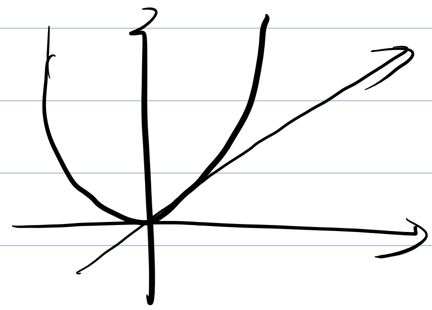
Youtube Link:

https://www.youtube.com/watch?v=oGZK3yGF-6k

Gradient

Suppose $y = x^T x$

$\Rightarrow dy = (2x)^T \cdot dx$ [dot product, small change of dx (vector)]



$$\underline{y - y_0 = 2x_0^T (x - x_0)}$$

Linearization $\Rightarrow$ plane

Matrix and vector product rule

$d(AB) = (dA) \cdot B + A \cdot (dB)$ (usually do not communicate)

Specially, $d(x^T x) = dx^T \cdot x + x^T \cdot dx$, since dot product can commute. $\Rightarrow d(x^T \cdot x) = 2x^T \cdot dx$

slight generally $d(u^T \cdot v) = du^T \cdot v + u^T \cdot dv$

($\text{note } du^T \cdot v = u^T \cdot dv$)

Gradient Derivation

$$f(x) = (Ax - b)^T (Ax - b)$$

$$df(x) = (A\,dx)^T (Ax - b) + (Ax - b)^T (A\,dx)$$

$$\text{(dot product)} = 2(Ax - b)^T A\,dx$$

$$= \underline{(2A^T(Ax - b))^T \cdot dx}$$

$\Downarrow$

Why this form?

geometric of gradient $\Rightarrow$ $\boxed{\text{hyperplane}}$ a dot product of $\boxed{O^T \cdot dx}$

$\Rightarrow$ setting $\frac{df}{dx} = 0 \Rightarrow A^TAx = A^Tb$ (least square)

$\nabla x^Tx = 2x$

$\nabla (\|Ax-b\|^2) = 2A^T(Ax-b)$

Trace

① a linear transformation from matrix to real

② Cyclic property $\text{trace}(ABC) = \text{trace}(BCA) = \text{trace}(CAB)$

③ $tr(A^T) = tr(A)$

④ $x^TAy = tr(x^TAy) = tr(yx^TA)$

$= tr((xy^T)^T \cdot A)$

$= \text{sum}((xy^T).^*A)$

If f is scalar and linear on $m\times n$ matrices, then there exists an $m\times n$ matrix M, such that $f(A) = \sum_{ij} M_{ij} \cdot A_{ij} = \text{sum}(M.^*A)$

$= tr(M^TA)$

Gradients of functions from matrices to scalars

eg: $f(A) = \text{trace}(A^2)$

$df = \text{trace}(A \cdot dA + dA \cdot A)$

$= tr((2A^T)^T \cdot dA)$ (cyclic rule)

gradient $= 2A^T$

in general $df = tr(M^TdA) \Rightarrow$ gradient $= M$, no indices needed

Side: Jacobian and Gradient

vector function R^n to R^m

$$f = \begin{pmatrix} f_1(x_1, \cdots, x_n) \\ f_2(x_1, \cdots, x_n) \\ \vdots \\ f_m(x_1, \cdots, x_n) \end{pmatrix}$$

n input scalars

m output scalars

$$\text{Jacobian} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

eg: $f(A) = (Ax-b)^T(Ax-b)$

$df = (dAx)^T(Ax-b) + (Ax-b)^T \cdot (dAx)$

$= tr(2(Ax-b)^T(dA)x]$

$= tr(2x(Ax-b)^T \cdot dA)$

$= tr([2(Ax-b)x^T]^T \cdot dA)$

So the gradient is $\underline{2(Ax-b)\cdot x^T}$ (elementary sum)

Jacobian Matrix

If $x \in R^n$ and $f(x)$ is a differentiable function whose values are in R^m, then the linearization is given by Jacobian matrix

$df = J \cdot dx$ where $J_{ij} = \partial f_i / \partial x_j$

eg: $f(x) = h((Wx-b))$ (h is a scalar function)

$df = \underbrace{h'(Wx-b) .^*}_{\text{element-wise}} (W dx)$ (chain rule)

$J = \text{diagonal}(h'(Wx+b)) * W$

Relationship to volumes (any dimension)

$df = J\,dx$

We have a small area around x is transformed

to a small area scaled by $|J|$.

Since J depends on x for nonlinear transformation,

the scaling is not uniform.

$d(A^{-1})$ $\quad X = A^{-1}$

$X \cdot A = I$

$d(X \cdot A) = d(I) = 0$

$dX \cdot A + X \cdot dA = 0$

$dX = -X \cdot dA \cdot A^{-1}$

$= -A^{-1} \cdot dA \cdot A^{-1}.$