. À partir de là on calcule, pour un nombre fixé de données certaines supplémentaires, combien de données incertaines supplémentaires seraient nécessaires pour obtenir la même performance qu'avec les données certaines supplémentaires. On constate que ce nombre de données incertaines augmente linéairement en fonction du nombre de données certaines, et que le ratio entre ces deux quantités dépend de la fiabilité (voir code).

Dans la Figure 4, on représente donc le ratio entre ces deux quantités en fonction de la fiabilité. Plus ce ratio est faible, plus les données incertaines ajoutées sont utiles pour la classification. Cette courbe est proposée pour trois valeurs de $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ différentes, la difficulté de la tâche de classification étant d'autant plus faible que $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ est grand. On constate plusieurs choses :

- Les données incertaines sont moins utiles lorsque la difficulté augmente.
- Les données peu fiables (fiabilité< 0.6) apportent peu d'information de plus que les données non étiquetées, et le ratio évolue peu.
- A l'inverse, pour les données plus fiables (fiabilité> 0.6), le ratio diminue sensiblement avec la fiabilité, montrant l'intérêt d'utiliser des données incertaines.

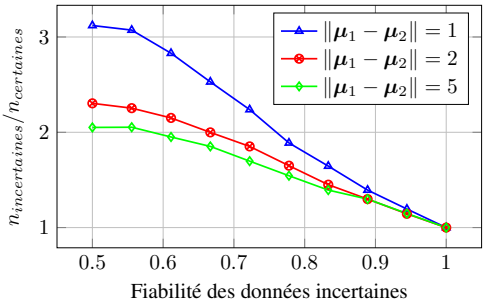

FIGURE 4 : Nombre de données incertaines supplémentaires nécessaires pour obtenir la même performance théorique qu'avec une donnée certaine supplémentaire, en fonction de la fiabilité, et pour différentes difficultés. La fiabilité des données permet de faire la jonction entre les données non étiquetées (*i.e.* de fiabilité 0.5) et les données étiquetées (*i.e.* de fiabilité 1).

4 Conclusion

La méthode d'étiquetage que nous avons étudiée ici permet d'aller plus loin que l'apprentissage semi-supervisé, en dépassant la dichotomie faite usuellement entre les données étiquetées et les données non étiquetées. Grâce aux outils issus de la théorie des matrices aléatoires, nous avons pu faire cela sans sacrifier la simplicité et la flexibilité de notre modèle. Il peut ainsi être utilisé pour des cas d'applications pratiques où l'étiquetage est difficile (pour les données médicales par exemple). A l'inverse des outils d'apprentissage profond, la compréhension théorique du modèle nous permet d'avoir un contrôle précis des performances.

Références

[1] M. Lelarge et L. Miolane *Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting* arXiv preprint arXiv :1907.03792, 2019.

[2] X. Mai et R. Couillet *Consistent Semi-Supervised Graph Regularization for High Dimensional Data* The Journal of Machine Learning Research, 2021.

[3] M. Seddik, C. Louart, M. Tamaazouti et R. Couillet *Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures* arXiv preprint arXiv :2001.08370, 2020.

[4] M. Tiomoko, R. Couillet et H. Tiomoko *Large Dimensional Analysis and Improvement of Multi Task Learning* arXiv preprint arXiv :2009.01591, 2020.

[5] D. Zhou, O. Bousquet, T. Lal, J. Weston et B. Schölkopf *Learning with Local and Global Consistency*, Advances in Neural Information Processing Systems, 2003.

[6] T. Joachims *Transductive Learning via Spectral Graph Partitioning* Proceedings of the Twentieth International Conference on Machine Learning, 2003.

[7] K. Avrachenkov, A. Mishenin, P. Gonçalves et M. Sokol *Generalized Optimization Framework for Graph-based Semi-supervised Learning*, Proceedings of the 2012 SIAM International Conference on Data Mining.