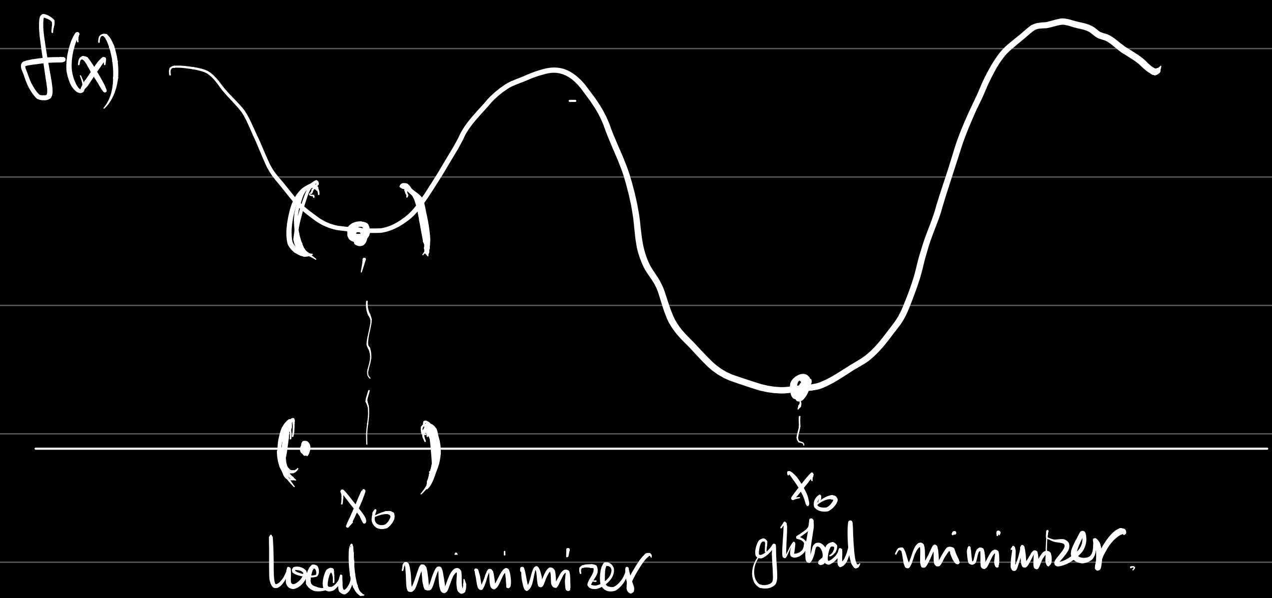
Recall: (1) a point $x_0 \in D(f)$ is called a local minimizer of f if $\exists\, \epsilon > 0$ s.t.

$$f(x_0) \leq f(x) \quad \text{for any } x \in \{x \in D(f) : \|x - x_0\| < \epsilon\}$$



(2) a point $x_0 \in D(f)$ is called a global minimizer of f if

$$f(x_0) \leq f(x) \quad \text{for any } x \in D(f)$$

(3) For $f \in C^1$, a point $x_0 \in D(f)$ is called a critical point of f if

$$\nabla f(x_0) = 0.$$

Remark: For $f \in C^1$, a necessary condition for x_0 being a local minimizer of f is that x_0 is a critical point.

Prop: For a convex optimization problem, a local minimizer is also a global minimizer.

Proof: Let $\min_{D} f(x)$ be a convex optimization problem where

$$D \stackrel{\text{def}}{=} D(f) \cap \{x : h_i(x) \leq 0,\ i=1,\cdots,I\} \cap \{x : g_j(x) = 0,\ j=1,\cdots,J\}.$$

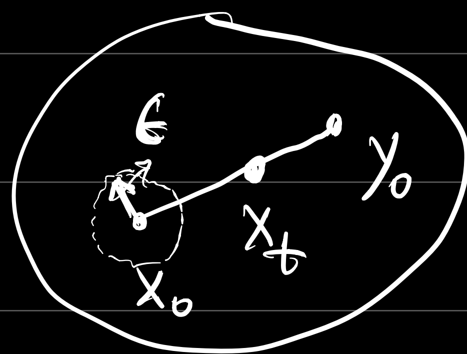
(D is convex)

Let $x_0 \in D$ be a local minizer of f, then $\exists\, \epsilon > 0$ s.t.

$$f(x_0) \leq f(x) \quad \text{for any } x \in \{D : \|x - x_0\| < \epsilon\}$$

Suppose x_0 is not a global minimizer, i.e., $\exists\, y_0 \in D$ s.t.

$$f(y_0) < f(x_0),$$



then $x_t \overset{\text{def}}{=} t y_0 + (1-t) x_0$, $\quad 0 \leq t \leq 1$.

We have $x_t \in D$ for any $0 \leq t \leq 1$.

In particular, if t is close to 0, x_t is close to x_0.

Choose t small s.t. $\|x_t - x_0\| < \epsilon$, but $\overset{\text{def of } x_t}{f(x_t) = f(ty_0 + (1-t)x_0)}$

$$\underset{f \text{ convex}}{\leq} t f(y_0) + (1-t) f(x_0)$$

$$\underset{f(y_0) < f(x_0)}{<} t f(x_0) + (1-t) f(x_0) = f(x_0)$$

contradicting that x_0 is a local minimizer

∎

We record a few rules of differentiation for vector-valued func.

Lemma: Let $x, y \in \mathbb{R}^n$ be two independent vectors and $A \in \mathbb{R}^{n \times n}$ be a sym. matrix. Then

(1) $$\frac{\partial (x \cdot y)}{\partial x} = \frac{\partial (x^T y)}{\partial x} = \frac{\partial (y^T x)}{\partial x} = y$$

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

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

Proof: (1) $x \cdot y = x^T y = y^T x = x_1 y_1 + \cdots + x_n y_n$

so $\frac{\partial (x \cdot y)}{\partial x_i} = \frac{\partial (x_1 y_1 + \cdots + x_i y_i + \cdots + x_n y_n)}{\partial x_i} = y_i$

i.e, i-th component of $\nabla_x (x \cdot y)$
= i-th component of y.

i.e. $\nabla_x (x \cdot y) = y$. ∎

(2) $x^T A x = \sum_{k,l=1}^{n} x_k A_{kl} x_l$

so $\frac{\partial}{\partial x_i}(x^T A x) = \sum_{k,l=1}^{n} \frac{\partial}{\partial x_i}(x_k A_{kl} x_l)$

$$= \sum_{k,l=1}^{n} \left(\frac{\partial x_k}{\partial x_i} A_{kl} x_l + x_k A_{kl} \frac{\partial x_l}{\partial x_i} \right)$$

$$= \sum_{k,l=1}^{n} \delta_{ki} A_{kl} x_l + \sum_{k,l=1}^{n} x_k A_{kl} \delta_{li}$$

$$= \sum_{l=1}^{n} A_{il} x_l + \sum_{k=1}^{n} x_k A_{ki}$$

$$= (Ax)_i + (A^T x)_i$$

$$\overset{A \text{ sym}}{=} 2(Ax)_i$$

Thus $\frac{\partial}{\partial x}(x^T A x) = 2Ax$, where

Kronecker delta: $\delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$

(3) $$\frac{\partial^2 (x^T A x)}{\partial x_i \, \partial x_j} = \frac{\partial (2Ax)_i}{\partial x_j} = \frac{\partial \left(2 \sum_{k=1}^{n} A_{ik} x_k \right)}{\partial x_j}$$

$$= 2\sum_{k=1}^{n} A_{ik}\delta_{kj} = 2A_{ij}$$

Thus the Hessian of $x^T A x = 2A$ ∎

2.2.2 Theory.

$$\beta^{ridge} = \arg\min_{\beta} \underbrace{\|y - x^T\beta\|^2 + \lambda\|\beta\|_2^2}_{f(\beta)}$$

where $f(\beta) \overset{def}{=} \|y - x^T\beta\|^2 + \lambda\|\beta\|_2^2$

$$= (y - x^T\beta)^T(y - x^T\beta) + \lambda\beta^T\beta$$

$$= y^T y - y^T X^T\beta - \beta^T X y + \beta^T X X^T\beta$$

$$+ \lambda\beta^T\beta$$

$$= \beta^T (XX^T + \lambda I)\beta - 2y^T X^T \beta + y^T y$$