Asymptotic Bayes risk of semi-supervised learning with uncertain labeling

Victor Léger
Laboratoire d'Informatique de Grenoble
Université Grenoble alpes
Grenoble, France
victor.leger@univ-grenoble-alpes.fr

Romain Couillet
Laboratoire d'Informatique de Grenoble
Université Grenoble alpes
Grenoble, France
romain.couillet@univ-grenoble-alpes.fr

***Abstract*—This article considers a semi-supervised classification setting on a Gaussian mixture model, where the data is not labeled strictly as usual, but instead with uncertain labels. Our main aim is to compute the Bayes risk for this model. We compare the behavior of the Bayes risk and the best known algorithm for this model. This comparison eventually gives new insights over the algorithm.**

***Index Terms*—classification, random matrix theory, semi-supervised learning, statistical physics**

I. Introduction

Semi-supervised learning (SSL) is an extension of the conventional supervised learning paradigm by augmenting the (labeled) training data set with unlabeled data, which then "unsupervisably" serve to boost learning performance. SSL has long been considered to be a powerful tool to make use of large amounts of unlabeled data [1].

However, some work also point out the lack of theoretical understanding of these methods [2]–[4]. Even by considering a mere Gaussian mixtures model (which is one of the simplest possible parametric model one could consider for a classification problem), well-known methods such as Laplacian regularization appears to be uneffective to learn from unlabeled data [5].

Fortunately, advances in Random Matrix Theory (RMT) has been exploited to design better methods, by proposing fundamental corrections of known algorithms [6], and even extend them, for instance by considering uncertain labeling [7].

Simultaneously, another field of research has focused on analysing Gaussian mixtures model with statistical physics. Such analysis brings an optimal bound for a given problem, meaning that any possible algorithm cannot reach a better performance [8], [9]. These optimal bounds are a precious tool to understand whether an algorithm has poor performances because of its design or because of the inherent difficulty of the problem it tries to solve.

Therefore, the objectives of this article are twofolds :

- Compute the Bayes risk in the case of uncertain data labeling, inspired by the work of [9].
- Use the knowledge of this optimal bound to further understand the behavior of the algorithm of [7], which performances have been proven to be close to the optimal bound.

For simplicity reasons, the model presented in this article is a single-task model, but it is worth to note that most of the conclusions remain true in a multi-task setting, as the previous works it is based on are multi-task models [7], [9].

The remainder of the article is organized as follows. Section II introduces the model, the assumptions and the aim of the next sections. Section III states our main theorem, and gives interpretation of this theorem. Section IV gives a succinct proof of the main theorem. Finally, Section V displays simulations of both the optimal bound and the algorithm presented in [7].

II. Model and Main Objective

We consider a semi-supervised binary classification task with training samples $\mathbf{X} = [\mathbf{X}_\ell, \mathbf{X}_u] \in \mathbb{R}^{p\times n}$ which consists of a set of n_ℓ labeled data samples $\mathbf{X}_\ell = \{\mathbf{x}_i\}_{i=1}^{n_\ell}$ and a set of n_u unlabeled data points $\mathbf{X}_u = \{\mathbf{x}_i\}_{i=n_\ell+1}^{n}$. Each labeled data point $\mathbf{x}_i$ has an associated couple (d_{i1}, d_{i2}) of *pre-estimated* probabilities that the vector belongs to one class or the other, such that $d_{i1}+d_{i2} = 1$. The goal of the classification task is to predict the genuine class of unlabeled data $\mathbf{X}_u$. In this context, we are interested in computing the Bayes risk of the classification task, *i.e.,* the minimal classification error achievable for each unlabeled sample $\mathbf{x}_i$ with the available data :

$$\inf_{\hat{y}_i} \mathbb{P}(\hat{y}_i \neq y_i) \tag{1}$$

where $\hat{y}_i = \mathbb{E}[y_i|\mathbf{X}]$ is the label prediction made for the sample $\mathbf{x}_i$.

Assumption 1 (On the data distribution) *The columns of the data matrix $\mathbf{X}$ are independent Gaussian random variables. Specifically, the data samples $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are i.i.d. observations such that $\mathbf{x}_i \in \mathcal{C}_j \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{I}_p)$ where $\mathcal{C}_j^t$ denotes the Class j. We assume that the number of data in each class is the same. We further define the quantity $\lambda = \frac{1}{4}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, which is called the* signal to noise ratio *(SNR).*

We study our model in a large dimensional setting, where the dimension and the amount of data have the same order of magnitude, which is practically the case with modern data.

Assumption 2 (Growth Rate) *As $n \to \infty$:*

- $p/n \to c > 0$
- $n_\ell/n \to \eta$

With our notations, in a single-task setting, and assuming that the probability couples of labeled data are either $(0,1)$ or $(1,0)$ (*i.e.,* the data is labeled with complete certainty), it has been proved in [9] that under the previous assumptions, as $p \to \infty$, the Bayes risk converges to

$$\mathcal{Q}(\sqrt{q_u}), \tag{2}$$

where $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} \,\mathrm{d}u$, and the couple (q_u, q_v) satisfies the following equations:

$$q_u = \lambda \frac{\lambda c q_v}{1 + \lambda c q_v} \tag{3}$$

$$q_v = \eta + (1-\eta) F(q_u), \tag{4}$$

with $F(q) = \mathbb{E}\left[\tanh(\sqrt{q} Z + q)\right]$, $Z \sim \mathcal{N}(0,1)$.

Our goal in the remainder of this article is to derive an equivalent result in the more general case where data is not labeled with certainty. Let us define, for each datapoint $\mathbf{x}_i$, $\varepsilon_i = d_{i2} - d_{i1} \in [-1, 1]$. This quantity is enough to charaterize the couple of probabilities, as $d_{i1} + d_{i2} = 1$. We observe that $|\varepsilon| = 1$ means that the data is labeled with certainty in a class, while $\varepsilon = 0$ means that the data is unlabeled.

To get equation (4), it is needed to compute the quantity $\hat{y}_i = \mathbb{E}\left[y_i | \mathbf{X}\right]$, which is the estimation of y_i with the available data $\mathbf{X}$. In the proof (presented in Section IV), a trick allows to compute this quantity as a function of q_u, and the labeling information is expressed through the prior distribution of y.

- For data labeled with certainty, it is know that $y_i = -1$ or $+1$, so the prior is either $\delta(t-1)$ or $\delta(t+1)$.
- For unlabeled data, the prior is uniform over $\{-1, +1\}$. Or equivalently, the distribution function is
$$\frac{1}{2}\delta(t-1) + \frac{1}{2}\delta(t+1).$$
- When the data is not labeled with certainty but with a probability couple (d_{i1}, d_{i2}), the prior distribution of y_i becomes
$$d_{i1}\delta(t-1) + d_{i2}\delta(t+1) \tag{5}$$

The following section states our main theorem using this last prior distribution.

III. Main Results

Theorem 1 *Under the previous assumptions, as $p \to \infty$,*

- *The Bayes risk converges to*
$$\mathcal{Q}(\sqrt{q_u}), \tag{6}$$
where $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} \,\mathrm{d}u.$
- *The overlaps q_u, q_v satisfy the following equations*
$$q_u = \lambda \frac{\lambda c q_v}{1 + \lambda c q_v} \tag{7}$$
$$q_v = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} F_{\varepsilon_i}(q_u) \tag{8}$$

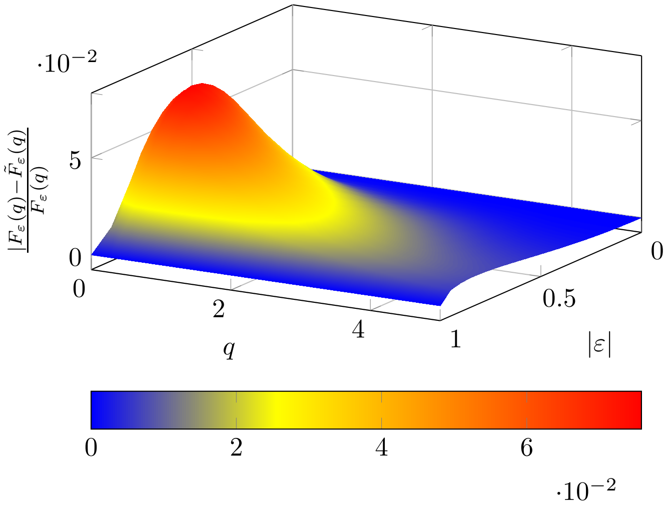

Fig. 1. Relative error of the approximation $\tilde{F}_\varepsilon(q) \simeq F_\varepsilon(q)$. The error is at most 7%, and shrinks for either $\varepsilon = 0$, $\varepsilon = 1$ or large q

with $F_\varepsilon(q) = \mathbb{E}\left[\psi_\varepsilon(q + \sqrt{q} Z)\right]$, $Z \sim \mathcal{N}(0,1)$ *and*

$$\psi_\varepsilon(t) = \frac{tanh(t) + \varepsilon^2\left(1 - tanh(t) - tanh^2(t)\right)}{1 - \varepsilon^2 tanh^2(t)}.$$

A sketch of the proof of Theorem 1 is given in Section IV. The function F_ε is similar to the previous function F, but the expression of ψ_ε is not easy to understand as it is.

Remark 1 *The function ψ_ε can be put in the following (more convenient) form :*

$$\psi_\varepsilon(t) = tanh(t) + \varepsilon^2\left(1 - tanh(t)\right) \\ - (1-\varepsilon^2)(1 - tanh(t)) \sum_{k \geq 1} \varepsilon^{2k} tanh^{2k}(t).$$

Remark 2 *The function F_ε can be approximated by :*

$$\tilde{F}_\varepsilon(q) = \mathbb{E}\left[\tilde{\psi}_\varepsilon(q + \sqrt{q} Z)\right], \tag{9}$$

with $Z \sim \mathcal{N}(0,1)$ and

$$\tilde{\psi}_\varepsilon(t) = tanh(t) + \varepsilon^2(1 - tanh(t)). \tag{10}$$

Figure 1 gives an idea of the quality of the approximation made in Remark 2. However, it is worth to note that its purpose is not to replace the original formula from Theorem 1, as that formula is already tractable and can be computed easily. Instead, Remark 2 intend to bring a simpler formula that conveys an understanding of the key role of quantity ε^2.

As q_u and q_v are related to each other, one could also worry that a small error in equation (8) could lead to a completely different solution for q_u and q_v. Fortunately, if one replaces the solution $(q_u^\star, q_v^\star)$ of the system by another solution $(q_u^\star + \Delta q_u, q_v^\star + \Delta q_v)$, then we have $|\frac{\Delta q_u}{q_u^\star}| \leq |\frac{\Delta q_v}{q_v^\star}|$. This means that a small variation of q_v leads to an even smaller variation of q_u.

Corollary 1 *With the previous approximation of the function F_ε, one can approximate the equation* (8) *:*

$$q_v \simeq \bar{\varepsilon}^2 + (1-\bar{\varepsilon}^2)F(q_u) \tag{11}$$

with

$$\bar{\varepsilon}^2 = \frac{1}{n}\sum_{i=1} {\varepsilon_i}^2$$
$$F(q) = \mathbb{E}\left[tanh(q+\sqrt{q}Z)\right]$$
$$Z \sim \mathcal{N}(0,1)$$

Proof: The function $\tilde{F}_{\varepsilon_i}$ described in Remark 2 can be expressed with function F :

$$\begin{aligned}\tilde{F}_{\varepsilon_i}(q) &= \mathbb{E}\left[\tanh(q+\sqrt{q}Z) + {\varepsilon_i}^2(1-\tanh(q+\sqrt{q}Z))\right] \\ &= {\varepsilon_i}^2 + \mathbb{E}\left[\tanh(q+\sqrt{q}Z)\right](1-{\varepsilon_i}^2) \\ &= {\varepsilon_i}^2 + (1-{\varepsilon_i}^2)F(q)\end{aligned}$$

By mixing the results of Theorem 1 and Remark 2, one gets asymptotically

$$\begin{aligned}q_v &\simeq \frac{1}{n}\sum_{i=1}^{n}\tilde{F}_{\varepsilon_i}(q_u) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[{\varepsilon_i}^2 + (1-{\varepsilon_i}^2)F(q)\right] \\ &= \bar{\varepsilon}^2 + \left(1-\bar{\varepsilon}^2\right)F(q)\end{aligned}$$

Corollary 1 enables an easy interpretation of Theorem 1. Indeed, the value of q_v given by equation (11) is similar to equation (4), with $\eta = \bar{\varepsilon}^2$. One can check that :

- $\varepsilon^2 = 0 \leftrightarrow$ unlabeled data $\leftrightarrow \eta = 0$
- $\varepsilon^2 = 1 \leftrightarrow$ data labeled with certainty $\leftrightarrow \eta = 1$

If all samples are labeled with the same value ε, then it is equivalent to a task for which one would have a proportion ε^2 of data labeled with certainty and a proportion $1-\varepsilon^2$ of unlabeled data.

To go further, $F(q_u)$ can be understood as a quantity that expresses how useful unlabeled data are, relatively to labeled data. Indeed, if $F(q_u) = 1$, unlabeled data brings as much information as labeled data. Interestingly, this quantity $F(q_u)$ only depends on q_u, which itself related to the Bayes risk of the classification task. Thus, usefulness of unlabeled data only depends on how well the task can be performed. Figure 2 displays the quantity $F(q_u)$ as a function of Bayes risk.

IV. Sketch of the proof

The proof of Theorem 1 is really similar to the one performed in [9], as (2) and (3) remain the same, but the main difference lays in the expression of q_v, that must be adapted. As in the original proof, $q_v = \langle\hat{\mathbf{y}}, \mathbf{y}\rangle$ is the *overlap* of the signal $\mathbf{y} = (y_i)_i$, and we have asymptotically, through the law of large numbers :

$$q_v = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\hat{y}_i y_i\right] \tag{12}$$

where $\hat{y}_i = \mathbb{E}\left[y_i|\mathbf{X}\right]$ is the MMSE estimator of y_i.

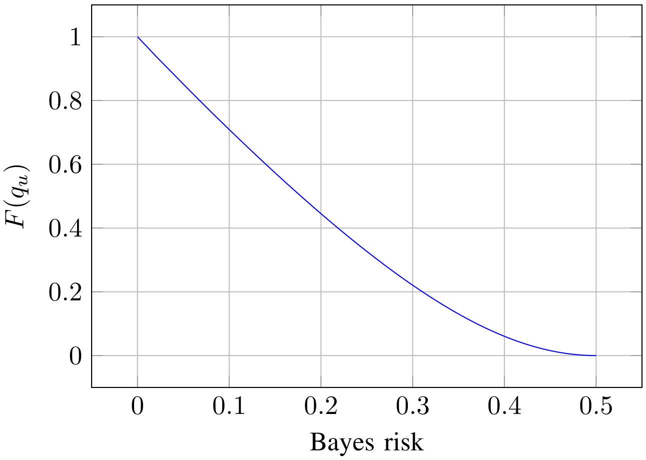

Fig. 2. Usefulness of unlabeled data as a function of the Bayes risk of the task. Interestingly, the only criterion to determinate the effectiveness of unlabeled data is how solvable the task is. The lower the Bayes risk is, the more unlabeled data are useful to perform the task.

Then, a key lemma for the proof is the following.

Lemma 1 *Estimating y_i from $\mathbf{X}$ is asymptotically equivalent to estimating the signal y_i from the output of a Gaussian channel with SNR q_u. Let us consider the following Gaussian channel*

$$U_i = \sqrt{\lambda}S_i + Z_i$$

with $\lambda = q_u$ the SNR, S_i the signal and $Z_i \sim \mathcal{N}(0,1)$. Then computing the overlap $\mathbb{E}\left[\hat{y}_i y_i\right]$ of y_i is equivalent to computing the overlap $\mathbb{E}\left[\hat{S}_i S_i\right]$ of S_i, with $\hat{S}_i = \mathbb{E}\left[S_i|U_i\right]$

Thus, one has

$$\mathbb{E}\left[\hat{y}_i y_i\right] = \mathbb{E}\left[\hat{S}_i S_i\right] \tag{13}$$

The signal S_i follows the same distribution than y_i :

$$S_i \sim \begin{cases} -1 \text{ with probability } d_{i1}, \\ +1 \text{ with probability } d_{i2} \end{cases}$$

$$\mathbf{E}\left[S_i|U_i\right] = \frac{d_{i1}e^{\sqrt{\lambda}U_i} - d_{i2}e^{-\sqrt{\lambda}U_i}}{d_{i1}e^{\sqrt{\lambda}U_i} + d_{i2}e^{-\sqrt{\lambda}U_i}}$$

Using $\varepsilon_i = d_{i1} - d_{i2}$, one gets $\mathbf{E}\left[S_i|U_i\right] = f_{\varepsilon_i}(\sqrt{\lambda}U_i)$, with

$$f_\varepsilon(t) = \frac{\tanh(t)+\varepsilon}{1+\varepsilon\tanh(t)}$$

$$\begin{aligned}&\mathbf{E}\left[S_i\mathbf{E}\left[S_i|U_i\right]\right] = \mathbf{E}\left[S_i f_{\varepsilon_i}(\lambda S_i + \sqrt{\lambda}Z_i)\right] \\ &= d_{i1}\mathbf{E}\left[f_{\varepsilon_i}(\lambda + \sqrt{\lambda}Z_i)\right] - d_{i2}\mathbf{E}\left[f_{\varepsilon_i}(-\lambda + \sqrt{\lambda}Z_i)\right] \\ &= d_{i1}\mathbf{E}\left[f_{\varepsilon_i}(\lambda + \sqrt{\lambda}Z_i)\right] - d_{i2}\mathbf{E}\left[f_{\varepsilon_i}(-(\lambda + \sqrt{\lambda}Z_i))\right] \\ &= \mathbf{E}\left[\psi_{\varepsilon_i}(\lambda + \sqrt{\lambda}Z_i)\right]\end{aligned} \tag{14}$$

with $\psi_{\varepsilon_i}(t) = d_{i1}f_{\varepsilon_i}(t) - d_{i2}f_{\varepsilon_i}(-t)$

More precisely,

$$\begin{aligned}\psi_{\varepsilon_i}(t) &= d_{i1}\frac{\tanh(t)+\varepsilon_i}{1+\varepsilon_i\tanh(t)} - d_{i2}\frac{-\tanh(t)+\varepsilon_i}{1-\varepsilon_i\tanh(t)}\\ &= (d_{i1}+d_{i2})\frac{\tanh(t)-\varepsilon_i{}^2\tanh(t)}{1-\varepsilon_i{}^2\tanh^2(t)}\\ &+ (d_{i1}-d_{i2})\frac{\varepsilon_i(1-\tanh^2(t))}{1-\varepsilon_i{}^2\tanh^2(t)}\\ &= \frac{\tanh(t)+\varepsilon_i{}^2\left(1-\tanh(t)-\tanh^2(t)\right)}{1-\varepsilon_i{}^2\tanh^2(t)}\end{aligned}$$

Combining (12), (13) and (14), we obtain (8).

V. Simulations and Applications

The objective of this section is to confront theoretical results of Section III and the algorithm described in [7], which will be from now on refered to as *optimal algorithm*. The common idea of the following experiments is that the optimal algorithm is expected to behave similarly to the optimal bound studied in Section III.

As we have seen in Section III, different labeling settings can lead to the same value of $\bar{\varepsilon}^2$. Let us start with only unlabeled data and data labeled with certainty. Then,

$$\bar{\varepsilon}^2 = \eta, \tag{15}$$

and we obtain a classification error E. Now, let us assume that all the labeled data is labeled with the same confidence $\kappa < 1$. The total number of data n stays unchanged. For different values of κ, if one wants to achieve the same error E, then more labeled data will be needed as κ decreases. Indeed, reaching the same performance means obtaining the same q_u, and therefore q_v, as the task does not change beyond that. We recall that $q_v \simeq \bar{\varepsilon}^2 + (1-\bar{\varepsilon}^2)F(q_u)$, and in our context, $F(q_u)$ does not change. Consequently, to obtain the same error E, $\bar{\varepsilon}^2$ must stay constant. Moreover, we have

$$\bar{\varepsilon}^2 = \frac{n_\ell}{n}(2\kappa-1)^2 \tag{16}$$

By combining (15) and (16), one gets

$$n_\ell = \frac{\eta}{(2\kappa-1)^2}n \tag{17}$$

For a given value of η, $(2\kappa-1)^2$ must be no smaller than η, otherwise even a fully labeled dataset would not allow to reach $\bar{\varepsilon}^2 = \eta$.

Figure 3 displays both empirical and theoretical values of n_ℓ as a function of κ in this setting. The theoretical value is given by (17), and the empirical one is computed by incrementing n_ℓ (and decrementing n_u) until the error given by the optimal algorithm gets below E. The match between empirical and theoretical curves show that Corollary 1 helps to understand the behavior of the algorithm.

The other takeaway message of Section III is that the usefulness of unlabeled data, expressed through the quantity $F(q_u)$, only depends on the Bayes risk of the task, and therefore the final classification error of the optimal algorithm. In order to understand the contribution of unlabeled data, one could be interested in computing the reduction of the classification error by using the semi-supervised version of the optimal algorithm instead of the fully supervised one. With the aim in mind, we will consider the two following quantities:

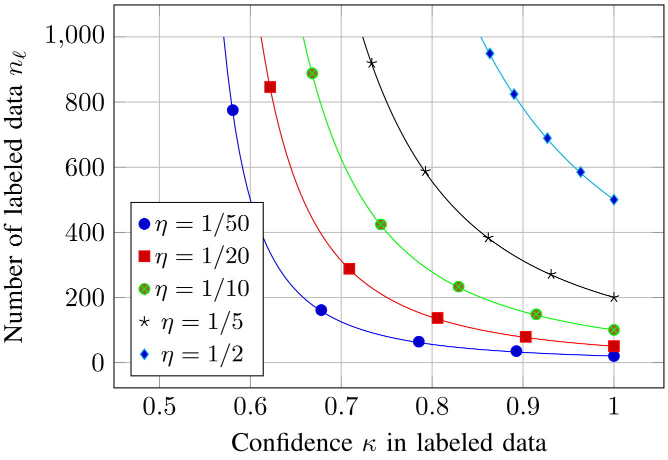

Fig. 3. Number of labeled data n_ℓ needed to perform the same performance, as a function of the confidence in the data labeling, for different values of η ($n = 1000, p = 200, \lambda = 0.25$). The empirical values are displayed in dots, and theoretical prediction (built on the results of Section III) in plain line. The least reliable the data is, the more data is needed to reach the same performance.

- The absolute error reduction
$$\frac{E_{\text{sup}} - E_{\text{semi-sup}}}{E_{\text{sup}}} \tag{18}$$
which is the tangible error reduction one can expect by adopting the semi-supervised method instead of the fully supervised one.
- The error reduction relatively to oracle bayes risk
$$\frac{E_{\text{sup}} - E_{\text{semi-sup}}}{E_{\text{sup}} - E_{\text{oracle}}} \tag{19}$$
which reflects how much of the way to oracle error has been done by adopting the semi-supervised method instead of the fully supervised one,

where E_{oracle} is the bayes risk one can expect when the centers of distributions $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are known. More precisely, oracle error is given by the formula

$$E_{\text{oracle}} = \mathcal{Q}(\sqrt{\lambda}). \tag{20}$$

Leaving out the parameter η, the main parameters that drive the final error are λ and c, as we can see in (7). Therefore, Figures 4 and 5 display the two kinds of error reduction presented above, respectively as functions of λ and c.

In Figure 4, it is clear that the error reduction is higher when λ grows, for both types of error reduction. Intuitively, a higher SNR means a lower final error, and consequently a higher contribution of unlabeled data to the classification.

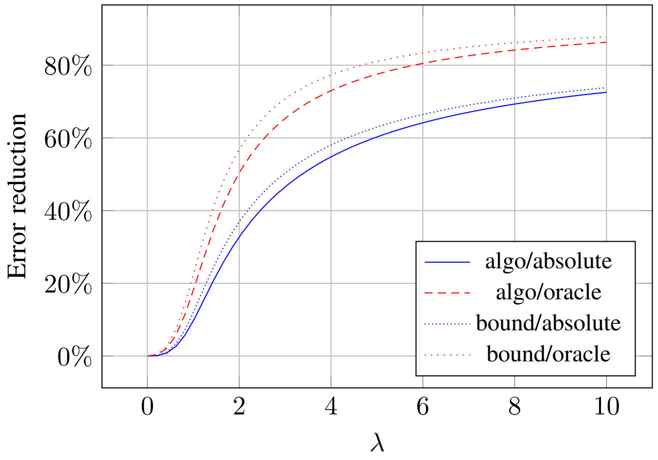

Fig. 4. Percentage of error reduction by using the semi-supervised algorithm instead of the supervised one, as a function of the SNR λ ($n = p = 200, \eta = 0.2$). The easier the task is, the higher the semi-supervised contribution is, because the classification error is lower.

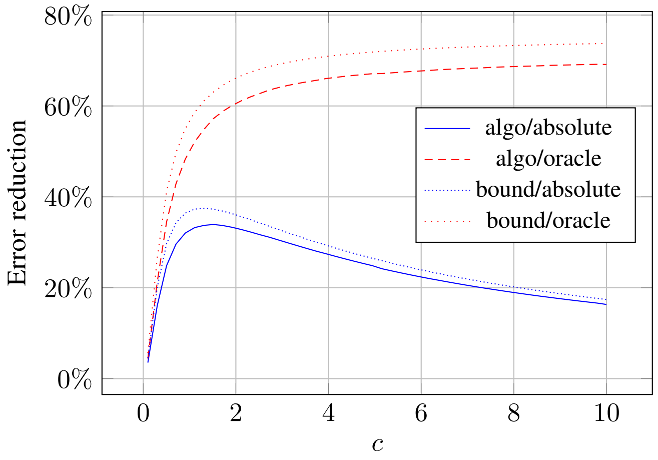

Fig. 5. Percentage of error reduction by using the semi-supervised algorithm instead of the supervised one, as a function of the ratio $c = n/p$ ($\lambda = 2, p = 200, \eta = 0.2$). As c grows, the semi-supervised algorithm is more and more effective comparatively to the supervised one, relatively to oracle error, because the classification error is lower and oracle error stays constant. However, if the oracle error ceases to be the reference, then the contribution of semi-supervised decreases for high values of c, because both algorithms edge closer to the oracle bound, which stays far from zero.

Meanwhile in Figure 5, the two types of error reduction do not behave similarly. In this case, the oracle error is constant, and both errors E_{sup} and $E_{\text{semi-sup}}$ get close to E_{oracle} as c grows. Therefore, there is not much to gain by adopting the semi-supervised algorithm, as we are already close to the oracle bound with the supervised one. However, the error reduction relatively to oracle still increases when c grows. We see that the understanding of what remains to be improved plays a key role in Figure 5.

VI. Conluding remarks

Figuring out the link between the performances of an algorithm and its optimal bound gives precious insights. In our case, the algorithm behaves similarly to its optimal bound, giving strong insight that the algorithm is indeed near optimal. So if the algorithm gives poor performances, it simply means that the problem is inherently too hard to solve. Therefore, the interest of computing such optimal bounds is clear. By knowing in advance how far from optimal an algorithm is, one can avoid spending too much energy to solve a problem which turns out to be a dead-end.

Furthermore, the similarity of behavior between the algorithm and the bound allows to understand the algorithm from an other perspective. Indeed, results from Sections III and V provide a new understanding on when semi-supervised learning is truly useful, and when it is not.

References

[1] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. The MIT Press, 09 2006. [Online]. Available: https://doi.org/10.7551/mitpress/9780262033589.001.0001

[2] B. Shahshahani and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, 1994.

[3] F. G. Cozman and I. Cohen, "Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers," in *Semi-Supervised Learning*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:63547716

[4] S. Ben-David, T. Lu, and D. Pál, "Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning," in *Annual Conference Computational Learning Theory*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:7670149

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

[6] ——, "Consistent semi-supervised graph regularization for high dimensional data," *Journal of Machine Learning Research*, vol. 22, no. 94, pp. 1–48, 2021. [Online]. Available: http://jmlr.org/papers/v22/19-081.html

[7] V. Leger and R. Couillet, "A large dimensional analysis of multi-task semi-supervised learning," 2024.

[8] M. Lelarge and L. Miolane, "Asymptotic bayes risk for gaussian mixture in a semi-supervised setting," 2019.

[9] M.-T. Nguyen and R. Couillet, "Asymptotic bayes risk of semi-supervised multitask learning on gaussian mixture," 2023.