Progress Report
1-15 November 2011

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)

December 16, 2011

1 Summary

In the last two weeks, I read a couple of papers on compressed sensing and ℓ_1-norm minimization techniques because for the learning phase of our neural associative memory, we are interested in finding some sparse weight vectors that are orthogonal to a set of given patterns. If we put all such patterns in a big matrix $X_{M\times n}$, where M is the number of patterns and n is their length, then our problem is to find a *sparse* weight vector w such that $Xw = \mathbf{0}$. That's the reason I have been working on ℓ_1-norm minimization techniques because we have exactly the same problem in compressed sensing and there, in order to find a sparse answer, ℓ_1-norm minimization is used as ℓ_0-norm minimization is not possible. In this report, you can find the summary of the papers that I have read along with some comments and ideas.

I also have tried to implement different ideas on learning sparse matrices from training data set. Furthermore, I had several discussions with Mr. Amin Karbasi on the above topics, in the hope of extending some of the already available results to our case.

2 LASSO risk as convex programming

In [1] the authors address the linear regression problem subject to sparsity constraints, also known as LASSO [2]. More specifically, we have n vectors x_i of size m and n observations y_i, where each observation is a linear summation of the elements in each vectors, i.e. $y_i = \sum_{j=1}^{m} \beta_j x_{ij}$. Now we would like to find *a sparse* vector β such that $(y - \beta x)^2$ is minimized. Mathematically speaking, here is the problem:

$$\min_{\beta_1,\dots,\beta_m} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 \tag{1a}$$

subject to:

$$\sum_{j=1}^{m} |\beta_j| \leq t \tag{1b}$$

Although the problem addressed in this paper is not exactly the same as what we are interested in but the method may be helpful for our case as well. However, the suggested optimization approach does not seem feasible in neural networks. The authors treat this problem as a convex programming problem:

$$\min_{\beta} f(\beta) = \frac{1}{2}(y - X\beta)^T(y - X\beta) = \frac{1}{2}r^T r \tag{2a}$$

subject to

$$g(\beta) = t - \sum_{i=1}^{m} |\beta_i| \geq 0 \tag{2b}$$

Here, $r = r(\beta)$ is the residual vector.

Since problem (2) is convex and all the critical points of $f(\beta)$ lie outside the area for which $t < t_0$, the solution to problem (2) lies at its boundaries, i.e. $\|\beta^*\|_1 = t$. To solve (2), the authors define the Lagrangian and use a primal-dual approach [1]. They also prove that if β^* is a solution of (1) for $m > n$, then β^* has <u>at most</u> n non-zero entries.

Finally, the authors have proposed two iterative algorithms to solve the desired optimization problem. These problems can be used even in cases where the number of parameters exceeds the number of observations. orks.

3 LASSO for Gaussian matrices

In [3], the authors consider the problem of learning *sparse* coefficients $x_0 \in \mathcal{R}^N$ from some *noisy* linear observation $y = Ax_0 + w$, where $y \in \mathcal{R}^n$. The main contribution of this paper is an expression for the mean square error of LASSO. The authors have been able to find such an expression for the first time when the measurement matrix A has iid Gaussian entries. Furthermore, using simulations they show that their analysis also works for other type of measurement matrices (both random and deterministic).

The idea behind finding such an expression is to show that the mean square error of the Approximate Message Passing (AMP) methods [6] and LASSO converge to the same value. Therefore, by finding an expression for the mean square error of AMP we will have it for LASSO as well.

More specifically, we have a set of linear measurements $y \in \mathcal{R}^n$ from an unknown vector $x_0 \in \mathcal{R}^N$, i.e. $y = Ax_0 + w$. We would like to find a *sparse* solution $\hat{x}$ such that the reconstruction error $\|x - x_0\|^2$ or $\|Ax - y\|^2$ is minimized. Here, $A \in \mathcal{R}^{n\times N}$ is the measurement matrix. LASSO is one way of solving the above problem when seeking sparse solutions. More specifically, in LASSO we have the following cost function:

$$C(x) = \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 \tag{3}$$

with $\lambda > 0$. Now the solution of LASSO is given by:

$$\hat{x}(\lambda, N) = argmin_x C(x) \tag{4}$$

There are lots of algorithms to solve problem (4) or to enhance the optimality of $\hat{x}$. However, mathematical understanding of the core problem is still not very deep. Some preliminary work on the bounds for $\|\hat{x} - x_0\|^2$ has been done in [4] and [5]. The authors address this shortcoming by analytically investigating problem (4). The idea is to show that the mean square error of the Approximate Message Passing (AMP) [6] and LASSO converge to the same value. To this end, we have the following lemmas and theorem:

Lemma [3]: Under the conditions of theorem 1.4 and for $\lambda > 0$ let $\hat{x}(\lambda, N)$ and $\{x(t)\}$ denote the LASSO estimate and AMP estimates, respectively. Then there is a constant B such that $\underline{\text{for all}}$ $t \geq 0$ we almost surely have:

$$\lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{n}\|x(t)\|^2 < B \tag{5a}$$

$$\lim_{N\to\infty}\frac{1}{n}\|\hat{x}(\lambda, N)\|^2 < B \tag{5b}$$

Theorem [3]: Let $A(N)$ be the measurement matrix of the LASSO problem whose entries are $\underline{\text{iid zero-mean Gaussian variables with variance } 1/n}$. If $\hat{x}(\lambda, N)$ is the LASSO estimate of the optimal solution x_0 the we *almost surely* have: [1]

$$\lim_{N \to \infty} \frac{1}{N} \|\hat{x} - x_0\|^2 = E\left[(\eta(X_0 + \tau_* Z, \theta) - X_0)^2\right] = \delta(\tau_*^2 - \sigma^2) \tag{6}$$

where τ_* is the variance of the final solution $\hat{x}$.

Although the theorems are proved for *random Gaussian matrices* in the *asymptotic* scenario, simulation results show that they are also applicable to other ensembles of matrices as well as N in the order of a few hundred.

- The signal vector x_0 was generated from the alphabet $\{-1, 0, 1\}$ with $P(x_{0,i} = +1) = P(x_{0,i} = -1 = 0.064$.
- The entries of the noise vector was w was generated by iid Gaussian variables $N(0, 0.2)$.
- The aspect ratio of the measurement matrix, δ, was fixed to $\delta = 0.64$.

To verify the correctness of the algorithm, the authors have used the packages which solve convex programming methods to find $\hat{x}$ [9], [10].

Simulations show that the results for all the scenarios above are *indistinguishable* from those of random Gaussian matrices. This fact strengthen the conjecture on the correctness of the suggested approach even for non-Gaussian matrices.

4 Future Works

For near future, I am going to continue my search to find a proper algorithm for finding a sparse vector that is orthogonal to a set of given data vectors. For that, I will take a look at the Approximate Message Passing (AMP) method [6] and try to extend the method introduced in [3] to the case of non-Gaussian measurement matrices.

References

[1] M. R. Osborne, B. Presnell, B. A. Turlach, *On the LASSO and its Dual*, Journal of Computational and Graphical Statistics, Vol. 9, No. 2, 2000, pp. 319-337.

[2] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Royal Statistical Society, Series B, Vol. 58, 1996, pp. 267-288.

[3] M. Bayati, A. Montanari, *The LASSO risk for gaussian matrices*, ArXiv:1008.2581v1, 2010.

[4] E. Candes, J. K. Romberg, T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics, Vol. 59, 2006, pp. 12071223.

[5] E. Candes, T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics, Vol. 35, 2007, pp. 23132351.

[1] The original version of theorem in the paper is more general. More specifically, it is defined for any *pseudo-Lipschitz function* $\psi : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$. A function $\psi : \mathcal{R}^2 \to \mathcal{R}$ is pseudo-Lipschitz if there exists a constant $L > 0$ such that for all $x_1, x_2 \in \mathcal{R}^2$ we have $|\psi(x_1) - \psi(x_2)| \leq L(1 + \|x_1\|_2 + \|x_2\|_2)\|x_1 - x_2\|_2$.

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

[7] Y. Kabashima, T. Wadayama, T. Tanaka, *A typical reconstruction limit for compressed sensing based on lp-norm minimization*, J. Stat. Mech., 2009, L09003.

[8] D. Guo, D. Baron, S. Shamai, *A single-letter characterization of optimal noisy compressed sensing*, 47th Annual Allerton Conference (Monticello, IL), 2009.

[9] M. Grant, S. Boyd, *CVX: Matlab software for disciplined convex programming*, version 1.21, $http : //cvxr.com/cvx$, May 2010.

[10] G. Andrew, G. Jianfeng, *Scalable training of ℓ_1-regularized log-linear models*, Proc. 24th International Conference on Machine learning, 2007, pp. 3340.