Word Representations Concentrate and This is Good News!

Romain Couillet[1] and **Yagmur Gizem Cinar**[2] and **Eric Gaussier**[2] and **Muhammad Imran**[2]
LargeDATA chair, GIPSA-lab, University Grenoble-Alpes, Grenoble, France
LIG-lab, University Grenoble-Alpes, Grenoble, France

Abstract

This article establishes that, unlike the legacy tf*idf representation, modern natural language representations (word embedding vectors) exhibit a so-called *concentration of measure phenomenon*, in the sense that, as the representation size p and database size n are both large, their behavior is similar to that of large dimensional Gaussian random vectors. This phenomenon has profound consequences: machine learning algorithms for natural language data become tractable and amenable to improvement, thereby establishing first solid theoretical results in the field of natural language processing.

1 Introduction

The latest breakthroughs in machine learning, and in particular the recent emergence of algorithms reaching super-human performance triggered by the deep learning revolution, strongly suggest that "bigger is better". Specifically, while the first learning methods (the basic perceptron to start with) were deemed unstable and unreliable, modern algorithms able to handle large amounts of *large dimensional* data and hyperparameters are now extremely stable – even though formally proving this statement is still a theoretical riddle. But the large dimensional nature of data alone cannot explain the stability of algorithms; one may in particular smartly devise intricate data models which no modern algorithm could resolve. From a half philosophical-half mathematical standpoint, following Occam's razor's principle, (Lin et al., 2017) proposes that *natural data adhere to the laws of physics and the laws of physics are simple*, and that this very fact explains to a large extent the success of modern deep learning.

From a purely mathematical perspective then comes the question of a relevant model to analyze natural data and algorithms, sufficiently generic to encompass a broad spectrum of real data of different nature (images, video, sound, and maybe *text?*) but sufficiently focused for mathematical tractability and practical usefulness, and in particular able to explain the stability of modern statistical learning methods.

In a recent line of works, Couillet and co-authors suggest and theoretically support that *concentrated random vectors* may hold the answer. By definition, a concentrated random vector $\mathbf{x} \in \mathbb{R}^p$ is a vector which satisfies a concentration of measure phenomenon in the sense of (Ledoux, 2001): in essence, concentration means that $\mathbf{x}$ does *not* converge (quite the opposite) but any *scalar Lipschitz observation* $g(\mathbf{x}) \in \mathbb{R}$ of $\mathbf{x}$ converges around its statistical mean when the size p of $\mathbf{x}$ increases; Figure 1 schematically illustrates the concentration of measure phenomenon. In particular, a key property to the present article is that the distance between any two concentrated random vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ with "nice properties" converges to a constant value, which *only depends* on the data statistics, and is in particular independent of their random realization. This fundamental phenomenon, not true for small data, is at the core of our present study.

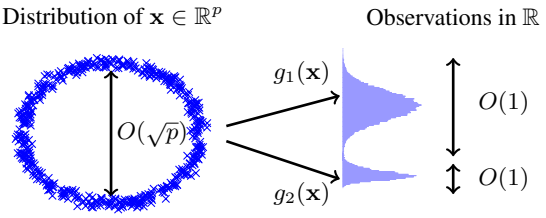

Figure 1: A visual interpretation of the concentration of measure phenomenon. **(Left)** Schematics of 500 realizations of p-dimensional Gaussian random vectors $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_p)$ (concentrated by definition). **(Right)** Concentration of two Lipschitz functionals ($g_1(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_p/\sqrt{p}$ and $g_2(\mathbf{x}) = \|\mathbf{x}\|_\infty$). While $\mathbf{x}$ "spreads out" in its ambient space, $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ converge.

In detail, Couillet and co-authors first demonstrated in (Seddik et al., 2020) that *natural images* and their modern representations (such as VGG, ResNet embeddings) can be appropriately modelled by concentrated random vectors: they precisely prove that the extremely realistic images produced by modern generative adversarial networks (GANs) are by definition concentrated random vectors. Besides, in (Louart and Couillet, 2018), the authors establish a *universality result* which proves that the performance of many machine learning algorithms – from support vector machines to (kernel) spectral clustering – applied to concentrated random data is asymptotically[1] the same as if the data had been Gaussian random vectors with the same first and second order statistics. These findings have tremendous consequences to modern machine learning: they in particular ensure, for the first time, that even involved algorithms applied to real data are analytically tractable, and that their performances can be anticipated and improved offline (without the need for cross-validation).

As (possibly large dimensional) vector representations of words and documents have become a basic building block of many natural language processing methods (Turney and Pantel, 2010), in particular since the success of word embeddings such as *word2Vec* (Mikolov et al., 2013) and *Glove* (Pennington et al., 2014), two natural questions arise: (i) do word (and document) representations exhibit concentration of measure phenomena?, and (ii) do some of the aforementioned findings on real images extend to words and textual documents?

The present article empirically investigates this question and claims to reach a positive answer.[2] Specifically, the main contributions of the article are as follows:

1. We empirically establish that modern word embedding representations can suffer a *distance concentration phenomenon*, typical of concentrated random vectors but usually considered as a manifestation of the curse of dimensionality;

2. We empirically confirm that these word embeddings, unlike tf*idf vectors, exhibit a *universality phenomenon* in the following sense: letting $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n words or document representations of dimension p, the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = f(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ for some smooth function f has the same behavior (entry-wise and spectral) as a matrix $\mathbf{K}'$ built out of Gaussian random vectors $\mathbf{x}'_i$ having the same statistical mean and covariance as the original data.[3]

3. As a concrete application, the classification performances achieved by a kernel (least-square) support vector machine applied to classes of documents of popular databases are shown to be theoretically predictable and to match the theory established on mere Gaussian random vectors, thereby confirming the universality property of word embedding representations and the possibility to use a simple Gaussian vector theory to predict the performance of machine learning algorithms for natural language processing.

Related works. Several works similarly tried to reinterpret word embeddings, either in terms of matrix factorization (Levy and Goldberg, 2014b) or latent models (Arora et al., 2016), and to account for the associations and analogies typical of the linear behavior of these embeddings (Levy and Goldberg, 2014a; Bolukbasi et al., 2016; Gittens et al., 2017; Ethayarajh et al., 2019a,b; Allen and Hospedales, 2019). In a different line of research, many attempts were made to understand the syntactic and semantic generalization capabilities of different deep learning models based on word embeddings, as in (Dessì and Baroni, 2019; Hewitt and Manning, 2019; Lakretz et al., 2019; Chi et al., 2020) to list a few. Our approach is however different in its trying to *statistically model* word embeddings so to grasp the behavior of related machine learning algorithms. To the best of the authors' knowledge, this the first time this original approach is being investigated.

2 Preliminaries and first observations

2.1 Asymptotics of learning

From a crude viewpoint, machine learning algorithms may be seen as functionals $F_\theta : \mathbb{R}^{p \times n} \times$

[1] In the limit of large number and dimension of the data.

[2] Recent *contextualized* word embeddings, such as BERT (Devlin et al., 2019), cannot be reasonably used without fine tuning. These are not considered in the present study, even though we do believe our conclusions also apply to them, a point to be investigated in future extensions.

[3] Those means and covariances being evaluated empirically from words and documents of a common class.

$\mathbb{R}^p \to \mathbb{R}$, $(\mathbf{X}, \mathbf{x}) \to F_\theta(\mathbf{X}, \mathbf{x})$ which, for an input training data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and a test datum $\mathbf{x}$ returns a soft scalar score or hard decision. Here θ accounts for the possible hyper-parameter vector used to fine-tune the algorithm. Assuming the training dataset $\mathbf{X} \in \mathbb{R}^{p \times n}$ to be a random matrix with some prescribed distribution (and similarly for $\mathbf{x}$), evaluating the performances of F_θ boils down to establishing the statistics of the *random variable* $F_\theta(\mathbf{X}, \mathbf{x})$. This has long been a cumbersome, if not impossible, task which has mainly been studied so far using the asymptotic statistics $n \to \infty$ and p fixed. Yet, these results have long remained of little use, not very expressive, and of limited interest when n is not much larger than p; this being in particular due to the non-linear (and often even implicit) nature of F_θ. Random matrix theory and statistical physics have recently changed this paradigm and managed to break the non-linearity barrier by showing that, as $n, p \to \infty$ simultaneously (thereby mimicking the modern large and numerous data setting), the performances of many non-trivial learning algorithms become tractable since they *converge*, as $n, p \to \infty$, to some deterministic limits.

These latest results are based on sufficiently "stable" random models for $\mathbf{X}$ (and $\mathbf{x}$): statistical physics uses isotropy and symmetries, which however often reduces to standard Gaussian data assumptions, while random matrix theory is richer and has lately exploited the *Lipschitz stability* offered by *concentrated random vector* models (Louart and Couillet, 2018). By definition, a random vector $\mathbf{z}$ in a vector space $\mathcal{S}$ is concentrated if, for all 1-Lipschitz functional $g : \mathcal{S} \to \mathbb{R}$, we have that for all $\varepsilon > 0$,

$$\mathbb{P}\left(|g(\mathbf{z}) - m_g| > \varepsilon\right) \leq Ce^{-c\varepsilon^2}$$

for some constant $C, c > 0$ and m_g a median of $g(\mathbf{z})$. That is, $\mathbf{z}$ itself may not converge in any usual sense (in general it does not: for instance $z \sim \mathcal{N}(0, \mathbf{I}_p)$ is concentrated but does not converge) but its Lipschitz functionals, also called *observations of* $\mathbf{z}$ do converge (e.g., $\frac{1}{\sqrt{p}}\|\mathbf{z}\| \to 1$ almost surely). Recall Figure 1 for a visual intuition. Concentrated random vector modelling is particularly convenient as it ensures that, if $\mathbf{X}$ is, say, a concentrated random matrix, then for any Lipschitz function G (that outputs either small or large dimensional data), $G(\mathbf{X})$ is still concentrated and in particular functionals $G : \mathbb{R}^{p \times n} \to \mathbb{R}$ are such that $G(\mathbf{X})$ almost surely converges. This is even extended in (Louart and Couillet, 2018) to non-Lipschitz operators G, and consequently to a wide range of machine learning "algorithms" F_θ.

Most crucially, it is proved in (Louart and Couillet, 2018) that, for a rich family of functionals F_θ, if $\mathbf{X}$ and $\mathbf{x}$ are concentrated, not only does $F_\theta(\mathbf{X}, \mathbf{x})$ converge, but it converges *to the same limit* as $F_\theta(\mathbf{X}', \mathbf{x}')$ for $\mathbf{X}'$ and $\mathbf{x}'$ random Gaussian matrix and vector having the same statistics (mean and covariance) as $\mathbf{X}$ and $\mathbf{x}$, respectively. This is a classical but fundamental result in random matrix theory, referred to as *universality*.

Remark 1 (When are n, p large enough?)**.** *If random matrix theory predicts the asymptotic convergence of algorithms as $n, p \to \infty$, these results are only useful if, in practice, n and p need not be extremely large. As a matter of fact, and quite surprisingly, the* large dimensional effects arise very rapidly *so that, in practice, n, p of the order of hundreds (sometimes even tens) is enough for an asymptotic behavior to emerge. This is explained by the numerous ($O(np)$) degrees of freedom inherent to the data which in particular induce rates of convergence, e.g., central limit theorems, at speed $1/\sqrt{np}$ instead of $1/\sqrt{n}$ when $n \to \infty$ alone. Word embedding vectors, of size $p \sim 100$ or more, natural enter this regime.*

2.2 How to testify of a concentration of measure phenomenon?

With this introductory overview in mind naturally arises the question of the relevance of a concentrated random vector modelling for practical data. As pointed out in the introduction, the synthetic images produced by GANs (Goodfellow et al., 2014) are by definition concentrated random vectors: this is because they are bounded Lipschitz functions (the Lipschitz operator being the pre-trained neural network) of a Gaussian random vector which is itself concentrated. Genuine images being so well approximated by GAN synthetic images, this strongly suggests that real images can be modelled as concentrated random vectors, which is confirmed by simulation results (Seddik et al., 2020).

But words and documents are so far not reliably produced by GANs and it is unclear whether they might embrace the concentration of measure phenomenon. The objective of the article is to empirically confirm that most of the pregnant phenomena occurring in concentration random vectors,

namely the convergence of distances between distinct vectors and the (Gaussian-like) universality behavior, are indeed met by word and document representations.

2.3 Concentration of distances, and kernel spectrum

2.3.1 Concentration of distance

A first phenomenon arising in concentrated random vectors, which disrupts standard machine learning intuition, is the *convergence of distances phenomenon*. Specifically, if $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are i.i.d. concentrated random vectors with $\mathbf{C} \equiv \mathrm{Cov}(\mathbf{x}_i)$ of bounded spectral norm, then, as $p, n \to \infty$ in such a way that n grows no more than polynomially with p (e.g., p/n is constant),

$$\max_{1 \leq i \neq j \leq n} \left| \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau_p \right| \to 0 \qquad (1)$$

almost surely, where $\tau_p \equiv \frac{2}{p}\mathrm{tr}\mathbf{C}$.

Besides, and most importantly, if the $\mathbf{x}_i$'s are drawn from a mixture of k distribution classes (with k fixed) such that $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O_p(1)$, with $\boldsymbol{\mu}_a = \mathbb{E}[\mathbf{x}_i]$ for $\mathbf{x}_i$ in Class a, and $\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b) = O_p(\sqrt{p})$, with $\mathbf{C}_a = \mathrm{Cov}(\mathbf{x}_i)$, then (1) remains valid: that is, the *data classes are asymptotically not distinguishable* from their distances. Here τ_p can be taken to be any $\frac{2}{p}\mathrm{tr}\mathbf{C}_a$, for $a \in \{1, \ldots, k\}$. The setting $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O_p(1)$ and $\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b) = O_p(\sqrt{p})$ is referred to as the *non-trivial classification regime*.

Remark 2 (On "non-trivial" classification)**.** *The above two assumptions $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O_p(1)$ and $\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b) = O_p(\sqrt{p})$, thoroughly discussed in (Couillet et al., 2016), are quite natural to model a* non-trivial*, that is neither too easy nor too hard, classification scenario. In other words, if either $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ or $\frac{1}{\sqrt{p}}\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b)$ were to increase with p, then a simple Bayesian analysis demonstrates that a trivial algorithm can achieve asymptotically perfect classification as p increases; conversely, if both $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ and $\frac{1}{\sqrt{p}}\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b)$ were to vanish as p increases, it is theoretically impossible to retrieve the classes with any algorithm. See (Couillet et al., 2018) for a detailed discussion. In practice, of course, p remains fixed so that the conditions $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O_p(1)$ and $\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b) = O_p(\sqrt{p})$ are mostly quantitative: in fact, "good" vector representations will tend to have rather large values of $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ and sometimes fall in a rather trivial regime (the classification task is then easy in general and most standard algorithms perform well), while other representations may be less discriminative, in which case classification is non-trivial and a well-tailored classification algorithm must be devised.*

Our first result consists in empirically confirming that the concentration of distances phenomenon of Equation (1) occurs with popular word and document representations. Specifically, Figure 2 displays the histogram of distances of a set of n vector observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ under four settings: (i) the $\mathbf{x}_i$'s are i.i.d. $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$ vs. $\mathcal{N}(-\boldsymbol{\mu}, \mathbf{T})$ for $\boldsymbol{\mu} = (4, 0, \ldots, 0)^\mathsf{T}$ (which satisfies the condition $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = O_p(1)$), $[\mathbf{T}]_{ij} = .4^{|i-j|}$ is a Toeplitz matrix, $n = 200$ and (i-a) $p = 4$ or (i-b) $p = 400$, which serves as a theoretical reference; or the $\mathbf{x}_i$'s correspond to $n = 1\,100$ balanced documents from two classes ("Christian" versus "Forsale") from the 20NewsGroup database[4]) obtained by selecting in each class the top $3\,500$ words according to their tf*idf scores, the idf being computed within the documents of the class, and encoded through (ii) Glove, (iii) word2vec, or merely through their (iv) tf*idf vectors. For comparison purposes, all datasets have been centered and normalized.

A first observation is that the distances between two-class distributions of both Glove and word2vec representations seemingly "concentrate around $\sqrt{2}$" instead of displaying a bi-modal distribution. Besides, and possibly more importantly, the distribution closely matches the distribution of distances obtained for mere large dimensional Gaussian random vectors. This "resemblance to large (rather than small) Gaussian vector behavior" provides a first hint into a behavior typical of concentrated random vectors. This conclusion does however not hold for tf*idf representations, the distance histogram of which is far from being symmetrically centered around $\sqrt{2}$, which is naturally explained by the sparse nature of the the tf*idf vectors. Together, these results are a first indicator of a peculiar concentration behavior of modern vector representations for documents, as opposed to tf*idf vectors.

As a side but important remark, we must add that the experiment and conclusions of Figure 2, here exemplified on the "Christian" versus "Forsale" classes of the 20NewsGroup database, have been replicated and consolidated on other classes and

[4] http://qwone.com/∼jason/20Newsgroups/

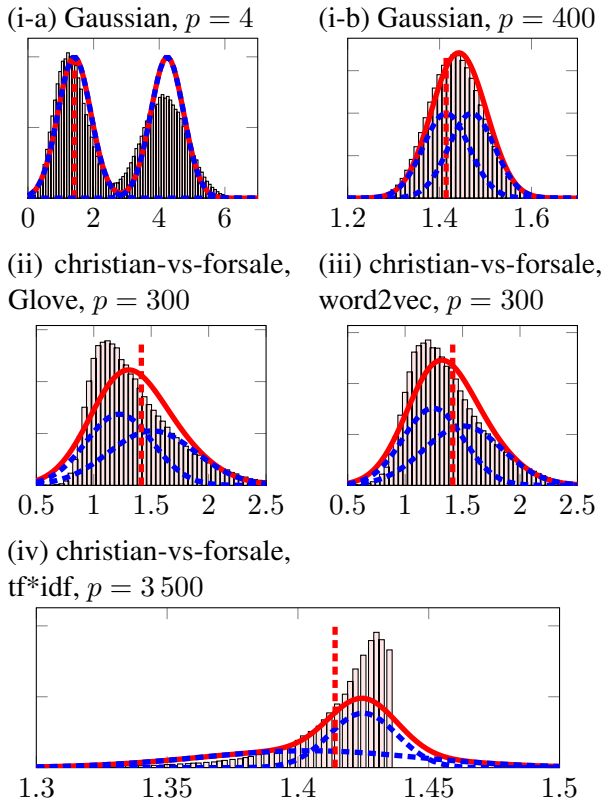

Figure 2: Distribution of (centered normalized) input data distances $\{\frac{1}{\sqrt{p}}\|\mathbf{x}_i - \mathbf{x}_j\|\}_{1 \leq i \neq j \leq n}$ for (i) two-class Gaussian mixture with mean $\pm\boldsymbol{\mu}$ of size (i-a) $p = 4$ or (i-b) $p = 400$, and two-class documents (20News-Groups, "Christian" vs. "Forsale") with (ii) Glove, (iii) word2vec, or (iv) tf*idf representations. In blue are displayed the intra- and inter-class distance distributions and in red the collective distance distribution, as if all data were Gaussian and all distances were independent (which they are not).[5] Dashed-red line pointing the $\sqrt{2}$ position (where distances theoretically concentrate).

databases. Section 3 will notably consider another database for experiments.

From a practical standpoint though, the mono-modal histograms of Figure 2 strongly suggest that "individual distance-based" document classification methods are likely to fail. The next section investigates this aspect by showing that more elaborate methods which treat data distances collectively rather than individually, such as spectral-based techniques, are more amenable to handle document vector classification than individual distance-based techniques.

[5] Exact calculus reveals that, for $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$ and $\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}_b, \mathbf{C}_b)$, under the aforementioned non-trivial regime, for $\tau_p = \frac{1}{p}\mathrm{tr}(\mathbf{C}_a + \mathbf{C}_b)$ here,

$$\frac{1}{\sqrt{p}}\|\mathbf{x}_i - \mathbf{x}_j\| \sim \mathcal{N}\left(\sqrt{\tau_p}, \sigma_{a,b}^2\right) + o_p(1)$$

$$\sigma_{a,b}^2 \equiv \frac{1}{\tau_p}\frac{1}{p^2}\mathrm{tr}(\mathbf{C}_a\mathbf{C}_b) + \frac{1}{2\tau_p}\frac{1}{p^2}\mathrm{tr}(\mathbf{C}_a^2) + \frac{1}{2\tau_p}\frac{1}{p^2}\mathrm{tr}(\mathbf{C}_b^2)$$

the quantities appearing in the variance $\sigma_{a,b}^2$ being consistently estimated from: $\frac{1}{p}\mathrm{tr}(\hat{\mathbf{C}}_1\hat{\mathbf{C}}_2) = \frac{1}{p}\mathrm{tr}(\mathbf{C}_1\mathbf{C}_2) + o_p(1)$ and $\frac{1}{p}\mathrm{tr}(\hat{\mathbf{C}}^2) = \frac{1}{p}\mathrm{tr}(\mathbf{C}^2) + \frac{1}{np}(\mathrm{tr}(\hat{\mathbf{C}}))^2 + o_p(1)$ with n the number of independent samples used to evaluate the sample covariance matrix $\hat{\mathbf{C}}$ of $\mathbf{C}$.

2.3.2 Kernel spectral behavior

A broad range of machine learning algorithms $F_\theta(\mathbf{X}, \mathbf{x})$ are of the form $G_\theta(\mathbf{K}, \mathbf{x})$ where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a *kernel matrix* of the input data $\mathbf{X}$ (kernel spectral clustering, kernel SVM, graph kernel semi-supervised learning, etc.). Typically, following our distance-based development, $\mathbf{K}_{ij} = f(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ for some smooth function f.

Studying the statistical behavior of such algorithms, even under a mere Gaussian mixture model setting, has long remained an open problem, due to the non-linearity of f and of the intricate dependence between the entries of $\mathbf{K}$. As a positive aftermath of the (a priori deleterious) concentration of distance phenomenon though, the authors in (El Karoui et al., 2010; Couillet et al., 2016) prove that, when $p, n \to \infty$, the involved matrix $\mathbf{K}$ is asymtotically well approximated by a form

$$\mathbf{K} = \mathbf{W} + \mathbf{P} + o_{\|\cdot\|}(1) \tag{2}$$

where $\mathbf{W}$ is a non-informative full-rank noise matrix and $\mathbf{P}$ is a low-rank[6] informative matrix which carries in its few eigenvectors the information about (a) the k data classes *only through* the first and second order statistics $\{\boldsymbol{\mu}_a\}_{a=1}^k$ and $\{\mathbf{C}_a\}_{a=1}^k$ of the classes, and (b) the kernel function f *only through* its local behavior around the joint distance concentration point τ_p. For instance, the popular radial-basis (RBF) kernel $f(t) = \exp(-\frac{t}{2\sigma^2})$ behaves theoretically the same as any other function (for instance a mere polynomial of order 2) having the same first two derivatives as f in τ_p. This finding, made explicit in (Couillet et al., 2016), notably opens the perspective to improve kernel-based algorithms based on a careful choice of the behavior of f around τ_p.

One of the main consequences of the approximation (2) is the theoretical ability to anticipate the *spectral behavior*, so in particular to describe

[6] Of rank usually equal or bounded by the number of classes in the dataset.

the statistics of the dominant eigenvectors[7] of $\mathbf{K}$, thereby allowing for a theoretical prediction of the performances of spectral learning (e.g., spectral clustering, manifold learning, etc.). These results are again universal in that they only depend on the statistical means and covariances of the data classes; see (Couillet et al., 2016) for details.

We wish here to demonstrate that kernel matrices built on natural language data similarly conform to the behavior of large dimensional Gaussian vectors. To this end, for the same two-class data benchmark introduced in the previous section, we design a matrix $\mathbf{K}$ for the popular RBF kernel $f(t) = \exp(-t/2)$ (that is with bandwidth $\sigma^2 = 1$) and extract its second dominant eigenvector $\mathbf{v}_2$.[8]

This is depicted in Figure 3, which it is convenient to compare to Figure 2. It is first observed that, while, according to Figure 2, the *entries of* $\mathbf{K}$, i.e., $\exp(-\cdot/2)$ applied to the distances $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2$, are not discriminating – the distance distribution being unimodal –, the *entries of* $\mathbf{v}_2$ are instead strongly informative – the eigenvector distribution is bi-modal –: this is in essence explained by a "redundancy" effect in the numerous data belonging to the same class which "gather energy" into an isolated eigenvalue with eigenvector $\mathbf{v}_2$.[9] This cumulative effect is not exploited by algorithms which treat data distances *one-by-one* (such a KNN kernel with few neighbors) rather than collectively.

A second observation, more to the point for our present demonstration, is that the histogram of the entries of $\mathbf{v}_2$ for genuine natural language data is a close match to the histogram of the synthetic Gaussian vector counterparts: this is a second manifestation of the *universality of concentration of measure*.

Remark 3 ("Behaving like" is not "being" a Gaussian)**.** *We wish to insist that this universality observation does* not *suggest that word and document vectors* look like *Gaussian vectors (this would be a mistake); it merely states that the observed* functional *of the learning data* $\mathbf{X}$ *(here the entries of an eigenvector of* $\mathbf{K}$*)* has the same asymptotic behavior *as with Gaussian vector inputs.*

[7] Those associated to the largest (or smallest) isolated eigenvalues of $\mathbf{K}$.

[8] Which is known from (Couillet et al., 2016) to be the best discriminating eigenvector in a two-class setting.

[9] In (Couillet et al., 2016), a mathematical argument using random matrix theory is provided to fully justify this observation.

These empirical results are strong indicators that natural language data representations behave similar to concentrated random vectors and may adequately be modelled as such. This implies that the *curse of dimensionality*, appearing here in the distance concentration phenomenon, is at play: as a main consequence, we expect many standard algorithms based on individual data distance evaluations to dramatically fail, where more elaborate techniques using spectral properties remain competitive and, in addition, are now prone to theoretical analysis. The next section investigates this claim in the specific case of SVMs.

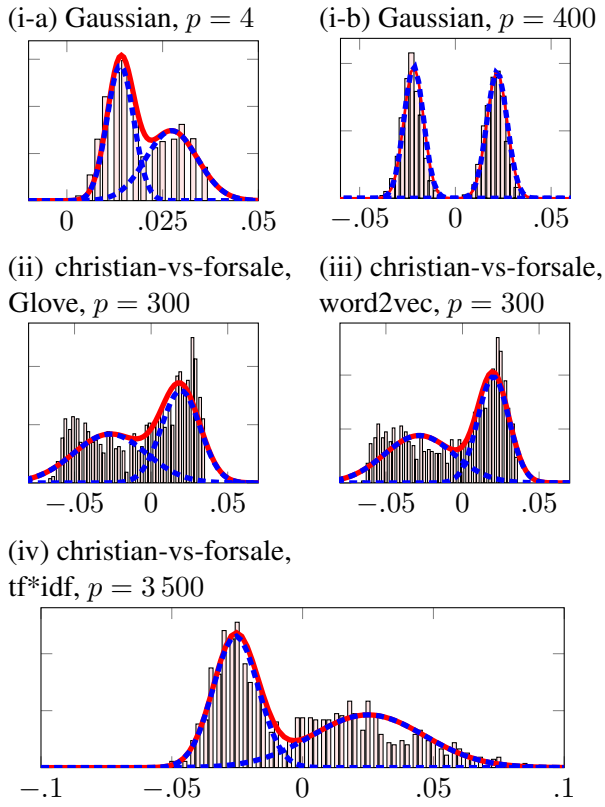

Figure 3: Histogram of the entries of the second dominant eigenvector $\mathbf{v}_2$ of $\mathbf{K} = \{\exp(-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2)\}_{1\leq i,j\leq n}$. **(Left)** Real data (same setting as in Figure 2; **(Right)** Gaussian vectors with the same (empirically estimated) first and second order statistics as their left counterpart. In red are displayed the theoretical distributions under Gaussian data input (according to (Couillet et al., 2016)).

3 Application to supervised learning

The concentration of measure phenomenon in real data (Equation (1)) has a fundamental advantage: the performance of many learning algorithms become predictable and, consequently, amenable to

improvement. The results of the previous section therefore strongly suggest that, for the first time to the authors' knowledge, one can predict to some extent (so long that the non-trivial conditions are met for the processed data) the performance of a host of machine learning algorithms for natural language processing.

Specifically, we consider here the standard least-square kernel support vector machine (LSSVM) classifier (used e.g., in (Mitra et al., 2007) for text classification with some refinement), with kernel $\mathbf{K} = \{f(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2)\}_{1\leq i,j\leq n}$, for some function f to be specified. The LSSVM classifier allocates the class of a new datum $\mathbf{x}$ based on its position with respect to a hyperplane in kernel space designed from the training set $\mathbf{X}$. Although not directly a spectral method (as in the unsupervised spectral clustering algorithm (Von Luxburg, 2007)), for large n, p, the LSSVM classifier inherently exploits the eigenspectrum of the kernel matrix $\mathbf{K}$ and its performance is proved in (Liao and Couillet, 2019) to be asymptotically predictable (for large enough p, n) and in closed form (which is thus simpler than the margin-based SVM, whose asymptotic performances do not admit a closed form).

Precisely, the class $\mathcal{C}_1$ or $\mathcal{C}_2$ allocated to $\mathbf{x}$ is the result of the binary test

$$g(\mathbf{x}) \underset{\mathcal{C}_1}{\overset{\mathcal{C}_2}{\gtrless}} \zeta_p$$

for some well-chosen threshold $\zeta_p \in \mathbb{R}$, where $g(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\mathbf{k}(\mathbf{x}) + b$, with

$$\boldsymbol{\alpha} = \mathbf{S}^{-1}(\mathbf{y} - b\mathbf{1}_n), \quad b = \frac{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{y}}{\mathbf{1}_n^\mathsf{T}\mathbf{S}^{-1}\mathbf{1}_n}$$

and $\mathbf{S} = \mathbf{K} + \frac{n}{\gamma}\mathbf{I}_n$, for $\mathbf{y} \in \{\pm 1\}^n$ the vector of training data labels, $\mathbf{k}(\mathbf{x}) = \{f(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}\|^2)\}_{1\leq i\leq n}$, and regularization $\gamma > 0$.

In (Liao and Couillet, 2019), the authors precisely show that, for a two-class mixture of concentrated random vectors with means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and covariances $\mathbf{C}_1$, $\mathbf{C}_2$, as $n, p \to \infty$ in the non-trivial regime described above, for $\mathbf{x}$ genuinely in class $\mathcal{C}_i$,

$$g(\mathbf{x}) \to \mathcal{N}(m_i, \sigma_i^2)$$

where m_1, m_2, σ_1^2 and σ_2^2 only depend on (a) the ratio $f'(\tau_p)/f''(\tau_p)$ and (b) scalar functionals of the statistical means and covariances (specifically, only $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, $\mathrm{tr}(\mathbf{C}_1 - \mathbf{C}_2)/\sqrt{p}$ and $\mathrm{tr}((\mathbf{C}_1 - \mathbf{C}_2)^2)/p$); see (Liao and Couillet, 2019) for details.[10] For instance, $f(t) = \exp(-t/2\sigma^2)$ is the standard radial-basis function kernel (RBF) with bandwidth σ^2, the asymptotic performances of which only depend on $f'(\tau_p)/f''(\tau_p) = -2\sigma^2$.

Of utmost relevance here is that the asymptotic performances are *identical for concentrated random vectors as for Gaussian random vectors* having the same first and second order statistics.

Figure 4 reports the performances of LSSVM as a function of the ratio $f'(\hat{\tau}_p)/f''(\hat{\tau}_p)$, where $\hat{\tau}_p \equiv \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is a consistent (and fast converging) estimate for τ_p, here for two kernels: (a) the second order polynomial kernel such that $f(\hat{\tau}_p) = 4$, $f''(\hat{\tau}_p) = 1$ and $f'(\tau_p)$ varying from -2 to 1, and (b) the RBF kernel with bandwidth σ^2 such that $-2\sigma^2$ varies from -2 to 0 (of course $-2\sigma^2$ cannot be positive).

The benchmark dataset are the Yahoo Answer classes "cult" versus "education", the feature vectors of which are either (ii) Glove embedding ($p = 300$), (iii) word2vec embedding ($p = 300$) and (iv) tf*idf representation (with dictionary size $p = 3\,000$). A comparison to (i) Gaussian input data vectors is also provided for reference ($p = 300$). In each experiment, the number of training data is $n = 500$ or $n = 2\,000$.

Figure 4 first shows a trend for the performances to converge, as the results for both $n = 500$ and $n = 2\,000$ are similar: as such, these performances are not random and then possibly amenable to theoretical analysis.

More in detail, Figure 4 demonstrates that, for the tf*idf representation, the theoretical equivalent for concentrated vectors (red) and the empirical performance (blue) are quite different, clearly confirming that tf*idf representations are not appropriately modelled by concentrated vectors. This is again no surprise as these vectors are intrinsically sparse, which concentrated vectors cannot be.

The case of word2vec and Glove is more interesting as Figure 4 reports an extremely accurate fit between theory and practice for $f'(\hat{\tau}_p)/f''(\hat{\tau}_p)$ below -1 and above .5. More crucially, in these regions, the performances for both the RFB kernel and the polynomial kernel with $f'(\hat{\tau}_p)/f''(\hat{\tau}_p) = -2\sigma^2$ perfectly coincide, so that real data performance

[10] Of particular interest, (Liao and Couillet, 2019) proves that the optimal threshold ζ_p *must* be around $\frac{n_2}{n} - \frac{n_1}{n}$, with n_a the number of elements of class $\mathcal{C}_a$ in the training dataset, and not around 0 as conventionally assumed.

corroborates the theory. Only the region $[-1, .5]$ shows a severe discrepancy. This is explained by two factors: (a) for any kernel (here for the polynomial kernel), (Liao and Couillet, 2019) shows that the region where $f'(\tau_p) \simeq 0$ is particularly unstable to "strongly mean-discriminative data", i.e., data mixtures strongly identifiable from their statistical means; this is what is observed here with a vanishing performance (dropping to 50%) when $f'(\hat{\tau}_p) = 0$, inducing instability; at this point of our analysis though, we cannot explain the performance increase near 0^- predicted by the theory while the empirical performance monotonously drops; (b) for the RBF kernel, in the vicinity of $\sigma^2 \sim 0$, the entries of $\mathbf{K}$ degenerate; $\mathbf{K}$ becomes sparse, which goes against concentration; this is already observed for Gaussian inputs (top display); this gap can only be covered with larger p, n values.

4 Concluding Remarks

The results of this article may scratch the surface of a new mathematical theory for harnessing modern natural language processing representations: modern word and document features (word2vec and Glove) were shown here to exhibit some key characteristics of concentrated random vectors, which tf*idf maps do not. This, as a consequence of recent works on the analysis of machine learning algorithms for concentration random vectors, opens the path to theoretical analyses, improved understanding, fine-tuned and new algorithms for natural language data processing.

Yet, the preliminary conclusions of the present article are less compelling than similar conclusions drawn for image representations (e.g., in (Liao and Couillet, 2019; Seddik et al., 2020), where the performance predictions on real images are extremely accurate for wide ranges of hyperparameters). This may be interpreted in two ways: either the modern document representation (Glove and word2vec) need be perfected to be as discriminative and "maximum entropic"[11] as VGG or ResNet are for images, or the *concentration power* of document embeddings is intrinsically weaker than image embeddings. If the latter hypothesis is correct, further mathematical efforts are needed to improve our understanding of these "weakly concentrated" data models. But both hypotheses at any rate strongly suggest that natural language processing is more amenable to theoretical probabilistic analysis than considered to this day, which brings hope for accompanying the recent word embedding revolution in NLP by a renewal of the entire NLP theory.

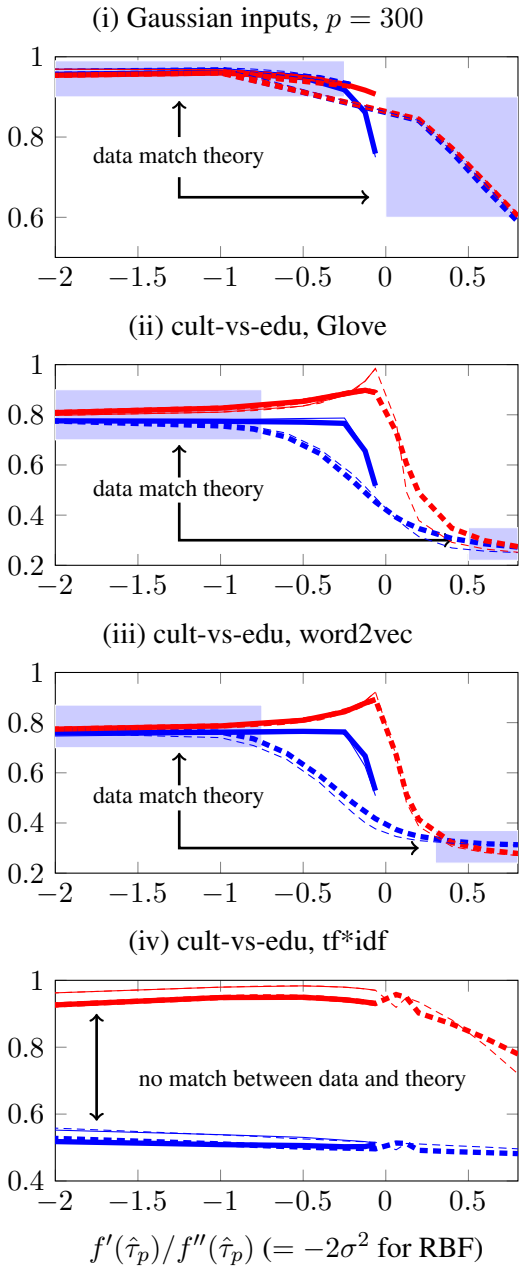

Figure 4: Comparative performance of LSSVM classification for various document vector representations: empirical results (**blue**) versus asymptotic theory (**red**), for $n = 500$ (**thick**) or $n = 2000$ (**light**), with either RBF (**plain**) or second-order polynomial (**dashed**) kernel. Good fit for word2vec and Glove embeddings away from unstability region of $f'(\hat{\tau}_p)/f''(\hat{\tau}_p)$, suggesting concentration (Gaussian-like) behavior; tf*idf data do not concentrate.

[11] We suggest here that good representations should extract all the information and leave residual noise as maximally uninformative; in the manner of isotropic Gaussian vectors.

References

Carl Allen and Timothy M. Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 223–231. PMLR.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Trans. Assoc. Comput. Linguistics*, 4:385–399.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5564–5577.

Romain Couillet, Florent Benaych-Georges, et al. 2016. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454.

Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. 2018. Classification asymptotics in the random matrix regime. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1875–1879. IEEE.

Roberto Dessì and Marco Baroni. 2019. Cnns found to jump around more skillfully than rnns: Compositional generalization in seq2seq convolutional networks. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3919–3923.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Noureddine El Karoui et al. 2010. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards understanding linear word analogies. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3253–3262.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding undesirable word embedding associations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1696–1705.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. 2017. Skip-gram - zipf + uniform = vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 69–76.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Yair Lakretz, Germán Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 11–20.

Michel Ledoux. 2001. *The concentration of measure phenomenon*. 89. American Mathematical Soc.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185.

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

Henry W Lin, Max Tegmark, and David Rolnick. 2017. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Vikramjit Mitra, Chia-Jiu Wang, and Satarupa Banerjee. 2007. Text classification: A least square support vector machine approach. *Applied Soft Computing*, 7(3):908–914.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

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

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 37:141–188.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.