Compressed Sensing Project

A compressed sensing view of unsupervised text embeddings

Salomé Do, Lucas Zanini

March, 22nd 2019

Introduction

Text documents have been, until recently, an underexploited source of information due to their complex, non-numerical structure. Low level natural language processing (NLP) tasks as text classification, sequence tagging (POS-tagging, named entity recognition, ...) and parsing have long been tackled through bag-of-words methods and hidden markov models. The application of deep learning to NLP tasks (started around 2010s) has opened new, complementary and effective ways to solve these problems, eventually leading to core innovations in high-level tasks such as question answering or language generation. Using feed-forward networks to generate low-dimensional representation of sparse, high dimensional word vectors; recurrent neural networks as LSTMs to fit the sequential nature of sentences; and later on attention mechanims, deep learning techniques have set new baselines on traditional benchmarks. However, the efficiency of these results is totally empirical and is not backed by theoretical results, in addition to having high computationnal costs. The article studied in this project [1] aims at giving some results on unsupervised text embeddings.

To do so, it focusses on a compressed sensing view of the text embedding problem, which aims at provinding low-dimensional representation for texts. As words can be viewed as very sparse vectors in the vocabulary space, providing a low-dimensional representation of a text can be seen as measuring words signals in a compressed way. The article explores this idea, and links compressed sensing to some LSTMs-learned text representations. In order to evaluate such representations (through their performances on text classification), authors adapts a result in [2] on the quality of low-dimensional compressed sensing representations as inputs for a classification problem. Showing that some embeddings are compressing matrix verifying the proper RIP conditions, the above compressed learning results apply.

In this work, we will first recall what are words and text embeddings, in order to recall the basic NLP background for a non-specialized reader, and we will present author's text embeddings, how they relate to compressed sensing, and the properties they verify (useful RIP conditions for the rest of the work). Then, we will focuss the compressed learning problem and its relation to LSTMs.

Contents

1 Text embedding as a Compressed Sensing problem

A text embedding is a vector representing a text. Usually, we want this representation to be low-dimensional, and to keep some information on the meaning of the text. By necessity, these representations are generally unsupervised : it would be hard to define a "standard" representation to be learned in a supervised way. In this section, we recall usual ways to learn to generate word embeddings - upon which lie LSTMs text embeddings -, text embeddings, and we explain in which ways they can be used for tasks as text classification.

1.1 Word embeddings

The very beginning of most popular text embeddings is word embeddings. In this part we will briefly explain how such embeddings are learned, and we will set the definitions and the general context for the rest of the work.

Supposing that we have a vocabulary $\mathcal{V} = \{w_1, ..., w_V\}$, of size $|\mathcal{V}| = V$, a natural vector representation of any word $w_i, i = 1, ..., V$ is the following :

$$w_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i, \quad w_i \in \mathbb{R}^V$$

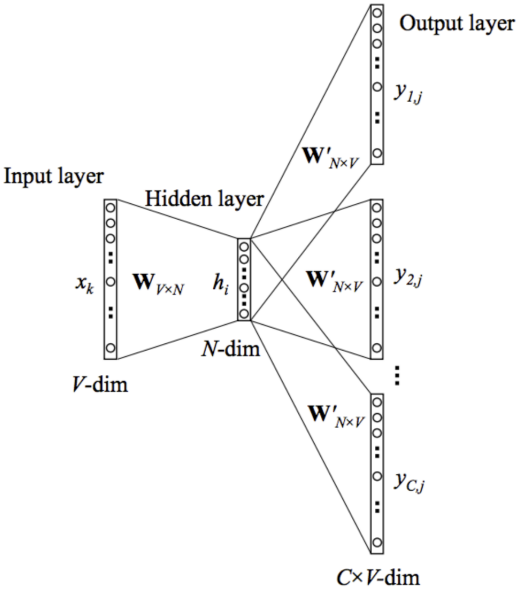
The problem with this simple representation is the dimension (V) and its sparsity ($||w_i||_0 = 1, \forall i = 1, ..., V$) of the vectors. To adress this issue, many techniques have been proposed, especially in recent deep learning developments. Non-deep learning techniques are based on co-occurence matrices : given a set $\mathcal{D}$ of documents, we first build a matrix containing the number of co-occurrences of every word in all documents in a context window of size c, for instance :

$$M = \begin{pmatrix} C(w_1, w_1) & \dots & C(w_1, w_V) \\ \vdots & \ddots & \vdots \\ C(w_V, w_1) & \dots & C(w_V, w_V) \end{pmatrix} \in \mathbb{R}^{V \times V}$$

Where $C(w_i, w_j)$ is the number of times that the word w_j appears in the context of w_i, the context being defined as the words far from at most c words from w_i in any of the documents in $\mathcal{D}$. This matrix is usually regularized using positive pointwise mutual information (PPMI). A word w_i is then represented as the i-th line of M, or its PPMI counterpart. However, this representation does not get rid of dimension neither sparsity problems. Assuming the PPMI matrix is very sparse, we use, as it is frequently done in high-dimensional statistics, a singular value decomposition $M^{\text{PPMI}} = W \times D \times C$, keeping only the k most important singular values and vectors, leading to a new truncated version of W, of size $V \times k$, which becomes the embedding matrix, each line being a word's embedding. This decomposition has good denoising and generalization properties.

In parallel, the most popular deep learning embedding is the Word2Vec embedding, proposed in [3]. This embedding is learned by trying to predict a word from its context (CBOW), or vice versa to predict a context from a given word (skip-gram); using a 1-hidden layer feed-forward neural network (Figure 1).

Figure 1: Skip-gram architecture for Word2Vec



Using some regularization and computationnal tricks during the training, as replacing the softmax output function computation by a hierarchical softmax computation, or as using negative sampling loss as an objective function, or doing frequency subsampling; which are detailed in [4]; Word2Vec embeddings achieved better generalization and denoising properties than SVD. The main idea behind both Word2Vec and SVD embeddings, which makes them coherent as a good semantic representation, is called the distributionnal hypothesis. This linguistic hypothesis is built upon Firth's statement that *"a word is characterized by the company that it keeps"*. Thus, words employed in similar contexts are more likely to have similar meanings. However, latest deep learning developments seem to have abandoned this approach, replaced by language modelling as in [5, 6, 7], which are outside the scope of this project.

The evaluation of such embeddings is difficult, as they are learned in an unsupervised way. However, two techniques are generally used to assess embeddings representation quality : the evaluation of the embedding through a downstream task, i.e. for instance the evaluation of classic text classifiers using the evaluated embedding as an input; and an evaluation through word similarities, with the objective that near vectors (regarding cosine distance) represent words with similar meanings. This evaluation techniques are used in the same manner for text embeddings.

1.2 From word embeddings to text embeddings

In natural language processing, and more specifically in high-level tasks, there is no interest on studying single words. Thus, a natural question when working on texts, is : how to transform word embeddings to a single text embedding? In this section we will describe 'natural' text embeddings, resembling the co-occurence matrix described above, used in the article; and LSTMs embeddings, which are now more commonly used.

1.2.1 Bag of n-Grams/Cooccurrences

In this part, we will suppose that we have a vocabulary $\mathcal{V} = \{v_1, ..., v_V\}$ and a document $d = w_1, ..., w_T$, where $w_i \in \mathcal{V}$ for any $i = 1, ..., T$, and a fixed integer $n \in \mathbb{N}$. We recall that a n-gram is a sequence of n words, for instance the set of unigrams is $\{v_1, ..., v_T\}$, the set of bigrams is $\{(v_1, v_2), (v_2, v_3), (v_1, v_3)..., (v_{V-1}, v_V)\}$, etc. We let V_k be the number of possible k-grams over $\mathcal{V}$, independently of order ($V_1 = V$), and $V_n^{\text{sum}} = \sum_{k=1}^{n} V_k$.

For any $k \in \mathbb{N}$, let B_k be a vector indicating, for any possible k-gram over $\mathcal{V}$, the number of occurences of this k-gram in the document. A Bag-of-n-Grams, denoted x^{BonG} is the concatenation of all the B_k vectors for $k = 1, ..., n$:

$$x^{\text{BonG}} = [B_1, ..., B_N]$$

In order to simplify and get rid of order in the n-grams, author merge the k-grams in the vocabulary that are the same words in a different order. Then, we can write : $B_k^{\text{Co-oc}} = \sum_{t=1}^{T-k+1} e_{\{w_t, ..., w_{t+k-1}\}}$. A Bag-of-$n$-Co-occurrences (BonC) is then :

$$x^{\text{BonC}} = [B_1^{\text{Co-oc}}, ..., B_N^{\text{Co-oc}}]$$

These embeddings are sparse, V_n^{sum}-dimensional vectors.

1.2.2 DisC embeddings

This document embeddings relies on word embeddings. Supposing that we have learned word embeddings and that we denote $x_{v_i} \in \mathbb{R}^d, d << V$ the embedding of the word $v_i \in \mathcal{V}$, for any $i = 1, ..., n$; and that we still have our document $d = w_1, ..., w_T$; then the unigram embedding of the document is :

$$z^u = \sum_{t=1}^{T} x_{w_t}$$

z^u can be re-written as the product of a compression matrix A in which columns are word embeddings x_w; and the "Bag-of-1-grams" document embedding :

$$z^u = Ax^{\text{Bo1G}} = \sum_{t=1}^{T} Ae_{w_t} = \sum_{t=1}^{T} x_{w_t}$$

Authors extend this unigram definition to n-co-occurrences by using element-wise multiplication of word embeddings, defining:

$$\tilde{x}_{\{w_1, ..., w_n\}} = d^{(n-1)/2} \odot_{t=1}^{n} x_{w_t} \quad \in \mathbb{R}^d$$

Then, DisC (distributed co-occurence) embeddings are defined as the concatenation of:

$$z^{(n)} = \left[C_1 \sum_{t=1}^{T} \tilde{x}_{w_t}, ..., C_n \sum_{t=1}^{T-n+1} \tilde{x}_{\{w_t, ..., w_t+n-1\}} \right] \in \mathbb{R}^{nd}$$

With $C_1, ..., C_n$ being scaling factors, detailed later. As in the unigram case, DisC embeddings are directly related to compressed sensing as we can find $A^{(n)} \in R^{dn \times V_n^{\text{sum}}}$ such that:

$$z^{(n)} = A^{(n)} x^{BonC}$$

1.2.3 LSTMs

LSTMs, which stands for Long-Short-Term Memory are a kind of recurrent neural networks (RNN), introduced in [8]. As with classic RNNs, the point in LSTMs is to be able to handle sequential data, say for instance a document made of already embedded words :

$$x = \{x_1, ..., x_T\}$$

A given LSTM cell A 'evolves' with the time, and is caracterized in time by a *hidden state* h_t and a *memory* c_t. We denote $A^{(t)}$ the state of cell A at time t. $A^{(t)}$ has three inputs :

- h_{t-1}, the hidden state transmitted by the previous cell state $A^{(t-1)}$.
- c_{t-1}, the memory transmitted by the previous cell state $A^{(t-1)}$.
- x_t, the 'real' sample input

And two outputs :

- h_t, the updated hidden state transmitted to $A^{(t+1)}$.
- c_t, the updated memory transmitted to $A^{(t+1)}$.

These inputs and outputs are schematized in Figure 2 [1]. The core question of LSTMs is to define update rules for h_t and c_t. First, we want to know how to update memory, or in other words, which informations to forget and which informations to remember. To do so, we define :

- $i_t = \sigma\left(x_t U^i + h_{t-1} X^i\right) \in [0, 1]$, where U^i, W^i are weights to learn. This function is called the *input gate*. The input gate is a weighted sum of the sample data at time t and the previous hidden state of the cell, regularized by the sigmoid function. The inpute gate decides, what new information we are going to store in the updated memory, and in which proportion, depending on the precedent data.
- $\tilde{c}_t = \tanh\left(x_t U^c + h_{t-1} X^c\right) \in [-1, 1]$, where U^c, W^c are weights to learn. $\tilde{c}_t$ is a candidate for creating 'new c_t values'.
- $f_t = \sigma\left(x_t U^f + h_{t-1} X^f\right) \in [0, 1]$, where U^f, W^f are weights to learn. This is called the *forget gate*. The forget gate decides the proportion at which each element of the memory is going to be forgotten.

The memory can then be updated, with respect to the following rule:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

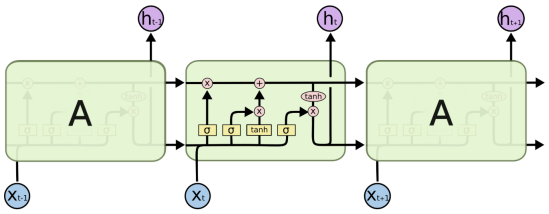
Finally, to update the hidden state, we define the *output gate* in the same fashion, and we update h_t :

$$\begin{aligned} o_t &= \sigma\left(x_t U^o + h_{t-1} X^o\right) && \in [0, 1] \\ h_t &= \tanh(c_t) * o_t && \in [-1, 1] \end{aligned}$$

Some variants of LSTM cells may define input, output and forget gates differently. For instance, the peephole version of LSTM implements the natural idea that the previous memory state c_{t-1} has to be considered when deciding the input, output and forget rates. GRU, Gated Recurrent Units are other variants of LSTMs.

[1] The figure is extracted from Christopher Olah's excellent blog : http://colah.github.io/

Figure 2: A representation of a LSTM cell in time



One question remains : what exactly is the hidden state h? $(h_t)_{t=1...T}$ can be viewed as the 'output' of the LSTM. In our case, the final hidden state h_T is directly used as the document embedding. This representation of the document is unsupervised, and the LSTM is usually trained on another task, regarding which $(h_t)_{t=1...T}$ are features (see for instance [9] for an example of LSTMs trained for named entity recognition).

Appart from having a very flexible and modular structure, LSTMs cells do not suffer of the same vanishing gradient problem as RNNs in practice, and are thus more effective in modelling sequential data. They are particularly adapted to NLP problems because of their sequential nature.

2 Learning to classify in the compressed domain

Texts can be viewed as a collection of sparses measures (words) over a complex signal (ideas, a meaning, a sentiment, etc.). The text embedding problem is thus directly a compressed sensing problem : using a compression (=embedding) matrix A over a collection of words $x = w_1, ...w_T$ to produce a dense, compressed representation $z = Ax$, with good reconstruction qualities. As the "reconstruction" quality of the compressed signal cannot, in this case, be properly evaluated, we need results on this reconstruction quality through a downstream task as classification.

2.1 Results on compressed sensing for classification

As we are interested in the evaluation of text embeddings through text classification, the compressed learning approach proposed in [2] (generalized and modified in our article of interest [1]) gives us a useful result stated in 2.1. This result tells us how training a linear classifier over the compressed data instead of the original data penalizes the classification results.

We chose to reproduce the ideas of proof of this theorem here, with more detail in general, as we think that it is an very interesting result regarding the compressed sensing course.

Theorem 2.1. *For any subset $\mathcal{X} \subset \mathbb{R}^N$ containing the origin, let $A \in \mathbb{R}^{d \times N}$ be $(\Delta\mathcal{X}, \varepsilon)-RIP$, and m samples $S = \{(x_i, y_i)\} \subset \mathcal{X} \times \{-1, 1\}$ be drawn i.i.d. from the same distribution $\mathcal{D}$ over $\mathcal{X}$, with $||x||_2 \leq R$.*
If l is a λ-Lipschitz convex loss function and $w_0 \in \mathbb{R}^N$ is its minimizer over $\mathcal{D}$, then, with probability $1 - 2\delta$, the linear classifier $\hat{w}_A \in \mathbb{R}^d$ minimizing the l_2-regularized empirical loss function $l_{S_A}(w) + \frac{1}{2C}||w||_2^2$ over the compressed sample $S_A = \{(Ax_i, y_i)\}_{i=1}^m \subset \mathbb{R}^d \times \{-1, 1\}$ satisfies :

$$l_{\mathcal{D}}(\hat{w}_A) \leq l_{\mathcal{D}}(w_0) + \mathcal{O}\left(\lambda R||w_0||_2 \sqrt{\varepsilon + \frac{1}{m}\log\frac{1}{\delta}}\right) \tag{1}$$

for an appropriate choice of C, and recalling that $\Delta\mathcal{X} = \{x - x' : x, x' \in \mathcal{X}\}$, for any $\mathcal{X} \subset \mathbb{R}^N$

As the proof is a bit complex, we are going to explain its outline first. The idea is to first define an empirical (regularized) risk minimizer on both the sparse and the compressed domain, respectively $\hat{w}, \hat{w_A}$.
In parallel, we define the theoretical (regularized) risk minimizer (which minimizes the loss expectation over the whole distribution $\mathcal{D}$), also both in the sparse and in the compressed domain, respectively w^*, w_A^*.
On the compressed domain, we compare the loss of $\hat{w_A}$ with the loss of w_A^*. Using a lemma deduced from the RIP condition over A and a result from [10], we compare the losses of w_A^* and $A\hat{w}$. The same lemma enables us to compare the losses of $A\hat{w}$ and $\hat{w}$. Finally, we compare the losses of $\hat{w}$ and w^*.

First define the l_2-regularization of the loss function l as :

$$L(w) = l(w) + \frac{1}{2C}||w||_2^2$$

Lemma 2.2. *Let $\hat{w}$ be the classifier obtained minimizing : $L_S(w) = \frac{1}{m}\sum_{i=1}^m l(w^T x_i, y_i) + \frac{1}{2C}||w||_2^2$, i.e. $\hat{w}$ is the empirical (regularized) risk minimizer, where $l(.,.)$ is a convex λ-Lipschitz function in the first coordinate. Then :*

$$\hat{w} = \sum_{i=1}^m \alpha_i y_i x_i, \quad |\alpha_i| \leq \frac{\lambda C}{m}, \forall i$$

The result holds in the compressed domain.

Proof. By convexity, the only optimizer can be found by taking first-order conditions :

$$\partial_w L_S(w) = \frac{w}{C} + \frac{1}{m}\sum_{i=1}^m \partial_{w^T x_i} l(w^T x_i, y_i) x_i$$

Then, we have :

$$\begin{aligned}\hat{w} &= -\frac{C}{m}\sum_{i=1}^m \partial_{\hat{w}^T x_i} l(\hat{w}^T x_i, y_i) x_i \\ &= \frac{C}{m}\sum_{i=1}^m -y_i \partial_{\hat{w}^T x_i} l(\hat{w}^T x_i, y_i) y_i x_i\end{aligned}$$

As $y_i \in \{-1, 1\}$. l is a λ-Lipschitz function, thus its sub-gradient is bounded by λ, implying $|\partial_{\hat{w}^T x_i} l(\hat{w}^T x_i, y_i)| \leq \lambda$. Hence the result. □

Lemma 2.3.

$$x, x' \in \mathcal{X} \Rightarrow (1+\varepsilon)x^T x' - 2R^2\varepsilon \leq (Ax)^T(Ax') \leq (1-\varepsilon)x^T x' + 2R^2\varepsilon$$

Proof. In one of our hypothesis, A is $(\Delta\mathcal{X}, \varepsilon)$-RIP. Then, by definition :

$$(1-\varepsilon)||x - x'||_2 \leq ||A(x - x')||_2 \leq (1+\varepsilon)||x - x'||_2$$

Using this, we show the first side of the inequation, as the other side is symmetric :

$$\begin{aligned} ||A(x-x')||_2^2 &= ||Ax||_2^2 + ||Ax'||_2^2 - 2(Ax)^T(Ax') \\ (1-\varepsilon)||x-x'||_2^2 &\leq ||Ax||_2^2 + ||Ax'||_2^2 - 2(Ax)^T(Ax') \\ (1-\varepsilon)(||x||_2^2 + ||x'||_2^2 - 2x^Tx') &\leq ||Ax||_2^2 + ||Ax'||_2^2 - 2(Ax)^T(Ax') \end{aligned}$$

As $0_N \in \mathcal{X}$, A is also $(\mathcal{X}, \varepsilon)$-RIP, leading to having : $||Ax||_2^2 \leq (1+\varepsilon)||x||_2^2$, and thus :

$$(1-\varepsilon)(||x||_2^2 + ||x'||_2^2 - 2x^Tx') \leq (1+\varepsilon)(||x||_2^2 + ||x'||_2^2) - 2(Ax)^T(Ax')$$

Finally, using the hypothesis : $||x||_2 \leq R$,

$$(Ax)^T(Ax') \leq (1-\varepsilon)x^Tx' + 2R^2\varepsilon$$

□

Corollary 2.3.1.

$$||\hat{w}||_2^2 \leq \lambda^2 C^2 R^2 \text{ and } ||\hat{w}_A||_2^2 \leq \lambda^2 C^2 (1+\varepsilon)^2 R^2$$

Proof. Using 2.2 when calculating $||\hat{w}||_2^2$, and using $||x||_2 \leq R$, we find the first bound. Using 2.3, when calculating $||\hat{w}_A||_2^2$, and $||x||_2 \leq R$, we find the second bound. □

Lemma 2.4.

$$L_{\mathcal{D}}(A\hat{w}) \leq \ L_{\mathcal{D}}(\hat{w}) + \mathcal{O}(\lambda^2 C R^2 \varepsilon)$$

Proof. As this proof is very focused on calculations, we will only insist on most important points, and not whole computations.

First, we use 2.2 to have that :

$$(A\hat{w})^T(Ax) = \sum_{i=1}^{m} \alpha_i y_i (Ax_i)^T(Ax)$$

Using 2.3,

$$\hat{w}^Tx - 3\lambda C R^2 \varepsilon \leq (A\hat{w})^T(Ax) \leq \hat{w}^Tx + 3\lambda C R^2 \varepsilon$$

As l is λ-Lipschitz, we can take the expectations over $\mathcal{D}$:

$$l_{\mathcal{D}}(A\hat{w}) \leq l_{\mathcal{D}}(\hat{w}) + 3\lambda C R^2 \varepsilon \tag{2}$$

We now want to compute $||A\hat{w}||_2^2$:

$$\begin{aligned} ||A\hat{w}||_2^2 &= (A\hat{w})^T(A\hat{w}) \\ &= (\sum_{i=1}^{T} \alpha_i y_i x_i)^T (\sum_{j=1}^{T} \alpha_j y_j x_j) \\ &= \sum_{i=1}^{T}\sum_{j=1}^{T} \alpha_i \alpha_j y_i y_j (Ax_i)^T(Ax_j) \end{aligned}$$

Using 2.3,

$$||A\hat{w}||_2^2 \leq ||\hat{w}||_2^2 + 3\lambda^2 C^2 R^2 \varepsilon$$
$$\Leftrightarrow \frac{1}{2C}||A\hat{w}||_2^2 \leq \frac{1}{2C}||\hat{w}||_2^2 + \frac{3}{2}\lambda^2 C R^2 \varepsilon$$

With this inequality and using 2, we have the result. □

We have shown some results on $\hat{w}$, the empirical regularized risk (L_S) minizer. We are now interested in w^*, the theoretical regularized risk ($L_\mathcal{D}$) minimizer.

Lemma 2.5. *Letting w^* be the linear classifier minimizing $L_\mathcal{D}$, we have that, with probability $1-\gamma$:*

$$L_\mathcal{D}(\hat{w}) \leq L_\mathcal{D}(w^*) + \mathcal{O}\left(\frac{\lambda^2 C R^2}{m}\log\frac{1}{\gamma}\right)$$

This result holds in the compressed domain.

Proof. Corollary 2.1 tells us that :

$$||\hat{w}||_2^2 \leq \lambda^2 C^2 R^2 \text{ and } ||\hat{w}_A||_2^2 \leq \lambda^2 C^2 (1+\varepsilon)^2 R^2$$

In other words, $\hat{w}$ and $\hat{w_A}$ are contained in a closed convex subset independent of S. Authors use the fact that l is λ-Lipschitz to say that L is strongly $\frac{1}{C}$-strongly convex. Moreover, $||x||_2 \leq R$. A theorem in [10] can be applied, giving that with probability $1-\gamma$:

$$L_\mathcal{D}(\hat{w}) - L_\mathcal{D}(w^*) \leq 2[L_\mathcal{S}(\hat{w}) - L_\mathcal{S}(w^*)]_+ + \mathcal{O}\left(\frac{\lambda^2 C R^2}{m}\log\frac{1}{\gamma}\right)$$

By definition, $\hat{w}$ minimizes L_S, thus : $[L_\mathcal{S}(\hat{w}) - L_\mathcal{S}(w^*)]_+ = 0$, and the result follows directly. □

Proof of the theorem:

Proof. As $l_\mathcal{D}(\hat{w}_A) \leq l_\mathcal{D}(\hat{w}_A) + \frac{1}{2C}||w_A^*||_2^2 = L_{\mathcal{D}(\hat{w_A})}$. Lemma 2.5, applied to the compressed domain gives :

$$l_\mathcal{D}(\hat{w_A}) \leq L_\mathcal{D}(\hat{w_A}) \leq L_\mathcal{D}(w_A^*) + \mathcal{O}\left(\frac{\lambda^2 C R^2}{m}\log\frac{1}{\gamma}\right)$$

As w_A^* is the minimizer of $L_\mathcal{D}, L_\mathcal{D}(w*_A) \leq L_\mathcal{D}(A\hat{w})$. Thus

$$\begin{aligned} l_\mathcal{D}(\hat{w_A}) &\leq L_\mathcal{D}(A\hat{w}) + \mathcal{O}\left(\frac{\lambda^2 C R^2}{m}\log\frac{1}{\gamma}\right) \\ &\leq L_\mathcal{D}(\hat{w}) + \mathcal{O}(\lambda^2 C R^2 \varepsilon) + \mathcal{O}\left(\frac{\lambda^2 C R^2}{m}\log\frac{1}{\gamma}\right) \text{ (lemma 2.5)} \\ &\leq L_\mathcal{D}(\hat{w}) + \mathcal{O}\left(\lambda^2 C R^2\left(\varepsilon + \frac{1}{m}\log(\frac{1}{\gamma})\right)\right) \end{aligned}$$

In the sparse domain, lemma 2.5 assure that : $L_\mathcal{D}(\hat{w}) \leq L_\mathcal{D}(w^*) + \mathcal{O}\left(\frac{\lambda^2 C R^2}{m}\log\frac{1}{\gamma}\right)$. We put this inequality in the last inequality to have:

$$l_\mathcal{D}(\hat{w_A}) \leq L_\mathcal{D}(w^*) + \mathcal{O}\left(\lambda^2 C R^2\left(\varepsilon + \frac{1}{m}\log(\frac{1}{\gamma})\right)\right)$$

Finally, as w^* minimizes $L_{\mathcal{D}}$, $L_{\mathcal{D}}(w^*) \leq L_{\mathcal{D}}(w_0) = l_{\mathcal{D}}(w_0) + \frac{1}{2C}||w_0||_2^2$ Which leads to:

$$l_{\mathcal{D}}(\hat{w_A}) \leq l_{\mathcal{D}}(w_0) + \frac{1}{2C}||w_0||_2^2 + \mathcal{O}\left(\lambda^2 C R^2 \left(\varepsilon + \frac{1}{m}\log(\frac{1}{\gamma})\right)\right)$$

The theorem follows easily by optimizing C. □

2.2 Application on LSTMs

References

[1] S. Arora, M. Khodak, N. Saunshi, and K. Vodrahalli, "A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and lstms," in Proceedings of the 6th International Conference on Learning Representations (ICLR), 2018.

[2] R. Calderbank, "Compressed learning : Universal sparse dimensionality reduction and learning in the measurement domain," 2009.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Advances in Neural Information Processing Systems 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119, Curran Associates, Inc., 2013.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," pp. 1–12, 2013.

[5] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018.

[6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, pp. 1735–1780, Nov. 1997.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[10] K. Sridharan, S. Shalev-shwartz, and N. Srebro, "Fast rates for regularized objectives," in Advances in Neural Information Processing Systems 21 (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 1545–1552, Curran Associates, Inc., 2009.