Numerical Optimization

Unit 7: Least Square Problems

Che-Rung Lee

Department of Computer Science
National Tsing Hua University

October 28, 2022

Linear least squares

- Given samplings $\vec{a}_1, \vec{a}_2, \dots \vec{a}_m \in \mathbb{R}^n$ for observations $b_1, b_2, \dots b_m \in \mathbb{R}^1$, the linear least square method wants to find $\vec{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ s.t. $F(\vec{x}) = \sum_{i=1}^{m} (\vec{a}_i^T \vec{x} - b_i)^2$ is minimized.
- Let $A = \begin{pmatrix} \vec{a}_1^T \\ \vec{a}_2^T \\ \vdots \\ \vec{a}_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}$, $\vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$.
- Let $F(\vec{x}) = \|A\vec{x} - \vec{b}\|^2 = (A\vec{x} - \vec{b})^T (A\vec{x} - \vec{b})$. The problem can be written as

$$\min_{\vec{x}} F(\vec{x})$$

Normal equation

- The optimal condition of linear least squares is $\nabla F = 0$,

$$\nabla F(\vec{x}) = 2A^T(A\vec{x} - \vec{b}) = 0.$$

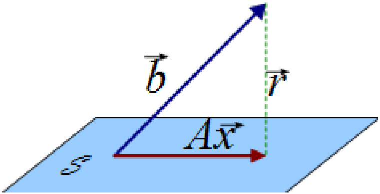
- The equation

$$A^T A\vec{x} = A^T \vec{b}, \tag{1}$$

  is called the *normal equation*.
- Matrix $A^T A$ is symmetric positive semi-definite. (why?)
- If $A^T A$ is SPD, we can solve (1) by the Cholesky decomposition.
- If $A^T A$ is ill-conditioned, solving (1) directly is not numerically stable.
- How to solve (1) if $A^T A$ is singular or ill-conditioned?
- A best way to solve the normal equation is by the QR method.

Geometrical interpretation of linear least square

- The problem $\min_{\vec{x}} \|A\vec{x} - \vec{b}\|^2$ is to find a linear combination of A's column vectors which is closet to $\vec{b}$.
- Let $\mathcal{S}$ be the subspace spanned by A's column vectors.
- If $\vec{b}$ is in $\mathcal{S}$, then there exists $\vec{x} \in \mathcal{S}$ s.t. $A\vec{x} = \vec{b}$.
- If $\vec{b}$ is not in $\mathcal{S}$, then $A\vec{x}$ is $\vec{b}$'s projection on $\mathcal{S}$. (why?)



- Moreover, $\|\vec{r}\| = \min_{\vec{x}} \|A\vec{x} - \vec{b}\|$.

Geometrical interpretation

- The vector $\vec{r}$ is orthogonal to all the column vectors in $A = [\vec{a}_1 \ \vec{a}_2 \ \dots \ \vec{a}_n]$, which means

$$A^T\vec{r} = \begin{bmatrix} \vec{a}_1^T\vec{r} \\ \vec{a}_2^T\vec{r} \\ \vdots \\ \vec{a}_n^T\vec{r} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow A^T(A\vec{x} - \vec{b}) = 0 \Rightarrow A^TA\vec{x} = A^T\vec{b}.$$

  which is the normal equation again.
- The column vectors of Q_1 form an orthogonal basis of $\mathcal{S}$. The vector that $\vec{b}$ projected to $\mathcal{S}$ is $Q_1Q_1^T\vec{b}$, where $Q_1^T\vec{b}$ is the coordinates of the projected vector in the Q_1 coordinate system.

QR method

The QR method for linear least square problem for $m \geq n$.

Algorithm 1: QR method

1. Compute A's QR decomposition:
$$AP = Q_1 \left[\begin{array}{cc} R_{k\times k} & T_{k\times(n-k)} \end{array} \right], \tag{2}$$
where Q_1 is an $m \times k$ matrix, $Q_1^T Q_1 = I$, R is full ranked upper triangular, and P is an $n \times n$ permutation matrix.
2. The inverse of P is P^T, so
$$A = Q_1 \left[\begin{array}{cc} R_{k\times k} & T_{k\times(n-k)} \end{array} \right] P^T. \tag{3}$$
3. The optimal solution $\vec{x}^* = P \left[\begin{array}{c} R^{-1} Q_1^T \vec{b} \\ \vec{0}_{n-k} \end{array} \right]$.

Matrix rank and orthogonal matrix

- Rank of a matrix: the number of linearly independent rows or columns of a matrix.
- Let Q_2 be the orthogonal complement of Q_1, and $Q = [Q_1 \;\; Q_2]$. The matrix Q is an $m \times m$ orthogonal matrix, which means $Q^T Q = I$, and $Q^{-1} = Q^T$. It implies $Q_1^T Q_1 = I_k$, $Q_2^T Q_2 = I_{m-k}$, $Q_1^T Q_2 = 0_{k\times(m-k)}$, and $Q_2^T Q_1 = 0_{(m-k)\times k}$.
- From the normal equation $A^T A\vec{x} = A^T\vec{b}$ and (3), we have

$$P \begin{bmatrix} R^T \\ T^T \end{bmatrix} Q_1^T Q_1 \begin{bmatrix} R & T \end{bmatrix} P^T \vec{x} = P \begin{bmatrix} R^T \\ T^T \end{bmatrix} Q_1^T \vec{b}$$

Let $P^T \vec{x}^T = [\vec{x}_1^T \;\; \vec{x}_2^T]$ where $\vec{x}_1 \in \mathbb{R}^k$ and $\vec{x}_1 \in \mathbb{R}^{n-k}$. It becomes

$$\begin{bmatrix} R^T R & R^T T \\ T^T R & T^T T \end{bmatrix} \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \end{bmatrix} = \begin{bmatrix} R^T Q_1^T \vec{b} \\ T^T Q_1^T \vec{b} \end{bmatrix} \tag{4}$$

Algebraic derivation

Let $Q = [Q_1 \; Q_2]$ be a full orthogonal matrix, where Q_1 and Q_2 are defined as in the QR method.

$$\begin{aligned} \|\vec{r}\|^2 = \|A\vec{x} - \vec{b}\|^2 &= \|Q^T(A\vec{x} - \vec{b})\|^2 \\ &= \|Q_1^T(A\vec{x} - \vec{b})\|^2 + \|Q_2^T(A\vec{x} - \vec{b})\|^2 \\ &= \|Q_1^T A\vec{x} - Q_1^T\vec{b}\|^2 + \|Q_2^T\vec{b}\|^2. \end{aligned}$$

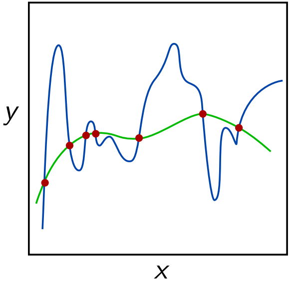
- We can control $\vec{x}$ and make the first term 0, but we cannot do anything about the second term.
- The first equation from (4) shows $R\vec{x}_1 + T\vec{x}_2 = Q_1^T\vec{b}$. One of the solution is to set $\vec{x}_1 = R^{-1}Q_1^T\vec{b}$ and $\vec{x}_2 = \vec{0}$, which gives us a solution

$$\vec{x}^* = P\begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \end{bmatrix} = P\begin{bmatrix} R^{-1}Q_1^T\vec{b} \\ \vec{0}_{n-k} \end{bmatrix}.$$

- The solution $\vec{x}^*$ is not unique, but $A\vec{x}^* = Q_1 Q_1^T\vec{b}$ is unique.

Overfitting problem and regularization

- The solution of a least square problem may generate a model that fits the given data well, but loses the generality. Such problem is called overfitting.
- Regularization is the process of adding information to prevent overfitting.



Two commonly used regularizations:

1. ℓ_2 regularization (ridge regression, Tikhonov regularization):

$$\min_{\vec{x}} \|A\vec{x} - \vec{y}\|^2 + \lambda \|\vec{x}\|_2^2.$$

2. ℓ_1 regularization (LASSO: Least Absolute Shrinkage and Selection Operator):

$$\min_{\vec{x}} \|A\vec{x} - \vec{y}\|^2 + \lambda \|\vec{x}\|_1$$

Ridge regression

- The objective of the ridge regression is

$$
\begin{aligned}
J(\vec{x}) =& \|A\vec{x} - \vec{y}\|^2 + \lambda \|\vec{x}\|_2^2 \\
=& (A\vec{x} - \vec{y})^T (A\vec{x} - \vec{y}) + \lambda \vec{x}^T \vec{x} \\
=& \vec{x}^T (A^T A + \lambda I)\vec{x} - 2\vec{y}^T A\vec{x} + \vec{y}^T \vec{y}
\end{aligned}
$$

- Since $A^T A$ is symmetric positive semi-definite and $\lambda > 0$, J is a convex function, which has a unique minimum point at $\nabla J = 0$.

$$\nabla J = 2(A^T A + \lambda I)\vec{x} - 2A^T \vec{y} = 0$$

$$\vec{x}^* = (A^T A + \lambda I)^{-1} A^T \vec{y}.$$

- This is just like the modified Newton method using shift.

LASSO

- The objective of LASSO is

$$J(\vec{x}) = \|A\vec{x} - \vec{y}\|^2 + \lambda \|\vec{x}\|_1,$$

where $\|\vec{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$.

- Since the absolute function is non-differentiable, it cannot be solved directly. However, we can reformulate it as a constrained optimization problem.

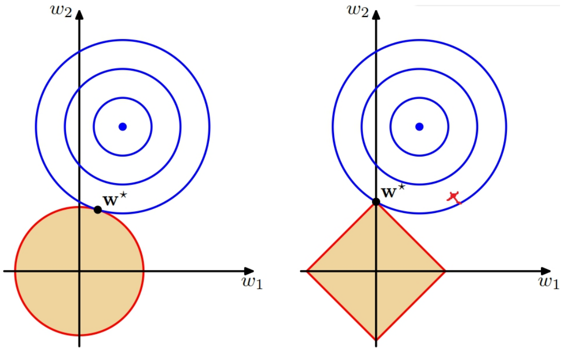
$$\min_{\vec{x},s} \|A\vec{x} - \vec{y}\|^2 + \lambda \sum_{i=1}^{n} s_i$$

$$\begin{aligned} \text{s.t.} \qquad x_i &\leq s_i \quad \text{for } i = 1, 2, \ldots, n \\ -x_i &\leq s_i \quad \text{for } i = 1, 2, \ldots, n \\ s_i &\geq 0 \end{aligned}$$

- We will learn how to solve this constrained optimization problem later.

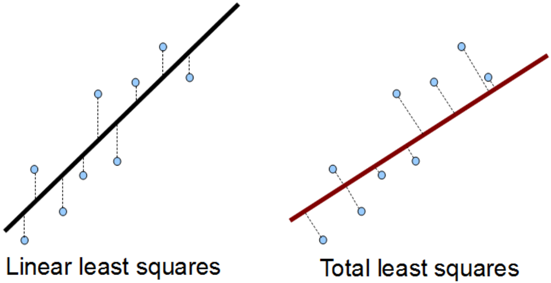
Comparison of those two regularization methods

- ℓ_2 regularization pushes all the element in $\vec{x}$ toward 0, but not exactly zero.
- ℓ_1 regularization makes the solution $\vec{x}$ sparse (many zeros) because of the natural of 1-norm.



Errors in observations and sampling points

- In the linear least square problems, we assume that the samplings, $\vec{a}_1, \vec{a}_2, \ldots \vec{a}_m$, have no bias and the only error comes from the observations $b_1, b_2, \ldots b_m$. What if the error is contributed by sampling and observations?
- The two dimensional problem: Suppose the sampling points are at $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$, and the observations are $y_1, y_2, \ldots y_m$.



Total least square problem for 2D

- Total least square: find a line $ax + by + c = 0$ such that the summation of the distance of all points $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ to this line is minimized.
- We need to find a, b, c. To make solution unique, we let $\sqrt{a^2 + b^2} = 1$.
- How to compute the distance from a point to a line?
  - The distance of a point (x_i, y_i) to the line $ax + by + c = 0$ is $|ax_i + by_i + c|$. (why?)
- Therefore, the total least squares can be formulated as

$$\min_{a,b,c} \sum_{i=1}^{m} (ax_i + by_i + c)^2,$$

where $a^2 + b^2 = 1$.

How to solve?

- Let $F(a,b,c) = \sum_{i=1}^{m}(ax_i + by_i + c)^2$. You may want to solve this problem by solving $\nabla F = 0$.

$$\nabla F = \begin{pmatrix} \partial F/\partial a \\ \partial F/\partial b \\ \partial F/\partial c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} 2x_i(ax_i + by_i + c) \\ \sum_{i=1}^{m} 2y_i(ax_i + by_i + c) \\ \sum_{i=1}^{m} 2(ax_i + by_i + c) \end{pmatrix}$$

- But this is not correct, since it has a constraint $a^2 + b^2 = 1$.
- Fortunately, the condition $\partial F/\partial c = 0$ is still held.
  - Let $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$. $(\bar{a}, \bar{b})$ is the centroid of data.
  - $(\bar{a}, \bar{b})$ must be on the solution line. (why?)
  - If we shift all the points to make $(\bar{a}, \bar{b}) = (0, 0)$, then the line equation becomes $ax + by = 0$.

The two dimensional problem example

- Let $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{y}_i = y_i - \bar{y}$. The problem becomes

$$\min_{a,b} \sum_{i=1}^{m} (a\tilde{x}_i + b\tilde{y}_i)^2 \text{ s.t. } a^2 + b^2 = 1$$

- Let matrix $A = \begin{pmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ x_2 - \bar{x} & y_2 - \bar{y} \\ \vdots & \vdots \\ x_m - \bar{x} & y_m - \bar{y} \end{pmatrix}$, and $\vec{x} = \begin{pmatrix} a \\ b \end{pmatrix}$.
- The problem can be expressed as

$$\min_{\vec{x}, \|\vec{x}\|=1} \vec{x}^T A^T A \vec{x}.$$

- In statistics, the matrix $A^T A$ is the covariance matrix of data $\{(x_i, y_i)\}_{i=1 \ldots m}$.

How to solve that?

- For the constrained optimization problem, the optimality condition is $\nabla f(\vec{x}) = \lambda \nabla c(\vec{x})$, where $c(\vec{x}) = 0$ is the constraint and λ is some scalar.
- Therefore, the optimal solution $\vec{x}^*$ must satisfy

$$A^T A \vec{x}^* = \lambda \vec{x}^*.$$

- The above equation says the solution is an eigenvector of $A^T A$, but which one?
- A faster way is using the singular value decomposition (SVD)

Singular value decomposition (SVD)

Theorem (Existence of SVD)

If A is a real $m \times n$ matrix, there exist orthogonal matrix $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that

$$U^T AV = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$$

where $p = \min(m, n)$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

Theorem (min-max of SVD)

If A is a real $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$, $p = \min(m, n)$, then for $k = 1, 2, \ldots, p$,

$$\sigma_k = \max_{\dim(S)=p-k+1} \min_{\vec{x}\in S} \frac{\|A\vec{x}\|}{\|\vec{x}\|}.$$

General form of least squares

- Let $f(\vec{x}) = \dfrac{1}{2}\sum_{j=1}^{m} r_j^2(\vec{x})$, in which $r_j(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$ is a smooth function, and $m \geq n$.
- Each $r_j = \phi(\vec{x}_j) - y_j$ is called a "residual", where function $\phi(\vec{x})$ is called the model function and y_j is an observation obtained at the sampling point $\vec{x}_j$.
- The least square problem is to solve

$$\min_{\vec{x}} f(\vec{x})$$

- If ϕ is nonlinear, the problem is called nonlinear least squares.

Vector function form

- Define a vector function $\vec{r}(\vec{x}) = \mathbb{R}^n \to \mathbb{R}^m$.

$$\vec{r}(\vec{x}) = \begin{pmatrix} r_1(\vec{x}) \\ r_2(\vec{x}) \\ \vdots \\ r_m(\vec{x}) \end{pmatrix}.$$

- The Jacobian $J(\vec{x})$ of $\vec{r}(\vec{x})$ is an $m \times n$ matrix

$$J(\vec{x}) = \begin{bmatrix} \nabla \vec{r}_1^T(\vec{x}) \\ \nabla \vec{r}_2^T(\vec{x}) \\ \vdots \\ \nabla \vec{r}_m^T(\vec{x}) \end{bmatrix} = \begin{bmatrix} \partial r_1/\partial x_1 & \partial r_1/\partial x_2 & \dots & \partial r_1/\partial x_n \\ \partial r_2/\partial x_1 & \partial r_2/\partial x_2 & \dots & \partial r_2/\partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial r_m/\partial x_1 & \partial r_m/\partial x_2 & \dots & \partial r_m/\partial x_n \end{bmatrix}$$

Nonlinear least square problems

- From the above definition, $f(\vec{x}) = \frac{1}{2}\vec{r}^T\vec{r}$.
- The gradient of $f(\vec{x})$ is

$$\nabla f(\vec{x}) = \sum_{j=1}^{m} r_j(\vec{x})\nabla r_j(\vec{x}) = J(\vec{x})^T\vec{r}(\vec{x})$$

- The Hessian of $f(\vec{x})$ is

$$\begin{aligned}\nabla^2 f(\vec{x}) &= \sum_{j=1}^{m} \nabla r_j(\vec{x})\nabla r_j(\vec{x})^T + \sum_{j=1}^{m} r_j(\vec{x})\nabla^2 r_j(\vec{x}) \\ &= J(\vec{x})^T J(\vec{x}) + \sum_{j=1}^{m} r_j(\vec{x})\nabla^2 r_j(\vec{x})\end{aligned}$$

- If ϕ is linear, $J(\vec{x}) = A$, $\vec{r}(\vec{x}) = A\vec{x} - \vec{b}$, and $\nabla^2 f(\vec{x}) = A^T A$.

Solve nonlinear least squares

We will present two algorithms to solve nonlinear least squares

- The Gauss–Newton method
- The Levenberg-Marquardt method.

The Gauss–Newton method

- Assume the residuals $r_j(\vec{x})$ are small, and we can approximate $\nabla^2 f(\vec{x}) \approx J^T J$.
- Use Newton's method to compute the search direction $\vec{p} = -H^{-1}\vec{g}$.
- It goes back to the linear least square method normal equation

$$(J^T J)\vec{p} = -J^T \vec{r}.$$

The Levenberg-Marquardt method

- It is under the trust-region framework. (See note 3.)
- The model is quadratic

$$m_k(\vec{p}) = \frac{1}{2}\|\vec{r}_k\|^2 + \vec{p}^T J_k^T \vec{r}_k + \frac{1}{2}\vec{p}^T J_k^T J_k \vec{p}$$

$$\min_{\vec{p}} \frac{1}{2}\|J_k\vec{p} + \vec{r}_k\|^2 \text{ s.t. } \|\vec{p}\| \le \Delta_k$$

- We will learn how to solve this kind of constrained problem in the rest of semester. Here are some clues.
  - If $\vec{z} = -(J_k^T J_k)^{-1}(J_k^T \vec{r}_k)$ and $\|\vec{z}\| < \Delta_k$, $\vec{p} = \vec{z}$.
  - Otherwise , there exists an λ s.t. $(J_k^T J_k + \lambda I)\vec{p} = -J_k^T \vec{r}_k$ and $\|\vec{p}\| = \Delta_k$. The remaining problem is how to find λ_k.

Other variations

Weighted least square problem

For a diagonal matrix W, the weighted least squares is to solve

$$\min_{\vec{x}} \|W(A\vec{x} - \vec{b})\|^2.$$

Lorentzian functions

- The square function is sensitive to outliers. Use Lorentzian function

$$L(\vec{r}) = \log(1 + \vec{r}^T \vec{r}/\sigma).$$

- The problem becomes $\min_{\vec{x}} L(A\vec{x} - \vec{b})$.

Constrained least squares

$$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|^2 \text{ s.t. } \|B\vec{x} + \vec{d}\| \leq \alpha.$$