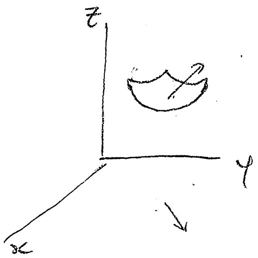
1/29/15 differentiation wrt vector & matrix

$f(x,y) = x^2 + 4xy + 3y^2 \qquad \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = (2x+4y,\ 4x+6y)$ "direction of max. change"



directional derivative: $D_u f = \nabla f \cdot u$ (unit vector)

e.g. u = unit vector in direction of gradient

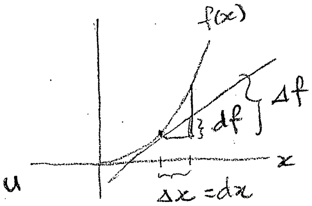
$D_u f = \nabla f \cdot \nabla f / |\nabla f| = |\nabla f|^2 / |\nabla f| = |\nabla f|$

$v = (x,y) \qquad f(v + \varepsilon u) \approx f(v) + \varepsilon D_u f$

$f(v+\varepsilon u) - f(v) \approx \varepsilon D_u f$

$\Delta f \approx (\Delta v)(D_u f)$

$\nabla f \cdot \Delta v,\ \Delta v = \varepsilon u$



increment $\{\Delta f \approx \Delta x f'(x)$

$\underbrace{df}_{\text{differential}} = dx f'(x)$

machine learning: optimization problems over $\mathbb{R}^d$ or $\mathbb{R}^{m \times n}$
can differentiate wrt each dimension separately,
but often easier to differentiate wrt the whole vector/matrix
since the derivative / differential can often be expressed
in terms of v / M.

differentiation wrt vector

def $x \in \mathbb{R}^d$, $f(x) \in \mathbb{R}^m$

$$\frac{\partial f}{\partial x} = \underbrace{\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & & \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_d}\end{pmatrix}}_{\text{Jacobian matrix}}$$

Sometimes defined as the transpose

ex. $x \in \mathbb{R}^d \quad A \in \mathbb{R}^{m \times d}$, not dependent on x.

$$\begin{pmatrix} a_{11} & \cdots & a_{1d} \\ & & \vdots \\ a_{m1} & \cdots & a_{md}\end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_d\end{pmatrix} = \begin{pmatrix} \sum_{i=1}^d a_{1i}x_i \\ \vdots \\ \sum_{i=1}^d a_{mi}x_i\end{pmatrix} \qquad \frac{\partial\left(\sum_{i=1}^d a_{ji}x_i\right)}{\partial x_k} = a_{jk}$$

$$f(x) = Ax \qquad \frac{\partial f}{\partial x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{md}\end{pmatrix} = A$$

ex. $f(x) = x^T A x$, where $A \in \mathbb{R}^{d \times d}$

$$x^T A x = \sum_{i=1}^d a_{1i}x_i x_1 + \sum_{i=1}^d a_{2i}x_i x_2 + \cdots + \sum_{i=1}^d a_{di}x_i x_d = \sum_{j=1}^d \sum_{i=1}^d a_{ji}x_i x_j$$

$$\frac{\partial f}{\partial x_k} = \sum_{i \neq k} a_{ki}x_i + 2a_{kk}x_k^2 + \sum_{j \neq k} a_{jk}x_j = \sum_{i=1}^d a_{ki}x_i + \sum_{j=1}^d a_{jk}x_j$$

= 1×d matrix, so vector is on the left and is transposed

$$\frac{\partial f}{\partial x} = x^T A^T + x^T A = x^T(A^T + A)$$

in particular: if A is symmetric, $\frac{\partial}{\partial x}(x^T A x) = 2x^T A$.

what about $\frac{\partial}{\partial s}\left((x-s)^T A (x-s)\right)$ — s represents a translation

easy if we have some sort of chain rule, but a pain to prove it

instead: $(x-s)^T A (x-s) = x^T A x - s^T A x - x^T A s - s^T A s$

diff. each term: $2x^T A - s^T A - s^T A \qquad (x^T A s = s^T A^T x = s^T A x)$

$= 2(x-s)^T A$

$x^T A x$ for symmetric A is called a quadratic form: represents a homogeneous polynomial of deg 2.

example from before, $x^2 + 4xy + 3y^2 = (x\ \ y)\begin{pmatrix}1 & 2 \\ 2 & 3\end{pmatrix}\begin{pmatrix}x \\ y\end{pmatrix} \qquad \frac{\partial f}{\partial v} = 2(x+2y,\ 2x+3y)$

$= (2x+4y,\ 4x+6y)$ ✓

ex. ridge regression objective function (generalization of linear regression)

$$J_\lambda(\theta) = \left[\sum_{i=1}^m (x_i^T\theta - Y_i)^2\right] + \lambda \sum \theta_i^2 \quad \text{where } x_i, \theta \in \mathbb{R}^d,\ Y \in \mathbb{R}^m,\ \lambda \in \mathbb{R}^+,\ 1 \le i \le m$$

$$= \|X\theta - Y\| + \lambda\|\theta\| \quad \text{where } X = \begin{pmatrix} x_1 \\ \vdots \\ x_m\end{pmatrix} \in \mathbb{R}^{m \times d}$$

$$\frac{\partial J_\lambda(\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\left[(X\theta - Y)^T(X\theta - Y) + \lambda\theta^T\theta\right] = \frac{\partial}{\partial\theta}\left[\theta^T X^T X\theta - Y^T X\theta - \theta^T X^T Y + Y^T Y + \lambda\theta^T\theta\right]$$

$$= 2\theta^T X^T X - 2Y^T X + 2\lambda\theta^T \xrightarrow{\text{set to } 0} \theta^T X^T X + \lambda\theta^T = Y^T X$$

$$\theta^T(X^T X + \lambda I) = Y^T X$$

$$\theta = \underbrace{(X^T X + \lambda I)}_{\text{positive diagonal, invertible}}{}^{-1} X^T Y \Leftarrow \theta^T = Y^T X (X^T X + \lambda I)^{-1}$$

solution to linear regression ($\lambda = 0$ case)

$\theta = (X^T X)^{-1} X^T Y$ } hat matrix, the inverse may not exist