9/6

- Previously, we saw some basic LP ideas
- Now we will look at nonlinear optimization

Example: least squares data-fitting

$f : \mathbb{R} \to \mathbb{R}$: some unknown function, TBD
$\quad y \mapsto f(x)$

$y_i = y(x_i)$: observations, $i = 1, \ldots, n$

$f(x) = c_0 + c_1 x + c_2 x^2$: simple quadratic model
$r_i = r(x_i) = y_i - f(x_i)$: residual

Goal: find c_0, c_1, c_2 st $\underbrace{\frac{1}{n}\sum_i r_i^2}_{}$ is minimized

MMSE: minimum mean-square error

or: minimize $\sum_i r_i^2$
subject to <nothing>

This is an unconstrained quadratic program (QP).

How to solve? Take pds, set to zero, solve for (c_0, c_1, c_2)

Note: $r_i = r_i(c_0, c_1, c_2)$

Have: $$\frac{\partial(r_i^2)}{\partial c_j} = 2r_i \frac{\partial r_i}{\partial c_j}, \quad j = 0, 1, 2$$

$$\frac{\partial r_i}{\partial c_0} = \frac{\partial}{\partial c_0}\{y_i - c_0 - c_1 x_i - c_2 x_i^2\} = -1$$

$$\frac{\partial r_i}{\partial c_1} = -x_i, \quad \frac{\partial r_i}{\partial c_2} = -x_i^2$$

So: solve:

$$0 = \frac{\partial(r_i^2)}{\partial c_0} = 2r_i \frac{\partial r_i}{\partial c_j} = -2(y_i - f(x_i))\frac{\partial r_i}{\partial c_j}$$

$$\Rightarrow \begin{cases} 0 = -\sum_{i=1}^{n} (y_i - f(x_i)) \\ 0 = -\sum_{i=1}^{n} (y_i - f(x_i))\, x_i \\ 0 = -\sum_{i=1}^{n} (y_i - f(x_i))\, x_i^2 \end{cases}$$

This is a linear system. Rewrite in matrix notation:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = -\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \end{bmatrix} \begin{bmatrix} y_1 - f(x_1) \\ y_2 - f(x_2) \\ \vdots \\ y_n - f(x_n) \end{bmatrix}$$

Let's define:

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, \quad \hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad c = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

Then: note:

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \begin{bmatrix} c_0 + c_1 x_1 + c_2 x_1^2 \\ \vdots \\ c_0 + c_1 x_n + c_2 x_n^2 \end{bmatrix} = Ac$$

So:

$$0 = -A^T(\hat{y} - Ac)$$

$$\Rightarrow \boxed{A^T A c = A^T y} \leftarrow \text{the normal equations}$$

$$\Rightarrow \boxed{c = (A^T A)^{-1} A^T y} \leftarrow \text{Least squares solution}$$

or: $c = A^{\dagger} y$, $\boxed{A^{\dagger} := (A^T A)^{-1} A^T}$ ← Moore-Penrose pseudoinverse

Exercise: when can we solve this system? (Maybe easier, when can't we solve this system?)

- More generally: a least squares problem is:

$$\boxed{\min_{c \in \mathbb{R}^n} \|y - Ac\|_2^2}$$

where:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{mn} & \cdots & a_{mn} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

Can we solve this problem a bit more easily? Yes

Consider:

$$f(c) = \|y - Ac\|_2^2 = (y-Ac)^T(y-Ac)$$

$$\Rightarrow \quad Df(c) = D\left[(y-Ac)^T(y-Ac)\right]$$

$$= 2(y-Ac)^T D(y-Ac)$$

$$= -2(y-Ac)^T A$$

$$\Rightarrow \quad A^T(y-Ac) = 0 \quad \Rightarrow \quad \boxed{A^Ty = A^TAc} \leftarrow \text{normal equations}$$

How do we know that $DAc = A$?

Remember that one way the derivative of a scaler function can be computed is indirectly

$$f(x+\delta x) = f(x) + f'(x)\,\delta x + \underbrace{O(\delta x^2)}$$

↖ Landau big-O notation

$\longrightarrow$ usually we go this way...

this is the Taylor expansion of f about x

Simple case: $f(c) = Ac$. Then:

$$f(c+\delta c) = A(c+\delta c) = Ac + A\delta c + \underbrace{0}_{O(\|\delta c\|^2)}$$

How do we interpret this in the vector case?

Let's recall the multivariate Taylor expansion for a scalar field: $f(x+\delta x)$ $f(x) + Df(x)$

$$f(x+\delta x) = f(x) + \underbrace{\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\delta x_i}_{\text{how to interpret?}} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}\delta x_i \delta x_j + o(\|\delta x\|^2)$$

little-o!

Two ways:

1) $$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\delta x_i = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \vdots \\ \delta x_n \end{bmatrix} = \underbrace{\text{"}Df\text{"}}_{\substack{\text{Jacobian} \\ \text{matrix...} \\ \text{a row vector}}} \cdot \delta x$$

2) $$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\delta x_i = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}^T \begin{bmatrix} \delta x_1 \\ \vdots \\ \delta x_n \end{bmatrix} = \underbrace{\text{"}Df\text{"}}_{\text{gradient vector}}{}^T \delta x$$

Jacobian a bit easier to think about for a vector field, like $c \mapsto Ac$. So: conclude that $Df = D(Ac) = DA$

since $$f(c+\delta c) = \underbrace{Ac}_{f(c)} + \underbrace{A\delta c}_{Df\cdot\delta c} + \underbrace{0}_{O(\|\delta c\|^2)}.$$