Matrix Computations **Fall 2020**

Lecture 4:

Lecturer: Prof.Yue Qiu & Prof Ziping Zhao *Scribe: Bing Jiang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Least Squares

problem: given y$\in$R^m,A$\in$R$^{m\times n}$,solve

$$\mathbf{x_{LS}} = \min_{\mathbf{x}\in\mathbb{R}^n} ||\mathbf{y} - \mathbf{Ax}||_2^2 \tag{1}$$

- called (linear) least squares (LS)
- find an x whose residual r=y-Ax is the smallest in the Euclidean sense

solution: suppose that A has full-column rank ($m \geq n$). The solution to (LS) is unique and is given by

$$\mathbf{X_{LS}} = (\mathbf{A^T A})^{-1}\mathbf{A^T y} \tag{2}$$

- complexity O($mn^2 + n^3$)
- LS solutions to an overdetermined system of equations y=Ax ($m > n$)
- if A is semi-orthogonal, the solution is simplified to $\mathrm{X}_{LS} = \mathbf{A^T y}$
- if A is square, the solution is simplified to $\mathrm{X}_{LS} = \mathbf{A^{-1} y}$
- unless specified, in this lecture we will assume A to have full column rank without further mentioning

2 LS Solution

Theorem 1 A vector $\mathbf{X_{LS}}$ is an optimal solution to the LS problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} ||\mathbf{y} - \mathbf{Ax}||_2^2 \tag{3}$$

if and only if it satisfies

$$\mathbf{A^T A X_{LS}} = \mathbf{A^T y} \tag{4}$$

- the optimality condition in (4) is true for any A, not just full-column rank A

- suppose that A has full column rank
  - (4) is a symmetric PD linear system
  - the Gram metrix $A^T Aisnon-singular$
  - the solution to (4) is uniquely given by $\mathbf{X_{LS}} = (\mathbf{A^T A})^{-1}\mathbf{A^T y}$
- (4) is called the normal equations
- the same result holds for the complex case, viz., $\mathbf{A^H A X_{LS}} = \mathbf{A^H y}$

3 Gradient Decent For LS

- consider a general unconstrained optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \tag{5}$$

where f is continuously differentiable

- Gradient Decent: given a starting point $x^{(0)}$, do

$$x^{(k)} = x^{(k-1)} - \mu\nabla f(x^{(k-1)}), k = 1, 2, \ldots \tag{6}$$

where $\mu > 0$ is a step size

- take an optimization course to get more details! It is known that
  - for convex f and under some appropriate choice of μ, gradient decent converges to an optimal solution
  - for non-convex f and under some appropriate choice of μ, gradient decent converges to a stationary point
- gradient decent for LS:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - 2\mu(\mathbf{A^T A x^{(k-1)}} - \mathbf{A^T y}), k = 0, 1, \ldots \tag{7}$$

- complexity for dense A
  - computing $\mathbf{A^T A}$ and $\mathbf{A^T y}$: O(mn^2) and O(mn), resp. (same as before)
    - * $\mathbf{A^T A}$ and $\mathbf{A^T y}$ are cached for subsequent use in gradient decent
  - complexity of each iteration: O(n^2)
- complexity for sparse A
  - computing $A^T y : \mathcal{O}(nnz(\mathbf{A}))$
  - complexity of each iteration: O(n+nnz(A))
    - * $\mathbf{A^T A}$ is not necessarily sparse, so we do $\mathbf{A}\mathbf{x}^{(k-1)}$ and then $\mathbf{A^T}(\mathbf{A}\mathbf{x}^{(k-1)})$

4 Regularized LS

- Intuition: replace $\mathbf{X_{LS}} = \mathbf{A}^{\dagger}\mathbf{y} = (\mathbf{A^T A})^{-1}\mathbf{A^T y}$ by

$$\mathbf{X_{RLS}} = (\mathbf{A^T A} + \lambda \mathbf{I})^{-1}\mathbf{A^T y}, \tag{8}$$

  for some $\lambda > 0$, where the term λI is added to improve the system conditioning, thereby attempting to reduce noise sensitivity

- how may we make sense out of such a modification?

- L2 regularized LS: find an x that solves

$$\min_{\mathbf{x} \in \mathbb{R}^n} ||\mathbf{y} - \mathbf{Ax}||_2^2 + \lambda ||\mathbf{x}||_2^2 \tag{9}$$

  for some pre-determined $\lambda > 0$.

  - the solution is uniquely given by $\mathbf{X_{RLS}} = (\mathbf{A^T A} + \lambda \mathbf{I})^{-1}\mathbf{A^T y}$

  - the formulation says that we try to minimize both $||\mathbf{y} - \mathbf{Ax}||_2^2$ and $||\mathbf{x}||_2^2$, and λ controls which one should be more emphasized in the minimization.

- L1 regularized LS: given $\lambda > 0$, solve

$$\min_{\mathbf{x} \in \mathbb{R}^n} ||\mathbf{y} - \mathbf{Ax}||_2^2 + \lambda ||\mathbf{x}||_1 \tag{10}$$

- now consider applying Majorization-Minimization to the L2-L1 minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2}||\mathbf{y} - \mathbf{Ax}||_2^2 + \lambda ||\mathbf{x}||_1 \tag{11}$$

- The Majorization-Minimization method for solving problem updates $\mathbf{x}$ as

$$\mathbf{x^{(k+1)}} = \mathbf{soft}(\frac{1}{c}\mathbf{A^T}(\mathbf{y} - \mathbf{Ax^{(k)}}) + \mathbf{x^{(k)}}, \lambda/c),$$

  where **soft** is called the soft-thresholding operator and is defined as follows: if $\mathbf{z} = \mathbf{soft}(\mathbf{x}, \sigma)$, then $z_i = sign(x_i)\max\{|x_i| - \sigma, 0\}$.