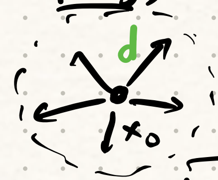
MATH FOR ML : UNIT 2 (CALC + OPT) REVIEW.

I. DERIVATIVE NOTIONS (total, gradient, Jacobian, etc.)
II. TAYLOR SERIES + IMPLICATIONS.
III. OPTIMIZATION w/ or w/o CONSTRAINTS.
IV. GRADIENT DESCENT.

DERIVATIVE NOTIONS

- Single-variable calculus : only two directions (pos. or neg.)
- Multi-variable calculus : inf. many directions



$$F: \mathbb{R}^d \to \mathbb{R}^n$$

only make sense AT a fixed $\vec{x}_0 \in \mathbb{R}^d$.

TOTAL DERIVATIVE
(Lin. Transformation $Df_{\underline{\underline{x_0}}}: \mathbb{R}^d \to \mathbb{R}^n$)

is — when $n=1$ — **Gradient**: $\nabla f(x_0)^T \in \mathbb{R}^{1\times d}$

$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_d}\right)$ written down

second Derivative: **HESSIAN** (For $F: \mathbb{R}^d \to \mathbb{R}$)

is — when n is arbitrary. — **Jacobian**: $\nabla f(x_0) \in \mathbb{R}^{n\times d}$

Gradient gives... **DIRECTIONAL DERIVATIVE**: $\boxed{\nabla f(x_0)^T d}$
(Scalar in $\mathbb{R}$ that says "how fast F changes in d direction.")

when $d = e_i$: **PARTIAL DERIVATIVE**
(change in F in a basis direction)

In general :

$$\begin{bmatrix} - \nabla f_1(x)^T - \\ \vdots \\ - \nabla f_n(x)^T - \end{bmatrix}$$

In Jacobian case :

$$\begin{bmatrix} - \nabla F - \\ \vdots \\ - \nabla F_n - \end{bmatrix} \begin{bmatrix} | \\ d \\ | \end{bmatrix} = \begin{bmatrix} \text{how much } F_1 \text{ changes} \\ \vdots \\ \text{how much } F_n \text{ changes} \end{bmatrix}$$

in direction d.

TAYLOR SERIES

At x_0: $\underbrace{\underbrace{F(x) = F(x_0) + F'(x_0)(x - x_0)}_{\text{Lin. Approximation.}} + \frac{F''(x_0)}{2!}(x - x_0)^2}_{\text{second order Approx.}} + \dots$

① **DESCENT LEMMA**
$f(x_t) \leq f(x_{t-1}) - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2$

+ Remainder: Bonded by smoothness on Quad. Form.

③ Proof of First order Convexity
$f(x) + \nabla f(x)^T(y - x) \leq f(y)$

$$F(x) = F(x_0) + \nabla F(x_0)^T(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 F(x_0)(x - x_0) + \dots$$

② Opt. Necessary cond. $\nabla F(x_0) = 0$, $\nabla^2 F(x_0)$ PSD. (otherwise: Candidate local min > some close point)

Opt. Suff. cond. $\nabla F(x_0) = 0$, $\nabla^2 F(x_0)$ PD.

(PD / PSD Notation: $A \succeq 0$ PSD, $A \succ 0$ PD)

OPTIMIZATION PROBLEMS.

Necessary: This has to true for P to be true.

Sufficient: If this is true then P is true.

$F: \mathbb{R}^d \to \mathbb{R}$.

x^* = global minimizer.

$$\boxed{\begin{array}{ll} \text{minimize} & F(\vec{x}) \\ \text{subject to} & \vec{x} \in \mathcal{C}. \end{array}}$$

local minima.

① UNCONSTRAINED ($\mathcal{C} = \mathbb{R}^d$)

(i) Necessary: $\nabla F(x^*) = 0$,
$\nabla^2 F(x^*)$ PSD.

(ii) Sufficient: $\nabla F(x^*) = 0$,
$\nabla^2 F(x^*)$ PD.

③ CONVEX

$\mathcal{C}$ = convex
F convex

All local minima are global minima.

GD is guaranteed to minimize!

② CONSTRAINED $\mathcal{C} \subset \mathbb{R}^d$

Ⓐ Equality Constraints

$$\begin{array}{ll} \text{minimize} & F(\vec{x}) \\ \text{subject to} & h_1(x) = 0 \\ & \vdots \\ & h_m(x) = 0 \end{array}$$

Solve using Lagrangian:

$$L(\vec{x}, \vec{\lambda}) = F(x) + \sum_{i=1}^{m} \lambda_i h_i(\vec{x})$$

(Mind regular points). Reduce: $L(x, \lambda)$

Ⓑ Inequality Constraints

Subject to $g_1(x) \le 0$ ... $g_m(x) \le 0$

Solve using Lagrangian:

$$L(x, \lambda, \mu) = F(x) + \sum \lambda_i h_i(x) + \sum \mu_j g_j(x)$$

complementary slackness

$\mu^* \ge 0$ & $\mu_j^* g_j(x^*) = 0$ $\forall j \in [r]$

GRADIENT DESCENT.

$$\boxed{f(x_t) \leftarrow f(x_{t-1}) - \eta \nabla f(x_{t-1}).}$$

- Taylor's Thrm
- smoothness of f (Bounded Hessian eigenvalues).

IF CONVEX F. ...

① DESCENT LEMMA

$$f(x_t) \le f(x_{t-1}) - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2$$

("Updates make the function smaller if $\nabla f(x_{t-1}) \neq 0$".)

② GD ON CONVEX

$$F(x_T) \lesssim \frac{B}{T}\|x_0 - x^*\|^2 + F(x^*)$$

("GD eventually reaches the global minimum".)

- use Descent Lemma.
- First-order Def. of convexity.
- Potential Argument.

QUADRATIC FORM.

Quadratic Form $\quad x^TAx \leftarrow$

In $d=1$:

$xax = ax^2$

Quadratic Function. $\quad x^TAx + b^Tx + c.$ (A is matrix, b vector, c scalar)

Taylor Series: $\overset{\mathbb{R}}{F(x)} = F(x_0) + F'(x_0)(x-x_0) + \frac{F''(x_0)}{2}(x-x_0)^2 + \ldots$

$\mathbb{R}^d$

$F(\underline{x}) = F(\underline{x_0}) + \nabla F(x_0)^T(x-x_0) + (x-x_0)^T \nabla^2 F(x_0)(x-x_0)$

MVC

① Quad. Function: $x^TAx + b^Tx + c$

$ax^2 + bx + b^2/4a$

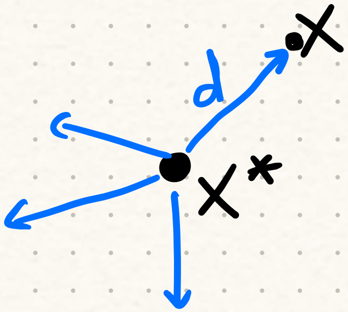
② Single variable quad: $ax^2 + \underline{b}x + \underline{c}. \longrightarrow (\sqrt{a}\,x + \frac{b}{2\sqrt{a}})^2 - \frac{b^2}{4a} + c$

$$\boxed{(\sqrt{a}\,x + b/2\sqrt{a})^2 - (\frac{b^2}{4a} - c).}$$

HS: $(x-\alpha)^2 + \beta \quad \Rightarrow$ so b and c are only responsible for shifting the parabola.

INTUITION FOR NECESSARY / SUFFICIENT.

From top down on $\mathbb{R}^d$



$$F(x) = F(x^*) + \nabla F(x^*)^T(x - x^*) + (x-x^*)^T \nabla^2 F(x^*)(x-x^*)$$

$$\Updownarrow$$

$$F(x^*+d) = F(x^*) + \nabla F(x^*)^T d + d^T \nabla^2 F(x^*) d.$$

SUFFICIENT: $\nabla F(x^*) = 0$, $\nabla^2 F(x^*)$ PD $\Longrightarrow$ Minimum.

$$\Longrightarrow$$

$$F(x^*+d) = F(x^*) + \underbrace{\cancel{\nabla F(x^*)^T d}}_{0} + \underbrace{d^T \nabla^2 F(x^*) d}_{>0}.$$

$F(x^*+d) = F(x^*) +$ Positive #

NECESSARY: $\nabla F(x^*) = 0$, $\nabla^2 F(x^*)$ PSD $\Longleftarrow$ Minimum.

$F(x^*+d) = F(x^*) + 0 +$ Negative #

$\hookrightarrow$ for some d.