Progress Report
3-19 January 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)

January 20, 2012

1 Summary

In the last two weeks or so, I was mainly busy with writing progress reports for previous months and documenting the MATLAB codes that we have. Besides that, together with Mr. Amin Karbasi, we were able to follow some ideas on learning a sparse vector that is orthogonal to a set of given training vectors. The algorithm is a combination of the message passing algorithm for compressed sensing by Donoho et al. [1] and that of Xu et al. for learning an orthogonal vector to a set of patterns in the training set [3]. In what follows, you can find more details about this idea.

2 Learning a sparse null basis

In our previous report, we proposed an algorithm based on the message passing metthod for compressed sensing in [1]. However, since in our special case we wanted to learn a vector that is orthogonal to a set of training patterns, then obviously the all-zero vector is a solution to this problem and it is the sparset one! Since this solution is not acceptable in our case, we added a penalty for the all-zero vector by requiring the ℓ_2-norm of the solution be greater than a constant ϵ. As a result, we formulated the problem as follows:

$$\min \|X.w\|_2 \tag{1a}$$

subject to

$$\|w\|_1 \leq q \tag{1b}$$

and

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

where $X_{\mathcal{S}\times n}$ is composed of non-negative integer entries, and rows of X constitute the patterns that belong to a subspace with dimension $k < n$. We are interested in finding a vector $w \in \mathcal{R}^n$ such that $X.w = \mathbf{0}$, where $\mathbf{0}$ is the all-zero vector. Finally, q determines the degree of sparsity and the constraint (1c) ensures that the algorithm will not converge to the all-zero solution.

Based on the above problem, we proposed the following iterative learning algorithm:

$$y(t) = \frac{X.w(t)}{\|X\|_2} \tag{2a}$$

$$w(t+1) = \eta\left((1+2\lambda_t)w(t) - 2\alpha_t \frac{X^T y(t)}{\|X\|_2}\right)_{\theta_t} \tag{2b}$$

subject to

$$\lambda_{t+1} = \left[\lambda_t + \gamma(\epsilon - \|w\|_2^2)\right] \tag{2c}$$

where t denotes the iteration number, X^T is the transpose of matrix X, γ and α_t are small step sizes, $[.]_+$ denotes $\max(.,0)$, θ_t is a positive threshold at iteration t and $\eta(.)_{\theta_t}$ is the *point-wise* soft-thresholding function given below:

$$\eta(u)_\theta = \begin{cases} u & \text{if } u > \theta; \\ u & \text{if } u < -\theta \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

2.1 Neural algorithms for learning the null vector

On the other hand, the approach proposed in [3] is designed for learning null vectors and ,hence, avoiding the all-zero solution has already taken care of in the weight update update rule. Their update rule in each iteration and in response to a training pattern x^μ is as follows:

$$y(\mu, t) = x^\mu.w(t) \tag{4a}$$

$$w(t+1) = w(t) - \alpha_t y(\mu, t)\left(x^\mu - \frac{y(\mu, t)w(t)}{\|w(t)\|^2}\right) \tag{4b}$$

The interesting idea in the above update rule is that $\Delta w(t) = w(t+1) - w(t)$ is orthogonal to $w(t)$. Hence, $\|w(t+1)\|_2 = \|w(t)\|_2 + \|\Delta w(t)\|_2$. Therefore, $\|w(t+1)\|_2 \geq \|w(t)$ which means if you start from a non-zero initial guess, you will not converge to the all-zero solution.

Now we would like to combine the two ideas to propose a new algorithm that does not need the penalty contraint as in equation (1c).

2.2 Combination of the two idea

Based on the above two ideas, we propose the following algorithm and try to prove its convergence.

$$y(t) = \frac{X.w(t)}{\|X\|_2} \tag{5a}$$

$$w(t+1) = \eta\left(w(t)\left(1 + 2\alpha_t \frac{\|y(t)\|_2^2}{\|w(t)\|_2^2}\right) - 2\alpha_t \frac{X^T y(t)}{\|X\|_2}\right)_{\theta_t} \tag{5b}$$

2.3 Avoiding the all-zero solution

We first have to show that the above update rule does not lead to the trivial all-zero soluion. To show that, we start by calculating $\|w(t+1)\|_2^2$. Let $w'(t) = (1 + 2\alpha_t \frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})w(t) - 2\alpha_t \frac{X^T y}{\|X\|_2}$. Furthermore, let the function $\chi(u; \theta_t)$ be $u - \eta(u)_{\theta_t}$. Rewriting equation (5b) we will have:

$$w(t+1) = w'(t) - \chi(w'(t); \theta_t) \tag{6}$$

Therefore:

$$\|w(t+1)\|^2 = \|w'(t)\|^2 + \|\chi(w'(t); \theta_t)\|^2 - 2(w'(t))^T \chi(w'(t); \theta_t)$$

Noting that the function $\chi(w'_t(t); \theta_t) = w'_i(t)$ if $|w'_i(t)| \leq \theta_t$ and $\chi(w'_t(t); \theta_t) = 0$ if $|w'_i(t)| > \theta_t$ we find out that $(w'(t))^T \chi(w'(t); \theta_t) = \|\chi(w'(t); \theta_t)\|^2$. Therefore:

$$\|w(t+1)\|^2 = \|w'(t)\|^2 - \|\chi(w'(t); \theta_t)\|^2 = \sum_{i=1}^{n} [w'_i(t)^2 - \chi(w'_i(t))^2]$$

Now since $[w'_i(t)^2 - \chi(w'_i(t))^2] = 0$ if $|w'_i(t)| \leq \theta_t$ and $[w'_i(t)^2 - \chi(w'_i(t))^2] = w'_i(t)^2$ if $|w'_i(t)| > \theta_t$, then we conclude that $\|w(t+1)\|^2 \neq 0$ if and only if there is at least one entry in $w'(t)$ whose

absolute value is bigger than θ_t. Now expanding $w'(t)$ and replacing the value of $y(t)$ from equation (5a) we find out:

$$\begin{aligned} w'(t) &= (1+2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})w(t) - 2\alpha_t\frac{X^Ty}{\|X\|_2} \\ &= \left((1+2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})\mathcal{I}_{n\times n} - 2\alpha_t\frac{X^TX}{\|X\|_2^2}\right)w(t) \\ &= B_tw(t) \end{aligned} \tag{7}$$

Where $B_t = (1+2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})\mathcal{I}_{n\times n} - 2\alpha_t\frac{X^TX}{\|X\|_2^2}$. Since $w(t)$ is the result of previous updates, we know that its entries are either zero or has an absolute value larger than θ_{t-1}. Let us assume $w(t)$ is κ_t-sparse, i.e. it has κ_t non-zero entries. Furthermore, without loss of generality we assume that $\|w(t)\|_{\max} = |w_1(t)|$. Now let $A = X^TX$. Expanding $w_1'(t)$ we will have:

$$w_1'(t) = w_1(t)\left(1+2\alpha_t\frac{\|Xw(t)\|_2^2}{\|w(t)\|_2^2\|X\|_2^2} - 2\alpha_t\frac{A_{11}}{\|A\|_2}\right) - 2\alpha_t\sum_{j=2}^{n}w_j(t)\frac{A_{1j}}{\|A\|_2} \tag{8}$$

Therefore,

$$\begin{aligned} |w_1'(t)| &= |w_1(t)\left(1+2\alpha_t\frac{\|Xw(t)\|_2^2}{\|w(t)\|_2^2\|X\|_2^2} - 2\alpha_t\frac{A_{11}}{\|A\|_2}\right) - 2\alpha_t\sum_{j=2}^{n}w_j(t)\frac{A_{1j}}{\|A\|_2}| \\ &\geq |w_1(t)\left(1+2\alpha_t\frac{\|Xw(t)\|_2^2}{\|w(t)\|_2^2\|X\|_2^2} - 2\alpha_t\frac{A_{11}}{\|A\|_2}\right)| - 2\alpha_t\sum_{j=2}^{n}|w_j(t)\frac{A_{1j}}{\|A\|_2}| \\ &\geq |w_1(t)|(1-2\alpha_t) - 2\alpha_t\sum_{j=2}^{n}|w_j(t)||\frac{A_{1j}}{\|A\|_2}| \\ &\geq |w_1(t)|(1-2\alpha_t) - 2\alpha_t\sum_{j=2}^{n}|w_j(t)| \\ &\geq |w_1(t)|(1-2\alpha_t) - 2\alpha_t(\kappa_t-1)|w_1(t)| \\ &= (1-2\alpha_t\kappa_t)|w_1(t)| \end{aligned} \tag{9}$$

The second inequality is true because $\frac{\|Xw(t)\|_2^2}{\|w(t)\|_2^2} \leq \|X\|_2^2$ and $\frac{A_{11}}{\|A\|_2} \leq 1$. Therefore, $1-2\alpha_t \leq 1+2\alpha_t\frac{\|Xw(t)\|_2^2}{\|w(t)\|_2^2\|X\|_2^2} - 2\alpha_t\frac{A_{11}}{\|A\|_2} \leq 1+2\alpha_t$. Assuming $\alpha_t \leq 0.5$, it means that $|1+2\alpha_t\frac{\|Xw(t)\|_2^2}{\|w(t)\|_2^2\|X\|_2^2} - 2\alpha_t\frac{A_{11}}{\|A\|_2}| \geq 1-2\alpha_t$. The third inequality follows because $\frac{A_{1j}}{\|A\|_2} \leq 1$. Finally, the fourth inequality follows because $|w_j(t)| \leq \|w(t)\|_{\max} = |w_1(t)|$.

Therefore, in order to have $|w_1'(t)| \geq \theta_t$, it is sufficient to have $(1-2\alpha_t\kappa_t)|w_1(t)| \geq \theta_t$. Which means we have to pick α_t according to the following inequality:

$$0 < \alpha_t < \min\left(0.5, \frac{1}{2\kappa_t}\left[1 - \frac{\theta_t}{\|w(t)\|_{\max}}\right]\right) \tag{10}$$

Note that $\frac{\theta_t}{\|w(t)\|_{\max}} < 1$ becuase $\theta_t < \theta_{t-1}$. And since we know that $\|w(t)\|_{\max} \geq \theta_{t-1}$, $\frac{\theta_t}{\|w(t)\|_{\max}} < 1$. Therefore, $1 - \frac{\theta_t}{\|w(t)\|_{\max}}$ is strictly positive and we always will have a proper choice of α_t. As a result, we can make sure that the algorithm does not converge to the all-zero solution.

2.4 Proof of convergence

Now we have to show that the algorithm converges to a proper solution. To this end, we let $E(t) = \|y(t)\|_{\max}$. We would like to show that $E(t+1) < E(t)$ for all iterations t. Following the same set of notations as before, we will have:

$$\begin{aligned}
E(t+1) &= \|y(t+1)\|_{\max} = \|\frac{X.w(t+1)}{\|X\|_2}\|_{\max} \\
&= \|\frac{X.w'(t)}{\|X\|} - \frac{X.\chi(w'(t);\theta_t)}{\|X\|_2}\|_{\max} \\
&\leq \frac{\|X.w'(t)\|_{\max}}{\|X\|_2} + \frac{\|X.\chi(w'(t);\theta_t)\|_{\max}}{\|X\|_2} \\
&\leq \frac{\|X.w'(t)\|_{\max}}{\|X\|_2} + \frac{\|X\|_{\max}\|\chi(w'(t);\theta_t)\|_{\max}}{\|X\|_2} \\
&\leq \frac{\|X.w'(t)\|_{\max}}{\|X\|_2} + \theta_t\frac{\|X\|_{\max}}{\|X\|_2} \\
&\leq \frac{\|X.w'(t)\|_{\max}}{\|X\|_2} + \theta_t \qquad (11)
\end{aligned}$$

where the last inequality follows because $\|X\|_{\max} \leq \|X\|_2$. Now expanding $\frac{X.w'(t)}{\|X\|}$ we will get

$$\begin{aligned}
\frac{X.w'(t)}{\|X\|_2} &= (1 + 2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})y(t) - 2\alpha_t\frac{XX^Ty(t)}{\|X\|_2^2} \\
&= \left((1 + 2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})\mathcal{I}_{\mathcal{S}\times\mathcal{S}} - 2\alpha_t\frac{XX^T}{\|X\|_2^2}\right)y(t) \qquad (12)
\end{aligned}$$

Letting $C_t = (1 + 2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2})\mathcal{I}_{\mathcal{S}\times\mathcal{S}} - 2\alpha_t\frac{XX^T}{\|X\|_2^2}$, we can further simplify inequality (11):

$$\begin{aligned}
E(t+1) &\leq \|C_ty(t)\|_{\max} + \theta_t \\
&\leq \|C_t\|_{\max}\|y(t)\|_{\max} + \theta_t \\
&= \|C_t\|_{\max}E(t) + \theta_t \qquad (13)
\end{aligned}$$

Therefore, in order to show that $E(t+1) < E(t)$, we must have $\|C_t\|_{\max} < 1$ and $\theta_t \to 0$ as $t \to \infty$. By setting $\theta_t = 1/t$ we will achieve the second requirement. Now to ensure $\|C_t\|_{\max} < 1$, we would like to have $|c_{ij}^t| < 1$ for all elements (i,j) of C_t. For non-diagonal elements this is not problem as

$$|-2\alpha_t\frac{D_{ij}}{\|D\|_2}| \leq 1 \Rightarrow |\alpha_t| < 0.5$$

Since $|\frac{D_{ij}}{\|D\|_2}| \leq 1$, where $D = XX^T$. For diagonal elements we have $C_{ii} = (1+2\alpha_t\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2}) - 2\alpha_t\frac{D_{ii}}{\|D\|_2}$. Letting $d_{\min} = \min_i D_{ii}/\|D\|_2$ and assuming $\frac{\|y(t)\|_2^2}{\|w(t)\|_2^2}) \leq e_{\max}$, we have $1 - 2\alpha_t \leq C_{ii} \leq 1 + 2\alpha_t(e_{\max} - d_{\min})$. Therefore, assumin $\alpha_t < 0.5$, if $e_{\max} \leq d_{\min}$ we will have convergence. Note that once for one iteration t we will intuitively have this condition for all iterations afterwards since $E(t+1) = \|y(t+1)\|_{max} \leq \|y(t)\|_{max} \leq \|y(t+1)\|_2$, which results in expecting that $\frac{\|y(t+1)\|_2^2}{\|w(t+1)\|_2^2}) \leq \frac{\|y(t)\|_2^2}{\|w(t)\|_2^2}) \leq e_{\max}$. Therefore, if we manage to pick a relatively good starting point, we will ensure convergence if $0 < \alpha_t < \min\left(0.5, \frac{1}{2\kappa_t}\left[1 - \frac{\theta_t}{\|w(t)\|_{\max}}\right]\right)$.

3 Future Works

The most important future work concerns completing the last step in the proof of convergence which we intuitively expressed. Besides that, tightening the upper bound in equation (10) is another subject which we have to address because having α as large as possible results in faster convergence.

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.