Least Squares Optimization

Objective function: $f(x) = \frac{1}{2}\sum_j r_j^2(x)$

$\downarrow$ residual : smooth

Ch 10 NW.

eg. Radioactive decay:

$$y(x,t) = x_1 + x_2 \exp(-x_3 t).$$

$t \to$ time.

unknowns : $x = \{x_1, x_2, x_3\}$.

Measure $\{t_j\}_{j=1}^{m} \to \{y_j\}_{j=1}^{m}$.

We want the "best" model to explain the data. ① $|y_j - y(x,t_j)|$

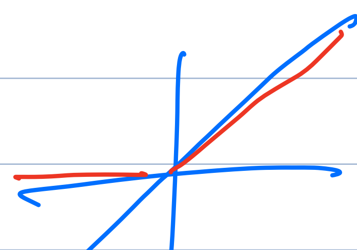
$\rightarrow$ ② $(y_j - y(x,t_j))^2$

motivates: $$\hat{x} = \underset{x}{\arg\min} \sum_{j=1}^{m} (y_j - y(x,t_j))^2$$

$\Rightarrow$ Least squares.

If this is affine w.r.t x then linear, else nonlinear.

$$y(x,t) = x_1 + x_2 t + x_3 t^2$$
linear problem.



Formalism

$$f(x) = \frac{1}{2} \sum_{j=1}^{m} r_j(x)^2$$

$x \in \mathbb{R}^{n}$

m → no of data points

n → no of vars

$\Rightarrow$ $m \geq n$ → realistic case

Aside $m < n$? → ∞ solns

min norm soln

pseudo inv soln.

a priori knowledge / domain knowledge

"Compressive Sensing"

Calculus

$$r(x) = [r_1(x), r_2(x), \ldots, r_m(x)]^T$$

$$f(x) = \frac{1}{2} \|r(x)\|^2 \; : \mathbb{R}^n \to \mathbb{R}$$

i/p o/p

$$r(x) : \mathbb{R}^n \to \mathbb{R}^m$$

Jacobian

$m \times n$

$$J_{ji} \equiv \left(\frac{\partial r_j}{\partial x_i} \right) = \begin{bmatrix} \nabla r_1(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}$$

$\frac{\partial r_1}{\partial x_1} \; \frac{\partial r_1}{\partial x_2} \; \ldots \; \frac{\partial r_1}{\partial x_n}$

$m \times n$

Gradient of f

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \end{bmatrix}$$

$$f = \frac{1}{2}\sum_j r_j(x)$$

$$\frac{\partial f}{\partial x_j} = \sum_j r_j(x)\frac{\partial r_j(x)}{\partial x_i} \qquad \begin{bmatrix} \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

$$\nabla f(x) = J(x)^T r(x).$$

Hessian: $$H_{ik} = \frac{\partial^2 f}{\partial x_i \partial x_k}$$

$$= \sum_j \frac{\partial r_j(x)}{\partial x_k}\frac{\partial r_j(x)}{\partial x_i} + \sum_j r_j \frac{\partial^2 r_j(x)}{\partial x_i \partial x_k}$$

$$\nabla^2 f(x) = J^T(x) J(x) + \sum_j r_j \nabla^2 r_j(x)$$

① r_j is small
↳ close to soln

② $\nabla^2 r_j$ is small
↳ r_j is affine

term usually lower in magnitude compared to the first term.

linear least squares : $\nabla^2 f = J^T J$

↳ Aside: Why squares?

$$y_j - y(x, t_j) \sim \varepsilon_j \quad \text{discrepancy}$$

I get the maximum likelihood estimate (MLE) when I minimize the sum of squares.

$$\sum_j |y_j - y(x, t_j)| \rightarrow \|r\|_1$$

$$\max_j |y_j - y(x, t_j)| \rightarrow \|r\|_\infty$$

Linear least squares

: $r_j(x)$ = Affine function

$$r(x) = \underbrace{A}_{m\times n} x - \underbrace{b}_{m\times 1}$$

$$f(x) = \frac{1}{2} \|Ax - b\|^2 ,$$

$$\nabla f(x) = A^T(Ax - b) = A^T r(x)$$

$$\Rightarrow A = J, \; b = y$$

$$\nabla f(x) = J^T r(x)$$

$$\nabla^2 f(x) = J^T J \rightarrow J \text{ is a const.}$$

f is a convex fn $\Rightarrow$ global minimizer? x^*

$$\nabla f(x^*) = 0 .$$

$$J^T r(x) = 0 = J^T(Jx^* - y) = 0$$

$$\Rightarrow \boxed{J^T J x = J^T y.} \rightarrow \text{Normal eqns.}$$

Solving using SVD

Says: $\underset{m\times n}{J} = \underset{m\times m}{U}\,\underset{m\times n}{S}\,\underset{n\times n}{V^T}$

$UU^T = I$ $\quad$ U, V : orthogonal matrices

S : diagonal, $\sigma_i \geq 0$

$m > n$ $\begin{bmatrix} \ddots & 0 \\ 0 & \end{bmatrix}$, $m < n$ $\begin{bmatrix} \ddots & 0 \\ 0 & \end{bmatrix}$

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$

$k = \min(m, n)$

$$J^T J = (VS^T U^T)(USV^T)$$

$$= V \underbrace{S^T S}_{n\times n} V^T$$

$n \times n$ → diagonal → σ_i^2, $i = 1$ to k

$m > n \rightarrow \sigma_1^2 \rightarrow \sigma_n^2$

$m < n \rightarrow \sigma_1^2 \cdots \rightarrow \sigma_m^2$

$$J^T J x^* = J^T y$$

$$\cancel{V} \underbrace{S^T S}\, V^T x^* = \cancel{V} S^T U^T y \qquad (\text{left} \times V^T)$$

① $m \geq n$
② $\sigma_n \neq 0$ $\Big\}$ $\Rightarrow$ J has full col. rank.

$$V^T x^* = \underbrace{(S^T S)^{-1} S^T}_{n \times m} U^T y$$

$$x^* = V D^{-1} U^T y$$

first n cols of U

square : $n \times n$ $\quad \frac{1}{\sigma_i}$

$$x^* = \sum_{i=1}^{n} \frac{1}{\sigma_i} (u_i^T y) v_i$$

$$\begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_n \end{bmatrix} \begin{bmatrix} u_1 & \text{---} \\ u_2 & \text{---} \\ u_n & \text{---} \end{bmatrix} \begin{bmatrix} \uparrow \\ y \\ \downarrow \end{bmatrix}$$

— x —