Progress Report
14-29 February 2012

Amir Hesam Salavati
E-mail: hesam.salavati@epfl.ch

Supervisor: Prof. Amin Shokrollahi
E-mail: amin.shokrollahi@epfl.ch

Algorithmics Laboratory (ALGO)
Ecole Polytechnique Federale de Lausanne (EPFL)

March 14, 2012

1 Summary

In the last two weeks, my primary focus was on extending the proof of the learning algorithm which I mentioned in November 2011 reports to more general cases so that it becomes ready for the journal paper. Corresponding to the proof, I read two elegant papers on proving the convergence of some iterative neural learning algorithms. I also started writing the text for the journal version of our ITW paper. Finally, I spent some time organizing MATLAB codes for simulating the learning algorithm and started implementing the idea of having a neural network which *adaptively* learns the constraints, i.e. starting with a few constraints and based on the performance learning more if needed.

In this report, I am going to explain the steps I have taken to prove the convergence of the sparse learning algorithm and the summary of the two papers I have read.

2 Journal Paper

2.1 Summary of the Learning Algorithm

In order to develop a simple iterative algorithm, we formulate the problem as an optimization framework and then use primal-dual approaches to iteratively find the solution. The problem to find a constraint vector W is given by equation (1).

$$\min \sum_{\mu=1}^{C} |x^{\mu} \cdot w|^2 \tag{1a}$$

subject to

$$\|w\|_0 \leq q \tag{1b}$$

and

$$\|w\|_2^2 \geq \epsilon \tag{1c}$$

where $q \in \mathbb{N}$ determines the degree of sparsity and $\epsilon \in \mathbb{R}^+$ prevents the all-zero solution.

Therefore, we first relax (1) as follows:

$$\min \sum_{\mu=1}^{C} |x^{\mu} \cdot w|^2 + \lambda(g(w) - q'). \tag{2a}$$

subject to:

$$\|w\|_2^2 \geq \epsilon \tag{2b}$$

In the above problem, we have approximated the constraint $\|w\|_0 \leq q$ with $g(w) \leq q'$ since $\|.\|_0$ is not a well-behaved function. The function $g(w)$ is chosen such that it favors sparsity. For instance one can pick $g(w)$ to be $\|.\|_1$, which leads to ℓ_1-norm minimizations. In this paper, we consider the function

$$g(w) = \sum_{i=1}^{n} \tanh(\sigma w_i^2)$$

where σ is chosen appropriately. The larger σ is, the closer $g(w)$ will be to $\|.\|_0$. By calculating the derivative of the objective function and primal-dual optimization techniques we obtain the iterative

Algorithm 1 Iterative Learning

Input: Set of patterns x^μ with $\mu = 1, \ldots, C$, stopping point p.
Output: w
  while $\max_\mu |y(\mu, t)| > p$ **do**
    Compute $y(\mu, t) = x^\mu.w(t)$
    Update $w(t+1) = w(t) - \alpha_t y(\mu, t) \left(x^\mu - \frac{y(\mu,t)w(t)}{\|w(t)\|^2} \right) - \lambda_t f(w(t))$.
    Update $\lambda_{t+1} = \left[\lambda_t + \delta(\epsilon - \|w\|_2^2) \right]$.
    $t \leftarrow t + 1$.
  end while

algorithm given by algorithm 1. Where $f(w) : \mathcal{R}^n \rightarrow \mathcal{R}^n = \nabla g(w)$ is the gradient of the penalty term for non-sparse solutions. Since we are interested in finding m orthogonal vectors, we have to do the above procedure m times in parallel.

Therefore, in order to prove the convergence, in each iteration we must show the MSE decreases. More specifically, denoting the MSE in iteration t by $E(t)$, we must show that:

$$E(t) = \frac{1}{C} \sum_{\mu=1}^{C} |w(t)^T x^\mu|^2 < E(t-1) \tag{3}$$

In what follows, I will explain the approaches which might help us in proving (3).

2.2 Approach 1

In the first approach, I am going to exploit the fact that the convergence of weight vectors to the corresponding null bases has been proven before in [4]. Since our algorithm is basically the algorithm given in [4] plus an additional sparsity constraint, we might be able to show that our algorithm converges if the algorithm mentioned in [4] converges. To this end, we show that $E(t)$ in our approach is less than or equal to the MSE at iteration t of the Anti Stochastic Gradient Ascent (ASGA) algortihm proposed in [3], which is an extension of the one in [4]. Let $w'(t) = w(t-1) - \alpha_t y(t-1)x(t)$, with $y(t) = w(t)^t x(t)$. Therefore, $w'(t)$ identifies the learning algorithm mentioned in [3], with normalizaion constraints being omitted for simplicity. Now, our learning algorithm can be defined as $w(t) = \eta(w'(t); \theta_t)$. As a result, if we can prove the following relationship, we have proven that in all the circumstances that the ASGA algorithm converges, ours converges as well. Suppose in iteration t, we have learned the pattern x'. Then, we would like to prove that:

$$P_1 = Pr\{E(t+1) = \mathbb{E}_x |\eta(w'(t); \theta_{t+1})^T x|^2 < E'(t+1) = \mathbb{E}_x |w'(t)^T x|^2\} \geq 1 - \epsilon, \ \forall x' \in \mathcal{X} \tag{4}$$

With ϵ being a small real number. In words, we claim that with high probability, the MSE of our algorithm in iteration t, $E(t)$, is less than or equal to the MSE of the ASGA algorithm in the same iteration, $E'(t)$.

Since we draw the patterns x' uniformly at random, we relax (4) into the following relationship:

$$P_2 = Pr\{E(t+1) = \mathbb{E}_{x'} \mathbb{E}_x |\eta(w'(t); \theta_{t+1})^T x|^2 < E'(t+1) = \mathbb{E}_{x'} \mathbb{E}_x |w'(t)^T x|^2\} \geq 1 - \epsilon \tag{5}$$

where $w'(t) = w(t-1) - \alpha_t y(t-1)x'$, with $y(t-1) = \sum_{j=1}^{n} w_j(t-1)x'_j$. Assming entries in $w(0)$ are i.i.d. Gaussian random variables and the entries of pattern vectors are i.i.d. over some probability distribution, we take the following steps to prove (5):

1. $w'_j(t)$ and x_j are uncorrelated, i.e. $\mathbb{E}\{w'_j(t)x_j\} = \mathbb{E}\{w'_j(t)\}\mathbb{E}\{x_j\}$.

2. Letting k denote the number of entries in $w'(t)$ whose absolute values are smaller than θ_t, we formulate (5) as an expectation over $k(t)$.

3. Letting $\bar{w}'(t) = \mathbb{E}_x\mathbb{E}_{x'}(w'(t))$, we show that the entries in $\bar{w}'(t)$ are correlated Gaussian random variables with mean $\mu_w(t)$, variance $\sigma_w(t)$ and correlation coefficient $\rho_w(t)$.

4. We find a closed form relationship for (5) conditioned on k, the number of *small* entries in $w'(t)$.

5. We analyze the assymptotic behavior of (5) as n tends to infinity and show that it holds for large enough n's.

2.2.1 Step 1: Proving uncorrelatedness

to start, we show that if the entries in x (or x') and $w(t-1)$ are independent, so will be the entries in x and $w'(t)$. We have:

$$\begin{aligned}
\mathbb{E}_x\{x_i w'_i(t)\} &= \mathbb{E}_x\{x_i w_i(t-1)\} - \alpha_t \mathbb{E}_x\{x_i x'_i \sum_{j=1}^{n} x'_j w_i(t-1)\} \\
&= \mathbb{E}_x\{x_i\}\mathbb{E}_x\{w_i(t-1)\} - \alpha_t \mathbb{E}_x\{x_i x'_i \sum_{j=1}^{n} x'_j w_i(t-1)\} \\
&= \mathbb{E}_x\{x_i\}\mathbb{E}_x\{w_i(t-1)\} - \alpha_t \mathbb{E}_x\{x_i\} x'_i \sum_{j=1}^{n} x'_j \mathbb{E}_x\{w_i(t-1)\} \\
&= \mathbb{E}_x\{x_i\}\left(\mathbb{E}_x\{w_i(t-1)\} - \alpha_t\{x_i\} x'_i \sum_{j=1}^{n} x'_j \mathbb{E}_x\{w_i(t-1)\}\right) \\
&= \mathbb{E}_x\{x_i\}\mathbb{E}_x\{w'_i(t)\} \qquad (6)
\end{aligned}$$

where the first equality follows from the assumption that x_i and $w_i(t-1)$ are uncorrelated. The second equality is the result of the fact that the pattern vectors are generated independently randomly.

Therefore and as a result of (6), if we generate the entries of $w(0)$ randomly and independent of x, we will maintain the uncorrelated assumption throughout the algorithm.

2.2.2 Step 2: Breaking equation (5) into smaller parts

Letting k denote the number of entries in $w'(t)$, we can simplify equation (5) as follows:

$$\begin{aligned}
P_1 &= Pr\{E(t+1) = \mathbb{E}_{x'}\mathbb{E}_x|\eta(w'(t);\theta_{t+1})^T x|^2 < E'(t+1) = \mathbb{E}_{x'}\mathbb{E}_x|w'(t)^T x|^2\} \\
&= \sum_{k=0}^{n} P_k Pr\{\mathbb{E}_{x'}\mathbb{E}_x|\sum_{j=1}^{n-k} w'_j(t)x_j|^2 < \mathbb{E}_{x'}\mathbb{E}_x|\sum_{j=1}^{n} w'_j(t)x_j|^2\} \qquad (7)
\end{aligned}$$

where P_k is the probability of having k entries in $w'(t)$ whose absolute values are smaller than θ_t. Now consider the following definitions:

$$A = \sum_{j=1}^{n-k} w'_j(t)x_j \tag{8}$$

$$B = \sum_{j=n-k+1}^{n} w'_j(t)x_j \tag{9}$$

Therefore, equation (7) can be rewritten as:

$$\begin{aligned} P_1 &= Pr\{\mathbb{E}_{x'}\mathbb{E}_x|A|^2 \leq \mathbb{E}_{x'}\mathbb{E}_x|A+B|^2\} \\ &= Pr\{0 \leq \mathbb{E}_{x'}\mathbb{E}_x(B^2+2AB)\} \\ &= Pr\{-2\mathbb{E}_{x'}\mathbb{E}_x(AB) \leq \mathbb{E}_{x'}\mathbb{E}_x(B^2)\} \end{aligned} \tag{10}$$

Now using the fact that $\mathbb{E}_x(B^2) \geq (\mathbb{E}_x(B))^2$, we see that $Pr\{-2\mathbb{E}_{x'}\mathbb{E}_x(AB) \leq \mathbb{E}_{x'}\mathbb{E}_x(B^2)\} \geq Pr\{-2\mathbb{E}_{x'}\mathbb{E}_x(AB) \leq (\mathbb{E}_{x'}\mathbb{E}_x(B))^2\}$. Therefore, in order to have (5) it is sufficient to have $P_2 \geq 1-\epsilon$, where P_2 is defined as follows:

$$\begin{aligned} P_2 &= \sum_{k=0}^{n} P_k Pr\{(\mathbb{E}_{x'}\mathbb{E}_x(B))^2 \geq -2\mathbb{E}_{x'}\mathbb{E}_x(AB)\} \\ &= \sum_{k=0}^{n} P_k Pr\{(\mathbb{E}_{x'}\mathbb{E}_x(B))^2 \geq -2\mathbb{E}_{x'}\mathbb{E}_x(A)\mathbb{E}_{x'}\mathbb{E}_x(A)\} \\ &= \sum_{k=0}^{n} P_k \left[Pr\{\mathbb{E}_{x'}\mathbb{E}_x(B) > 0\}Pr\{(\mathbb{E}_{x'}\mathbb{E}_x(B+2A) \geq 0\} + Pr\{(\mathbb{E}_{x'}\mathbb{E}_x(B) < 0\}Pr\{(\mathbb{E}_{x'}\mathbb{E}_x(B+2A) \leq 0\}\right] \\ &= \sum_{k=0}^{n} P_k \left[Pr\{\bar{B} > 0\}Pr\{\bar{B}+2\bar{A} \geq 0\} + Pr\{\bar{B} < 0\}Pr\{\bar{B}+2\bar{A} \leq 0\}\right] \end{aligned} \tag{11}$$

Where the second equality follows from the fact that the index of entries in A and B is non-overlapping and since x_i's are chosen i.i.d. at random, A and B are uncorrelated when it comes to expectations over x and x'. In addition, $\bar{B} = \mathbb{E}_{x'}\mathbb{E}_x(B)$ and $\bar{A} = \mathbb{E}_{x'}\mathbb{E}_x(A)$.

2.2.3 Step 3: Proving that $\bar{w'}$ is Gaussian

Letting $\bar{w}'(t) = \mathbb{E}_x\mathbb{E}_{x'}(w'(t))$ we see that:

$$\bar{w}'_i(t) = \mathbb{E}_x\mathbb{E}_{x'}(w'_i(t)) = (1-\alpha\tau_x)\bar{w}_i(t-1) - \alpha_t(\mu_x)^2 \sum_{j\neq i} \bar{w}_j(t-1) \tag{12}$$

Where $\tau_x = \mathbb{E}_x\{(x'_i)^2\}$. Therefore, if $\{\bar{w}_j(t-1)\}$ are gaussian random variables with mean $\mu_w(t-1)$, variance $\sigma_w^2(t-1)$ and covariance $\nu_w(t-1)$, $\{\bar{w}'_i(t)\}$ are also correlated Gaussian random variables with:

$$\mu_{w'}(t) = \left(1 - \alpha_t\tau_x - \alpha_t(n-1)(\mu_x)^2\right)\mu_w(t-1) \tag{13a}$$

$$\begin{aligned}\sigma^2_{w'}(t) &= \left((1-\alpha_t\tau_x)^2+(n-1)\alpha_t^2(\mu_x)^4\right)\sigma^2_w(t-1)\\ &+ \left((n-1)(n-2)\alpha_t^2(\mu_x)^4-2(n-1)\alpha_t(\mu_x)^2(1-\alpha_t\tau_x)\right)\nu_w(t-1) \qquad (13b)\end{aligned}$$

$$\begin{aligned}\nu_{w'}(t) &= \left((n-1)\alpha_t^2(\mu_x)^4-2\alpha_t\mu_x^2(1-\alpha_t\tau_x)\right)\sigma^2_w(t-1)\\ &+ \left((1-\alpha_t\tau_x)^2+(n-1)(n-2)\alpha_t^2(\mu_x)^4-2(n-2)\alpha_t(\mu_x)^2(1-\alpha_t\tau_x)\right)\nu_w(t-1) \qquad (13c)\end{aligned}$$

2.2.4 Step 4: closed from formula

From the definition of $\bar{A}$ and $\bar{B}$ we know that:

$$\bar{A} = \mu_x \sum_{j=1}^{n-k} \bar{w}'_j(t) \qquad (14)$$

and

$$\bar{B} = \mu_x \sum_{j=n-k+1}^{n} \bar{w}'_j(t) \qquad (15)$$

Knowing that $\{\bar{w}'_i(t)\}$ are correlated Gaussian random variables, we see that $\bar{A}$ and $\bar{B}$ are also Gaussian random variables with:

$$\mu_{\bar{A}}(t) = (n-k)\mu_x\mu_{w'}(t) \qquad (16a)$$

$$\mu_{\bar{B}}(t) = k\mu_x\mu_{w'}(t) \qquad (16b)$$

$$\sigma^2_{\bar{A}}(t) = (\mu_x)^2\left((n-k)\sigma^2_{w'}+(n-k)(n-k-1)\nu_{w'}(t)\right) \qquad (16c)$$

$$\sigma^2_{\bar{B}}(t) = (\mu_x)^2\left(k\sigma^2_{w'}+k(k-1)\nu_{w'}(t)\right) \qquad (16d)$$

Similarly, if one defines $\bar{D} = \bar{B} + 2\bar{A}$, it is obvious that $\bar{D}$ is also a Gaussain random variable with

$$\mu_{\bar{D}}(t) = (2n-k)\mu_x\mu_{w'}(t) \qquad (17a)$$

$$\sigma^2_{\bar{D}}(t) = (\mu_x)^2\left((4n-3k)\sigma^2_{w'}+\left(n^2+2nk-2k^2-n\right)\nu_{w'}(t)\right) \qquad (17b)$$

As a result, one can easily see that:

$$\begin{aligned}P_3 &= Pr\{\bar{B}\geq 0|k\} = 1-Q\left(\frac{\mu_{\bar{B}}(t)}{\sigma_{\bar{B}}(t)}\right)\\ &= 1-Q\left(\frac{k\mu_{w'}(t)}{\sqrt{k\sigma^2_{w'}+k(k-1)\nu_{w'}(t)}}\right)\\ &= 1-Q\left(\frac{\mu_{w'}(t)\sqrt{k}}{\sqrt{\sigma^2_{w'}+(k-1)\nu_{w'}(t)}}\right) \qquad (18)\end{aligned}$$

likewise:

$$\begin{aligned}P_4 &= Pr\{\bar{D}\geq 0|k\} = 1-Q\left(\frac{\mu_{\bar{D}}(t)}{\sigma_{\bar{D}}(t)}\right)\\ &= 1-Q\left(\frac{(2n-k)\mu_{w'}(t)}{\sqrt{(4n-3k)\sigma^2_{w'}+\left(n^2+2nk-2k^2-n\right)\nu_{w'}(t)}}\right) \qquad (19)\end{aligned}$$

Combining the above formulas, we can rewrite (5) as:

$$\begin{aligned} P_1 &= Pr\{\mathbb{E}_x|\eta(w'(t);\theta_{t+1})^T x|^2 < \mathbb{E}_x|w'(t)^T x|^2\} \\ &\geq \sum_{k=0}^{n} P_k \left[Pr\{\bar{B}>0\}Pr\{\bar{B}+2\bar{A}\geq 0\} + Pr\{\bar{B}<0\}Pr\{\bar{B}+2\bar{A}\leq 0\}\right] \\ &= \sum_{k=0}^{n} P_k \left[P_3P_4 + (1-P_3)(1-P_4)\right] \\ &= 1 - \sum_{k=0}^{n} P_k \left[2P_3P_4 - P_3 - P_4)\right] \end{aligned} \tag{20}$$

Where P_k is the probability that k entries in $w'(t)$ has absolute values smaller than θ_t. Now since $\{w'_i(t)\}$ are correlated Gaussian variables,

2.3 Current Issues

There are some major issues with the approach discussed above:

1. Equation (20) is too difficult to be handled analytically.
2. Even if it was possible to handle (20), the assumption $w_j(t)$'s being Gaussian is questionable as they are capped Gaussian random variables, i.e. those that are put to zero if their absolute value falls short of some threshold.
3. Statisticall properties in equation (18) and (19), such as $\sigma_{w'}$ of $\mu_{w'}$, depend on n and k as well. Therefore, in order to analyze the asymptotic behavior of equation (20) for $n \to \infty$ we must incorporate this dependence as well.

3 Papers

In the past two weeks, I also read two papers on proving the converges of learning algorithms similar to the one we use in our journal paper [7]. Both papers discuss a learning algorithm to learn the first largest eigenvectors of one or many subspaces from the sample vectors drawn from those subspaces.

3.1 Learning One Subspace

In [4], the authors prove the convergence of the Stochastic Gradient Ascent (SGA) learning algorithm to the largest eigenvectors of the correlation matrix in a given data set. This paper is of particular importance as the same approach is used in many learning methods such as [3] to identify the null basis of a set of given patterns, which were used later as a classification method. The authors use stochastic approximation techniques to prove the convergence of the proposed learning method and estimate the largest eigenvalues as well.

More specifically, the problem of interest is as follows: consider an $n \times n$ *almost surely symmetric real* matrix whose mean is denoted by $A \in \mathbb{R}^{n\times n}$. The goal is to compute the dominant eigenvectors and eigenvalues of this matrix in a situation where A is unknown, but a sequence of samples $\{A_k\}$

are available, where $A = \mathbb{E}\{A_k\}$. If A_k's were general, then one should use standard techniques such a the QR method. However, when A_k's are correlation matrices, i.e. they have the form $A_k = x_k x_k^T$, where $\{x_k\}$ is a random sequence of vectors, then an iterative method which updates the estimates each time a new vector arrives is simpler and have computational advantage.

The authors propose a learning algorithm (given by equation (21)) and prove the almost sure convergence of the weight vectors to the first largest eigenvectors of the correlation matrix A.

$$\tilde{W}_k = W_{k-1} + A_k W_{k-1} \Gamma_k \tag{21a}$$

$$W_k = \tilde{W}_k R_k^{-1} \tag{21b}$$

Now, the algorithm (21) translates into the following equation when we only consider one weight vector w_k in iteration k:

$$\tilde{w}_k = w_{k-1} + \gamma_k x_k x_k^T w_{k-1} \tag{22a}$$

$$w_w = \tilde{w_k} / \|\tilde{w}_k\| \tag{22b}$$

If γ_k is small enough, the normalization factor in equation (22b) can be absorbed into the main learning equation. Therefore, we might write:

$$w_k = w_{k-1} + \gamma_k [A_k w_{k-1} - (w_{k-1}^T A_k w_{k-1}) w_{k-1}] + \gamma_k b_k \tag{23}$$

where $b_k = O(\gamma_k)$.

The idea behind the proof of the convergence is as follows:

- First we can rewrite the learning equation (22) in terms of a differential equation. Note that there are two terms (given in equation (24)) plus a stochastic small term which is $O(\gamma_k)$, where γ_k is the step size in the learning algorithm.

$$\frac{dz}{dt} = Az - \frac{(z^T A z) z}{z^T z} \tag{24}$$

- The authors utilize the fact that this stochastic term is bounded and has some conditions which results in the corresponding differential equation to have a domain of attraction. Then, if w_k falls infinitely often in this domain with probability one, then the weight vector w_k converges to the solution of this differential equation (this is **Lemma 1**).

- Now the interesting thing is that the solution of this differential equation converges to largest eigenvectors of the correlation matrix *if* the projection of w_k onto the largest eigenvector is bounded away from zero infinitely often (this is **Lemma 2**).

- Finally the authors show that the required projection constraint above holds for the set of assumptions on γ_k's and the correlation matrix A (This is **Lemma 3**).

Given that we use a similar approach in [7] in order to learn the *sparse* null basis for a set of given patterns, this paper is of outmost importnace for our work.

3.2 Learning many subspaces

In [5], the authors propose a learning algorithm, which is very similar to the *Learning Subspace Method* (LSM) by Kohonen et al. [6]. The authors prove the almost sure convergence for this method. The difference between this work and [4] is that here, we have K classes, $\omega^{(1)}, \ldots, \omega^{(K)}$, which are reprented by subspaces $\{\mathbb{L}^{(i)}\}$, for $i = 1, \ldots, K$. We would like to learn the basis vectors of these subspaces (denoted by the first largest eigenvectors) using neural learning algorithms.

Let the patterns belonging to class i be defined by $\{x_k^{(i)}\}$, with $i = 1, \ldots, K$ and k being the pattern index. We have $x_k^{(i)} \in \mathbb{R}^n$. Now let $U_{k-1}^1, \ldots, U_{k-1}^K$ be a set of matrices, where the columns of U_{k-1}^i are orthonormal vectors which span the subspace defined by the set of vectors $x_0^{(i)}, \ldots x_{k-1}^{(i)}$, i.e.:

$$\mathbb{L}_{k-1}^{(i)} = \mathcal{R}(U_{k-1}^i) \tag{25}$$

where $\mathcal{R}$ defines the range of a matrix.

The proposed learning algorithm by the authors is as follows: At step k of the training phase, assume the input vector x belongs to class i, i.e. $x = x_k^{(i)}$. Then do the following:

$$\tilde{U}_k^{(i)} = (I + \mu_k^{(i,i)} x_k^{(i)} x_k^{(i)T}) U_{k-1}^{(i)} \tag{26a}$$

$$\tilde{U}_k^{(j)} = (I + \mu_k^{(j,i)} x_k^{(i)} x_k^{(i)T}) U_{k-1}^{(j)}, \forall j \neq i \tag{26b}$$

$$U_k^{(i)} = \tilde{U}_k^{(i)} V_k^{(i)}, \forall i \tag{26c}$$

Where the matrices $V_k^{(i)}$ perform Gram-Schmidt orthonormalization.

In words, equation (26a) *rotates* the basis vectors in $U_{k-1}^{(i)}$ more towards the correct subspace and equation (26b) rotates $U_{k-1}^{(j)}$ to minimize the projection of $x^{(i)}$ on $\mathbb{L}^{(j)}$.

In [6], all $\mu_k^{(j,i)}$'s are zero if pattern $x_k^{(i)}$ is correctly classified in $\mathbb{L}^{(i)}$. Otherwise, if the pattern is wrongly classified in $\mathbb{L}^{(j)}$, then only $\mu_k^{(j,i)}$ and $\mu_k^{(i,i)}$ are non-zero and all other coefficients are equal to zero. However, in this paper the authors assume the coefficients are independent of the intermediate results (i.e. classification results during the learning phase).

To prove the convergence of the suggested algorithm, the authors use stochastic approximation techniques again to write the projection matrix at iteration k ($P_k = U_k^{(1)} U_k^{(1)T}$) of the algorithm as a function of the projection matrix in iteration $k-1$, which is given by the following equation:

$$P_k = P_{k-1} + \mu_k [P_{k-1}\bar{C} + \bar{C}P_{k-1} - 2P_{k-1}\bar{C}P_{k-1}] \tag{27}$$

Then, the authors formulate equation (27) in terms of the differential equation, given below, and prove that P_k converges to the solution of this equation.

$$\frac{dP}{dt} = P\bar{C} + \bar{C}P - 2P\bar{C} \tag{28}$$

where P is a matrix. Finally, the authors show that the solution to (28) is

$$P(t) = e^{\bar{C}(t-t_0)} V (V^T e^{2\bar{C}(t-t_0)} V)^{-1} V^T e^{\bar{C}(t-t_0)} \tag{29}$$

with V being a $n \times p^{(1)}$ matrix such that $VV^T = P(t_0)$ and $V^TV = I$. Then they show that $P(t)$ converges to $\bar{F}$, the orthogonal projection matrix on the subspace $\mathcal{M}^{(1)}$, as t tends to ∞, where $\bar{C}$ is defined below and $\mathcal{M}^{(i)}$ is the subspace spanned by the eigenvectors of $\bar{C}^{(i)}$ corresponding to the $p^{(i)}$ largest eigenvalues. There is a number $\epsilon > 0$ such that in equation (26), the event $\{\|U_k^{(i)T} z\| \geq \epsilon, \forall z \in \mathcal{M}^{(i)} with \|z\| = 1\}$ occurs infinitely often with probability one.

$$\bar{C}^{(i)} = \theta^{(i,i)}\pi^{(i)}C^{(i)} - \sum_{j \neq i} \pi^{(j)}\theta^{(i,j)}C^{(j)} \tag{30}$$

in which $\mu^{(j,k)} = \theta^{(i,j)}\mu_k$.

4 Future Works

For the moment, the proof of convergence of the learning algorithm in our paper [7] remains unsolved and needs further work.

References

[1] D. L. Donoho, A. Maleki, A. Montanari, *Message passing algorithms for compressed sensing*, Proc. Nat. Acad. Sci., Vol. 106, 2009, pp. 1891418919.

[2] K.R. Kumar, A.H. Salavati and A. Shokrollahi, *Exponential pattern retrieval capacity with non-binary associative memory*, Proc. IEEE Information Theory Workshop, 2011.

[3] L. Xu, A. Krzyzak, E. Oja, *Neural nets for dual subspace pattern recognition method*, Int. J. Neur. Syst., Vol. 2, No. 3, 1991, pp. 169-184.

[4] E. Oja, J. Karhunen, *On stochastic approximation of eigenvectors and eigenvalues of the expectation of a random matrix*, J. Math. Analysis and Applications, Vol. 106, No. 1, 1985, pp 6984.

[5] E. Oja, J. Karhunen, *An analysis of convergence for a learning version of the subspace method*, J. Math. Analysis and Applications, Vol. 91, No. 1, 1983, pp. 102111.

[6] T. Kohonen, G. Nemeth, K. Bry, M. Jalanko, H. Riittine, *Spectral classification of phonemes by learning subspaces*, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 1979, pp. 97-100.

[7] K.R. Kumar, A.H. Salavati and A. Shokrollahi, *Exponential pattern retrieval capacity with non-binary learning associative memory*, Under preparation.