Data Sparse Matrix Computation - Lecture 20

Yao Cheng, Dongping Qi, Tianyi Shi

November 9, 2017

Contents

1 Introduction

Suppose we have some $A \in \mathbb{R}^{m\times n}(m < n)$, there are many solutions to $Ax = b$ and we want to pick one of them. There are many ways to do this: minimize $\|x\|_2$, minimize $\|x\|_1$ or minimize $\|x\|_0$ to get the sparsest solution.

The following is an heuristic idea about sparsity in signal processing due to the Heisenberg's uncertainty principle. Suppose that we have two representations of a signal. One is in time space as $f(x)$ and the other is its Fourier transform in frequency space as $\hat{f}(\omega)$. Assume that $\|f\|_2 = \|\hat{f}\|_2 = 1$, then we have [4].

$$\int_{-\infty}^{\infty} x^2|f(x)|^2dx \int_{-\infty}^{\infty} \omega^2|\hat{f}(\omega)|^2d\omega \geq \frac{1}{2}$$

There is also a discrete version of Uncertainty principle, according to [2]. Suppose that $(x_t)_{t=0}^{N-1}$ is a sequence of length N and let $(\hat{x}_w)_{w=0}^{N-1}$ be its discrete Fourier transform. Then we will have

$$\|(x_t)\|_0 \cdot \|(\hat{x}_w)\|_0 \geq N$$

which in some sense shows that the two reprsentations cannot be sparse both.

2 Theorems on Sparsity

2.1 Example: $A = [\Phi\ \Psi]$

Consider $A = [\Phi\ \Psi]$ where Φ, Ψ are unitary. Given b, $\exists\alpha, \beta$ s.t. $b = \Phi\alpha, b = \Psi\beta$.

Definition 1 *For arbitrary Φ, Ψ with columns ϕ_i, ψ_i, the mutual coherence of Φ, Ψ is*

$$\mu(A) = \max_{1 \le i,j \le m} \|\phi_i^* \psi_j\| (A = [\Phi \ \Psi]) = \textit{max absolute entry of } \phi^T \psi$$

The following is a theorem about the relation of sparsity and coherence. The proof basically follows [3] and readers can find more details there.

Theorem 2 *For arbitrary unitary Φ, Ψ with $\mu(A)$ and arbitrary $b \neq 0 \in \mathbb{R}^m$, if we have some vector α, β s.t. $b = \Phi\alpha, b = \Psi\beta$, then*

$$\|\alpha\|_0 + \|\beta\|_0 \ge \frac{2}{\mu(A)}$$

Proof. Without loss of generality, we can assume that $\|b\|_2 = 1$. Since $b = \Phi\alpha, b = \Psi\beta$ and $b^T b = 1$ we have

$$1 = b^T b = \alpha \Phi^T \Psi \beta = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j \phi_i^T \psi_j \le \mu(A) \sum_{i=1}^{n} \sum_{j=1}^{n} |\alpha_i||\beta_j|$$

which leads to

$$1 \le \mu(A) \cdot \|\alpha\|_1 \|\beta\|_1$$

Consider the following maximizing problem

$$\begin{aligned} \max \quad & \|\alpha\|_1 \\ \text{s.t.} \quad & \|\alpha\|_2^2 = 1 \\ & \|\alpha\|_0 = N \end{aligned}$$

We can assume that all the non-zero entries of α are its first N ones and introduce Lagrange multipliers as follows

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i + \lambda \Big(1 - \sum_{i=1}^{N} \alpha_i^2\Big)$$

Let the derivative equals to zero we have

$$1 - 2\lambda\alpha_i = 0, \quad i = 1, \cdots, N$$

which leads to the solution

$$\alpha_i = \frac{1}{2\lambda}, \quad i = 1, \cdots, N, \quad \lambda = \frac{\sqrt{N}}{2}$$

Therefore the optimal solution is

$$\alpha_i = \frac{1}{\sqrt{N}}, \quad i = 1, \cdots, N, \quad \alpha_i = 0, \quad i \ge N + 1$$

and the optimal value is $\sqrt{\|\alpha\|_0}$ By plugging into the former result and using basic inequality we have

$$\frac{1}{\mu(A)} \le \cdot \|\alpha\|_1 \|\beta\|_1 \le \sqrt{\|\alpha\|_0} \sqrt{\|\beta\|_0} \le \frac{\sqrt{\|\alpha\|_0} + \sqrt{\|\beta\|_0}}{2}$$

which is exactly the claim. ■

If there exists two different solutions $x_1 \neq x_2$ to the problem

$$[\Phi\ \Psi]x = b$$

we know that $e = x_1 - x_2$ is in the null space of A. Let

$$e = \begin{bmatrix} e_\Phi \\ e_\Psi \end{bmatrix}$$

where $e_\Phi, e_\Psi \in \mathbb{R}^m$. Then we can obtain $\Phi e_\Phi = -\Psi e_\Psi = y \neq 0$, since both Φ and Ψ are supposed to be unitary and if $y = 0$, then we have $e_\Phi = e_\Psi = 0$, which means that $x_1 = x_2$.

Then we can apply Theorem 2 to A and y and obtain

$$\|e\|_0 = \|e_\Phi\|_0 + \|e_\Psi\|_0 \geq \frac{2}{\mu(A)}$$

Therefore we can estimate the sparsity on x_1 and x_2 as

$$\|x_1\|_0 + \|x_2\|_0 \geq \|e\|_0 \geq \frac{2}{\mu(A)}$$

which give rise to the following theorem.

Theorem 3 *If a solution to $[\Phi\ \Psi]x = b$ has fewer than $\frac{1}{\mu(A)}$ non-zero entries, then it is the sparsest solution and in fact is unique.*

Proof. Suppose x^* is a solution to $[\Phi\ \Psi]x = b$ with $\|x^*\|_0 < \frac{1}{\mu(A)}$, then we have for any solution x to this problem

$$\|x\|_0 \geq \frac{2}{\mu(A)} - \|x^*\|_0 > \frac{1}{\mu(A)} > \|x^*\|_0$$

which means that x^* is the sparsest solution. ■

2.2 General Matrix A

Key idea: The "ideal" property of A is for the following to be large:

The smallest number of linearly dependent columns

Consider the following example. If all the columns of A are generated by two linearly independent vectors, and b can also be expressed as linear combinations of these two vectors (guarantees solutions exist), i.e.

$$[\phi\ \psi\ (\lambda\phi + \mu\psi)\ \cdots]x = [a\phi + b\psi]$$

then we can choose any two coordinates of x to be non-zero without influence the sparsity of x. This provides several sparsest solutions, which is what we do not want.

Definition 4 *The mutual coherence of a general $m \times n$ matrix A is*

$$\mu(A) = \max_{1 \leq i \neq j \leq n} \frac{|a_i^* a_j|}{\|a_i\|_2 \|a_j\|_2}$$

Theorem 5 *For $Ax = b$ with A of size $m \times n$ and A has unit 2-norm columns, if a solution x exists with $||x||_0 \leq \frac{1}{2}(1 + \frac{1}{\mu(A)})$, it is unique and solves both of the following problems:*

$$\begin{aligned} \min \quad & ||x||_0 \\ s.t. \quad & Ax = b \end{aligned}$$

and

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & Ax = b \end{aligned}$$

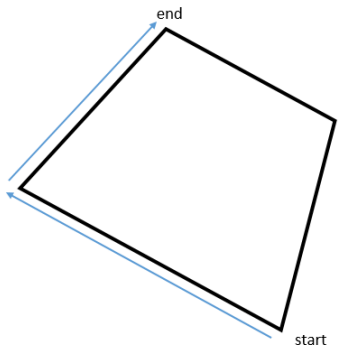
Proof. Here we present a sketch of the proof. For the detailed and complete proof, refer to [3].
From the previous theorem, we know x is the unique solution to the former problem, and we let $S = supp(x)$. Now suppose y is the optimizer for the latter problem and is different from x, and let $e = y - x$. Since x and y are both solutions to the problem $Ax = b$ we know $Ae = 0$. We also know from the optimality that $||e||_1 \leq 2||e_S||_1$. Besides, we can also derive $A^T Ae = 0$ and from this we know $\forall j, |e_j| \leq (1 + \mu(A))^{-1}\mu(A)||e||_1$. These two inequalities we have contradict our assumption. ■

From this theorem, we know that to find the desired solution x, we can solve the problem with 1-norm. There are several ways to solve this, and one way is to write it as a linear program (LP). Normally, to solve such a linear program, we would like to have $x \geq 0$ but we do not have such a constraint on x so we consider making two non-negative vectors the same dimension as x (denoting the positive and -negative parts) and rewrite the problem:
Let $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $u \geq 0$, $v \geq 0$ and

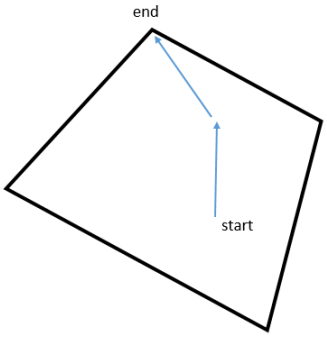
$$w = \begin{bmatrix} u \\ v \end{bmatrix} \tag{1}$$

$$\begin{aligned} \min \quad & 1^T w \\ \text{s.t.} \quad & [A \ -A]w = b \\ & w \geq 0 \end{aligned}$$

Two ways are commonly used to solve an LP problem, simplex method and interior point method. Systems of equations or inequalities of an LP problem often define a n-dimensional polytope and these two methods use two different ideas of finding solutions on the polytope[1].
Simplex method starts at a feasible vertex and move along the edges of the polytope to another vertex until it reaches the optimum solution. We can visualize it in a 2D graph:



Interior point method, comparatively, is a polynomial-time efficient method to solve an LP problem. It usually achieves optimization by going through the middle of the polytope rather than around the surface like the simplex method. We can also visualize it in a 2D graph:



3 Denoising and LASSO

Generally, in practical problems, we would not have a simple and pretty b. Instead, we will often encounter $y = b + \sigma z$ where σ is the noise level and z is a random number from Gaussian distribution with mean 0 and variance 1. In this case, solving $Ax = y$ for x might be a bad practice.
Our Basis Pursuit then becomes a Basis Pursuit Denoising (BPDN), an optimization problem: $\min_x \frac{1}{2}||Ax - y||_2^2 + \lambda||x||_1$, where λ is a parameter that controls the trade-off between sparsity and reconstruction fidelity.
We can tell that this optimization is sort of equivalently to least absolute shrinkage and selection operator (LASSO),a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces, which solves:

$$\begin{array}{ll} \min & \frac{1}{2}|Ax - b||_2^2 \\ \text{s.t.} & ||x||_1 \leq t \end{array}$$

We will cover more about LASSO and BPDN in future lectures.

References

[1] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[2] David L Donoho and Philip B Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[3] Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[4] Elias M Stein and Rami Shakarchi. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.