* Lecture 6

2.4.3.1 (In Book)

• Matrix Differentiation

Consider $f: \mathbb{R}^n \to \mathbb{R}^m$

$f(\vec{x}) = (f_1(\vec{x}), \dots, f_m(\vec{x}))$, $f_i: \mathbb{R}^n \to \mathbb{R}$, $1 \le i \le m$

$1 \le j \le n$

Define the Jacobian of f.

$$J(f) = \frac{\partial}{\partial \vec{x}} f(\vec{x}) = \left(\frac{\partial f_i(\vec{x})}{\partial x_j} \right)_{i,j=1,1}^{m,n} \in \mathbb{R}^{m \times n}$$

map $\mathbb{R}^3$ to $\mathbb{R}^2$

E.g. Consider $f(x,y,z) = \begin{pmatrix} x^2+yz \\ xyz \end{pmatrix} \in \mathbb{R}^2$

$f_1(\vec{x}) = x^2+yz$

$f_2(\vec{x}) = xyz$

$$J(f) = \frac{\partial}{\partial \vec{x}} f(\vec{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \end{pmatrix} = \begin{pmatrix} 2x & z & y \\ yz & xz & xy \end{pmatrix}$$

Remark: If $f: \mathbb{R}^m \to \mathbb{R}$ then $\frac{\partial f}{\partial \vec{x}}(\vec{x}) = \nabla_{\vec{x}} f(\vec{x})^T$ — this is the gradient.

E.g. $f(\vec{x}) = x^2+y^2+z^2$

$f: \mathbb{R}^3 \to \mathbb{R}$

$$\frac{\partial f}{\partial \vec{x}}(\vec{x}) = (2x \;\; 2y \;\; 2z) = \nabla_{\vec{x}} f(\vec{x})^T$$

E.g. $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and define $f(\vec{x}) = A\vec{x} + \vec{b}$

$f: \mathbb{R}^n \to \mathbb{R}^m$

$$\frac{\partial f}{\partial \vec{x}}(\vec{x}) = \left(\frac{\partial f_i}{\partial x_j}(\vec{x}) \right)_{i,j=1,1}^{m,n} = (A_{i,j})_{i,j=1,1}^{m,n} = A$$

$f_i(\vec{x}) = A_{i*}\vec{x} + b_i$

ith row of A times $\vec{x}$

$= (\sum_{i,j} A_{ij} x_j) + b_i$

Corollary: $\frac{\partial}{\partial \vec{x}} \vec{x} = I_{n\times n}$, $x \in \mathbb{R}^n$

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

$= I_{n\times n} \vec{x}$

• Rules ① $\frac{\partial}{\partial \vec{x}}(A\vec{x} + \vec{b}) = A$ ($A, \vec{b}$ constant)

② $\frac{\partial}{\partial \vec{x}}(\vec{x}^T A) = A^T$ (A constant)

A is a matrix compatible with $\vec{x}^T A$.

③ $\frac{\partial}{\partial \vec{x}}(\vec{x}^T A \vec{x}) = \vec{x}^T(A + A^T)$ (A constant)

f & g have compatible dimensions, so the products make sense.

④ $\frac{\partial}{\partial \vec{x}} f(\vec{x})^T g(\vec{x}) = g(\vec{x})^T \frac{\partial}{\partial \vec{x}} f(\vec{x}) + f(\vec{x})^T \frac{\partial}{\partial \vec{x}} g(\vec{x})$ similar to some product rules.

⑤ $\frac{\partial}{\partial \vec{x}} f(\vec{x})^T A g(\vec{x}) = g(\vec{x})^T A^T \frac{\partial}{\partial \vec{x}} f(\vec{x}) + f(\vec{x})^T A \frac{\partial}{\partial \vec{x}} g(x)$

• Proofs/more rules in reference "Matrix Calculus" on Canvas.

E.g. $\nabla_{\vec{y}}(-2\vec{x}^T D\vec{y} + \vec{y}^T\vec{y}) = 0$ (recall using this for PCA (from a past lecture))

$\Rightarrow \frac{\partial}{\partial \vec{y}}(-2\vec{x}^T D\vec{y} + \vec{y}^T\vec{y}) = 0$ can we rule ① here, to get.

linearity $\Rightarrow -2\frac{\partial}{\partial \vec{y}}(\vec{x}^T D\vec{y}) + \frac{\partial}{\partial \vec{y}}(\vec{y}^T\vec{y}) = 0$

use rule ⑤ $\overset{⑤}{\Rightarrow} -2\left(\vec{y}^T D^T \frac{\partial}{\partial \vec{y}}\vec{x} + \vec{x}^T D \frac{\partial}{\partial \vec{y}}\vec{y}\right) + \frac{\partial}{\partial \vec{y}}(\vec{y}^T\vec{y}) = 0$ ($\frac{\partial}{\partial \vec{y}}\vec{x} = 0$, $\frac{\partial}{\partial \vec{y}}\vec{y} = I$)

$\overset{③}{\Rightarrow} -2\vec{x}^T D I + \vec{y}^T(I + I) = 0$ ($I^T = I$)

$\Rightarrow -2\vec{x}^T D = -2\vec{y}^T$

$\Rightarrow \vec{y} = D^T\vec{x}$

Notice:

* $\frac{\partial}{\partial \vec{x}}(af(\vec{x}) + bg(\vec{x})) = a\frac{\partial}{\partial \vec{x}} f(\vec{x}) + b\frac{\partial}{\partial \vec{x}} g(\vec{x})$

more general form. (linearity)

• The Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$ is

$$H(f)(\vec{x}) = J(\nabla_x f(\vec{x})) = \frac{\partial}{\partial \vec{x}} \frac{\partial}{\partial \vec{x}} f(\vec{x})$$

$$= \frac{\partial^2}{\partial \vec{x}^2} f(\vec{x})$$

$$= \left(\frac{\partial^2}{\partial x_i \partial x_j} f(\vec{x}) \right)_{i,j=1..n}^{n,n}$$

① Remark: At points $\vec{x}$ s.t. all 2nd partials of f are continuous,

$$\frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f(\vec{x}) = \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} f(\vec{x})$$

so $H(f)(\vec{x})$ is symmetric, so $H(f)(\vec{x})$ has an orthonormal eigenvector decomposition and has real eigenvalues

② $\hat{u}$ unit vector

$\hat{u}^T H(f)(\vec{x}) \hat{u}$ is the 2nd derivatives of $f(\vec{x})$ in the direction $\hat{u}$.

Let's go back to gradient descent

(find $\underset{\vec{x}}{\operatorname{argmin}} f(\vec{x})$)

let $\vec{g} = \nabla_{\vec{x}} f(\vec{x})$ $\quad$ $q = \frac{\vec{x} - \vec{x}_0}{\|\vec{x} - \vec{x}_0\|}$ unit vector

$H = H(f)(\vec{x}_0)$

apply the Taylor expansion

$$f(\vec{x}) \approx f(\vec{x}_0) + \underbrace{(\vec{x} - \vec{x}_0)^T \vec{g}}_{\text{in } D_q f(\vec{x})} + \underbrace{\frac{1}{2}(\vec{x} - \vec{x}_0)^T H (\vec{x} - \vec{x}_0)}_{\text{in 2nd derivative of } f \text{ in direction } \hat{u}}$$

$D_{\hat{u}} f(\vec{x}) = \hat{u} \cdot \nabla_{\vec{x}} f(\vec{x})$

$\downarrow$

$\hat{u}^T$

*Lecture X

More on Gradient Descent.

Let $\vec{x}_0 \in \mathbb{R}^n$, $f: \mathbb{R}^n \to \mathbb{R}$

$\vec{g} = \nabla_{\vec{x}} f(\vec{x}_0)$ $\quad \hat{u} = \frac{\vec{x} - \vec{x}_0}{\|\vec{x} - \vec{x}_0\|_2}$

$H = H(f)(\vec{x}_0)$

Then $f(\vec{x}) \approx f(\vec{x}_0) + (\vec{x} + \vec{x}_0)^T \vec{g} + \frac{1}{2}(\vec{x} - \vec{x}_0)^T H (\vec{x} - \vec{x}_0)$

Gradient descent step

$f(\vec{x}_0 - \alpha \vec{g}) \approx f(\vec{x}_0) - \alpha \vec{g}^T \vec{g} + \frac{1}{2}\alpha^2 \vec{g}^T H \vec{g}$

keep going along the direction of the gradient, unless we reach a stopping position.

☺ tells how close $\vec{x}_0$ & $\vec{g}$ are depending on what α is. (these?)

α is the stopping condition

case 1 $\vec{g}^T H \vec{g} > 0$ (guaranteed if H is positive definite).

we are trying to minimize f to get the largest possible drop

Compute the optimal step size:

$\frac{\partial}{\partial \alpha} f(\vec{x}_0 - \alpha \vec{g}) = -\vec{g}^T \vec{g} + \alpha \vec{g}^T H \vec{g} \Rightarrow \boxed{\alpha = \frac{\vec{g}^T \vec{g}}{\vec{g}^T H \vec{g}}}$

Hessian?

case 2 $\vec{g}^T H \vec{g} < 0$ (previous case gives a local max ... no good)

- set sth. wrong
- the problem is now meaningless.

Taylor series says increasing & forever decreases f
(probably NOT true ... Taylor series is only accurate locally)

when $\vec{g}^T H \vec{g}$, probably looking @ sth. that's a saddle

2nd Derivative Test also a multi-variable version

suppose $\nabla_{\vec{x}} f(\vec{x}_0) = \vec{0}$

① If $H(f)(\vec{x}_0)$ is positive definite, then $\vec{x}_0$ is a local min
② If " " negative " " " " " local max
③ If $H(f)(\vec{x}_0)$ has both positive and negative eigenvalues, $\vec{x}_0$ is a saddle point.

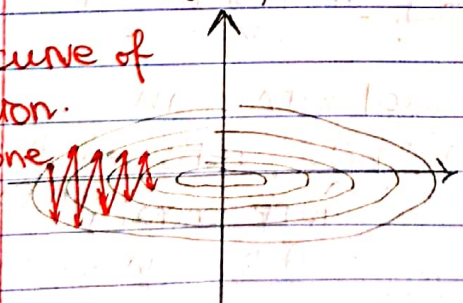
(④. $H(f)(\vec{x}_0) = 0$ (?) inconclusive, same as not saying anything .....)

Remark: The condition number $\max_{i,j} \left| \frac{\lambda_i}{\lambda_j} \right|$ of H indicates how well a gradient descent step will perform.

large → bad
small (1) → good.

suppose $m \geq 2$,

$H(f) = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$



the level curve of the function.
Idea: in one direction, stretch very far.

- horrible path.
- goes "forever"
- terrible performance when step size is fixed & large.

Fixes: ① Use Hessian info (algorithm centric)
② Normalize dataset (data centric)

take a look @ mean & sdv for each variable; then perform some transformation that sdv of each is similar. (more normalized :))

Q: when can we gaurantee that gradient descent (or some variation) will converge?

Answer: In deep learning, almost never, BUT for some the function is so complicated that it never has the best properties (e.g. neuro network).

special functions, we may get guarantees.

Defn: A function $f: \mathbb{R}^n \to \mathbb{R}^m$ is L-Lipschitz continuous if $\exists L > 0$, s.t. $\forall \vec{x}, \vec{y} \in \mathbb{R}^n$,

$$\|f(\vec{x}) - f(\vec{y})\|_2 \leq L \|\vec{x} - \vec{y}\|_2.$$

L: a positive constant (bounds the possible change in function value based on size of change in inputs.)

can be Lipschitz cts with many diff. values of L, BUT the most imp is the smallest L. usually knowing the function satisfy some L the properties will follow :)

E.g. $f(\vec{x}) = A\vec{x} + \vec{b}$

$$f(\vec{x}) - f(\vec{y}) = A\vec{x} + \vec{b} - A\vec{y} - \vec{b} = A(\vec{x} - \vec{y})$$

$$\therefore \max_{\text{here :)}} \|f(\vec{x}) - f(\vec{y})\|_2 = \|A(\vec{x} - \vec{y})\|_2 \leq \|A\|_2 \|\vec{x} - \vec{y}\|_2$$

$$\|A\|_2 = \max_{\vec{x} \neq \vec{0}} \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2}$$

so $\forall \vec{x} \in \mathbb{R}^n$, $\|A\|_2 \|\vec{x}\| \geq \|A\vec{x}\|_2$

so f is $\|A\|_2$-Lipschitz continuous.

this is the minimum vale of L as well.

Defn: $f: \mathbb{R}^n \to \mathbb{R}$ is convex if $H(f)(\vec{x})$ is positive semi-definite for all $\vec{x} \in \mathbb{R}^n$.

" " strict convexity " positive definite

" " " extra ensurity such as unique local minimum.

E.g. Consider $f(\vec{x}) = \vec{b}^T\vec{x} + \vec{x}^T A \vec{x} + C$. Let's see if this is convex!

to be general :) A is a symmetric matrix.

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

$$= \vec{b}^T + 2\vec{x}^T A$$

$$\frac{\partial^2 f}{\partial \vec{x}^2}(\vec{x}) = 2A^T = 2A.$$

the rule says

so f is convex iff A is positive semi-definite.

E.g. $f(\vec{x}) = \vec{x}^T\vec{x} = \|\vec{x}\|_2^2$ is convex.

$A = I$ here, so positive definite $\therefore$ strictly convex.

*Lecture 8

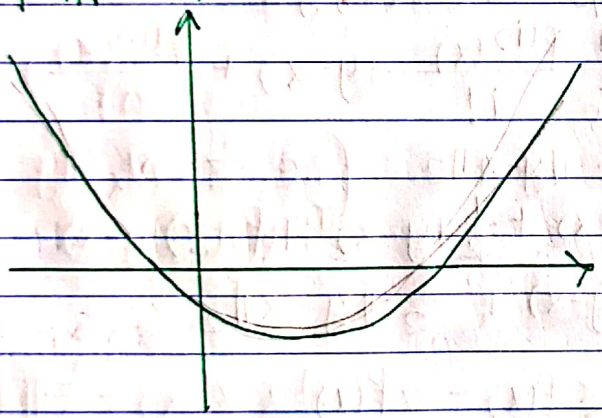
Theorem 1:

Remark: The definition I gave for convexity can be made more general.

Defn: $f: \mathbb{R}^n \to \mathbb{R}$ is convex if $\forall \vec{x}, \vec{y} \in \mathbb{R}^n \ \forall 0 \le t \le 1$
$f(t\vec{x} + (1-t)\vec{y}) \le tf(\vec{x}) + (1-t)f(\vec{y})$

Theorem: If f is twice differentiable on $\mathbb{R}^n$ then f is convex $\Leftrightarrow$ $H(f)(\vec{x})$ positive semi-definite. $\forall \vec{x} \in \mathbb{R}^n$

E.g. $f: \mathbb{R} \to \mathbb{R}$



Q: How to see general defn here?

Theorem: suppose $f: \mathbb{R}^n \to \mathbb{R}$ is twice differentiable convex, and has L-Lipschitz continuous gradient, and f has a global min $x^* \in \mathbb{R}^n$. Then if x_k is the k^{th} step of gradient descent, with fixed step size $\alpha < 1/L$, then

$$f(x_k) - f(x^*) \le \frac{\|x_0 - x^*\|_2^2}{2\alpha k}$$

proof: ∇f L-Lipschitz $\Rightarrow$ $H(f)(\vec{x}) - LI$ negative semi definite.

proof: $f_{\vec{x},\hat{u}}(t) = f(\vec{x} + t\hat{u}) \quad t \in \mathbb{R}$

then $f'_{\vec{x},\hat{u}}(0) = \hat{u}^T \nabla f(\vec{x})$

$\hat{u}^T(H(f)(\vec{x}) - LI)\hat{u} = \hat{u}^T H(f)(\vec{x})\hat{u} - \hat{u}^T L I \hat{u}$

$= \quad " \quad - L$

$= f''_{\vec{x},\hat{u}}(0) - L = \lim_{h \to 0} \frac{f'_{\vec{x},\hat{u}}(h) - f'_{\vec{x},\hat{u}}(0)}{h} - L$

$$= \lim_{h\to 0} \frac{\hat{u}\cdot[\nabla_{\vec{x}} f(\vec{x}+h\hat{u}) - \nabla_{\vec{x}} f(\vec{x})]}{h} - L$$

$$\leq \lim_{h\to 0} \frac{\|\nabla_{\vec{x}} f(\vec{x}+h\hat{u}) - \nabla_{\vec{x}} f(\vec{x})\|_2}{h} - L$$

$$\leq \lim_{h\to 0} \frac{L\|h\hat{u}\|_2}{h} - L = 0 \qquad \frac{\partial}{\partial ?}$$

$\therefore \forall x,y,z \in \mathbb{R}^n . (x-y)^T (H(f)(z) - LI)(x-y) \leq 0$

Let $x, y \in \mathbb{R}^n$. By Taylor's Remainder Theorem,

(In $\mathbb{R}$. TRT says $\exists z \in [x,y]$

$$f(y) = \sum_{i=0}^{n-1} \frac{f^{(i)}(x)}{i!}(y-x) + \frac{f^{(n)}(z)}{n!}(y-x)$$)

$\exists z$ s.t. $\|z-x\|_2 \leq \|y-x\|_2$ (i.e. $z \in N_x(\|y-x\|_2)$)

$$f(y) = f(x) + \nabla f_x(x)^T (y-x) + \frac{1}{2}(y-x)^T H(f)(z)(y-x)$$

$$\leq f'(x) + \nabla_x f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2$$

Now, $f(x - \alpha \nabla_x f(x)) \leq f(x) - \nabla f(x)^T \alpha \nabla f(x) + \frac{L}{2}\|\alpha \nabla_x f(x)\|_2^2$

$$= f(x) - (1 - \frac{L\alpha}{2})\alpha \|\nabla_x f(x)\|_2^2$$

$0 < \alpha < 1/L$.

Since $0 < \alpha < 1/L \iff 0 < \frac{\alpha L}{2} < \frac{1}{2}$

$$\iff 0 > \frac{-\alpha L}{2} > \frac{-1}{2}$$

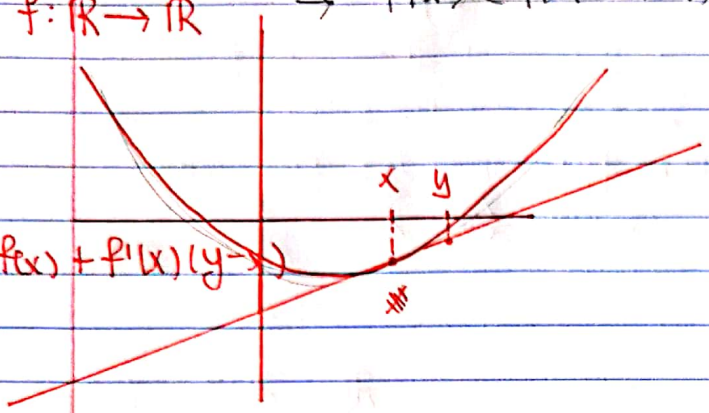
$$\iff 1 > \boxed{1 - \frac{\alpha L}{2} > 1 - \frac{1}{2}}$$

$\neq 1/2$

$f(x - \alpha \nabla_x f(x)) \leq f(x) - \frac{\alpha}{2}\|\nabla_x f(x)\|_2^2$ ($\Rightarrow$ sequence $(f(x_i))_{i\geq 0}$ weakly decreasing)

f convex $\Rightarrow$ $f(x^*) \geq f(x) + \nabla_x f(x)^T (x^* - x)$

note: $f: \mathbb{R} \to \mathbb{R}$ $\Rightarrow$ $f(x) \leq f(x^*) + \nabla_x f(x)^T (x - x^*)$.



$$f(x_{i+1}) = f(x_i - \alpha \nabla_x f(x_i)) \leq f(x_i) - \frac{\alpha}{2} \| \nabla_x f(x_i) \|_2^2$$
$$\leq f(x^*) + \nabla_x f(x_i)^T (x_i - x^*) - \frac{\alpha}{2} \| \nabla_x f(x_i) \|_2^2$$

$$x_{i+1} = x_i - \alpha \nabla_x f(x_i) \Rightarrow \nabla_x f(x_i) = \frac{x_i - x_{i+1}}{\alpha}$$

$$f(x_{i+1}) \leq f(x^*) + \nabla_x f(x_i)^T (x_i - x^*) - \frac{\alpha}{2} \| \nabla_x f(x_i) \|_2^2$$
$$= f(x^*) + \frac{1}{\alpha} (x_i - x_{i+1})^T (x_i - x^*) - \frac{\| x_i - x_{i+1} \|_2^2}{2\alpha}$$
$$= f(x^*) + \frac{1}{\alpha} (x_i - x_{i+1})^T (x_i - x^*) - \frac{1}{2\alpha} (x_i - x_{i+1})^T (x_i - x_{i+1})$$

Complete the square

$$= f(x^*) - \frac{1}{2\alpha} [(x_i - x_{i+1})^T (x_i - x_{i+1}) - 2(x_i - x_{i+1})^T (x_i - x^*)$$
$$+ (x_i - x^*)^T (x_i - x^*) - (x_i - x^*)^T (x_i - x^*)]$$

$$= f(x^*) - \frac{1}{2\alpha} [((x_i - x_{i+1}) - (x_i - x^*))^T ((x_i - x_{i+1}) - (x_i - x^*)) - \| \quad]$$

$$= f(x^*) - \frac{1}{2\alpha} [\| x^* - x_{i+1} \|_2^2 - \| x_i - x^* \|_2^2]$$

$$\sum_{i=0}^{k} (f(x_i) - f(x^*)) = \sum_{i=0}^{k} \frac{1}{2\alpha} [\quad \| \quad]$$

$$= \frac{1}{2\alpha} [\| x^* - x_0 \|_2^2 - \| x^* - x_k \|_2^2]$$
$$\leq \frac{1}{2\alpha} \| x^* - x_0 \|_2^2$$

$(f(x_i))$ decreasing $\Rightarrow f(x_k) - f(x^*) \leq f(x_i) - f(x^*) \quad \forall i < k$

$$\therefore k(f(x_k) - f(x^*)) = \sum_{i=1}^{k} (f(x_k) - f(x^*))$$

$$\leq \sum_{i=1}^{k} (f(x_i) - f(x^*))$$

$$\leq \frac{1}{2\alpha} \| x_0 - x^* \|_2^2$$

■