Higher order derivatives

Consider $f: \mathbb{R}^n \to \mathbb{R}$, assume it is differentiable, so all partial derivatives $\frac{\partial f}{\partial x_i}: \mathbb{R}^n \to \mathbb{R}$ exist.

If this function is differentiable, we can take its derivative:

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

These are called second order partial derivatives.

⚠ in general, we cannot change the order of derivatives.
$$\frac{\partial^2 f}{\partial x_i \partial x_j} \neq \frac{\partial^2 f}{\partial x_j \partial x_i}$$

Example: $f(x,y) = \dfrac{x \cdot y^3}{x^2+y^2}$

$$\nabla f(x,y) = \left(\frac{y^3(y^2-x^2)}{(x^2+y^2)^2}, \frac{xy^2(3x^2+y^2)}{(x^2+y^2)^2} \right)$$

Have: $\frac{\partial f}{\partial x}(0,y) = y \quad \forall y$, $\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = 1$

$\neq$

$\frac{\partial f}{\partial y}(x,0) = 0 \quad \forall x$, $\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = 0$

Def: We say that $f: \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable if all partial derivatives exist and are continuous.

We say that f is twice continuously differentiable if f is continuously differentiable and all its partial derivatives $\frac{\partial f}{\partial x_i}$ are again continuously differentiable.

Analogously: k times continuous differentiable

Notation: $\mathcal{C}^k(\mathbb{R}^n, \mathbb{R}^m) = \{ f: \mathbb{R}^n \to \mathbb{R}^m \mid k \text{ times cont. differentiable} \}$

$$\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R}^m) = \{ f: \mathbb{R}^n \to \mathbb{R}^m \mid \infty \text{ often cont. diff.} \}$$

Theorem (Schwartz): Assume that f is twice continuously differentiable. Then we can exchange the order in which we take partial derivatives: $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$

Analogously: k times cont. diff. $\Rightarrow$ can exchange order of first k partial derivatives.

⚠ Caution about derivatives

$f: \mathbb{R}^n \to \mathbb{R}$ ← function

$\nabla f: \mathbb{R}^n \to \mathbb{R}^n$ ← first derivative : n partial derivatives. $\frac{\partial f}{\partial x_i}$

$Hf: \mathbb{R}^n \to \mathbb{R}^{n \times n}$ ← second derivative: n^2 "partial derivatives" $\frac{\partial^2 f}{\partial x_i \partial x_j}$

Def: Hessian matrix

$f: \mathbb{R}^n \to \mathbb{R}$, then we define the Hessian of f at point x by,

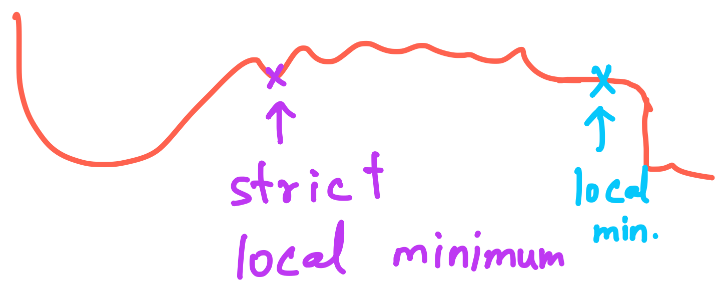
$$(Hf)_{ij}(x) := \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad i,j = 1,2,\dots n$$

Minima / Maxima

Def: $f : \mathbb{R}^n \to \mathbb{R}$ differentiable. If $\nabla f(x) = 0$ then we call x a critical point.

f has a local minimum at x_0 if there exists $\varepsilon > 0$, such that $\forall x \in B_\varepsilon(x_0) : f(x) \geq f(x_0)$

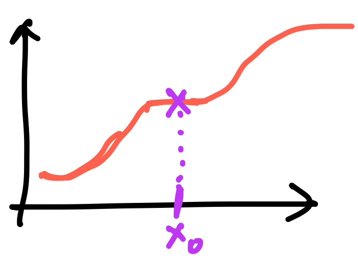
f has a strict local minimum at x_0 $\exists \varepsilon > 0$ such that $\forall x \in B_\varepsilon(x_0) : f(x) > f(x_0)$

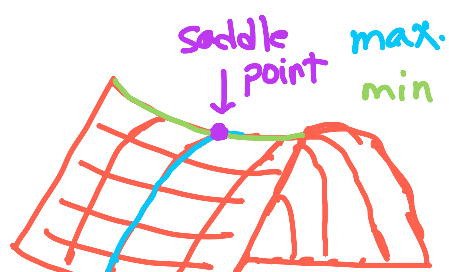


f has a local maximum (resp, a strict local max)

$$\forall x \in B_\varepsilon(x_0) : \quad f(x) \underset{<}{\leq} f(x_0).$$

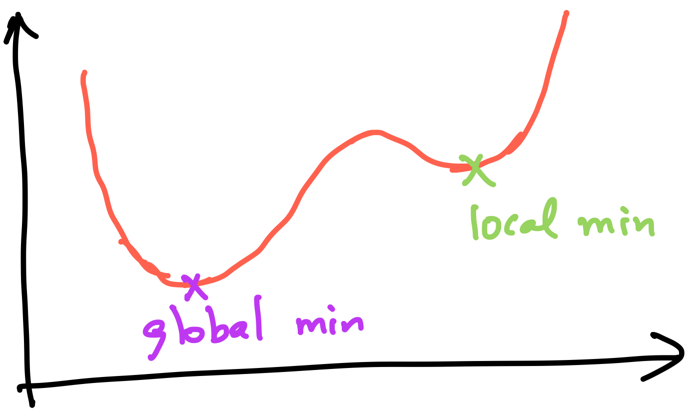
If f is diff. and x_0 is a critical point that is neither a local min. / local max. we call it a saddle point.





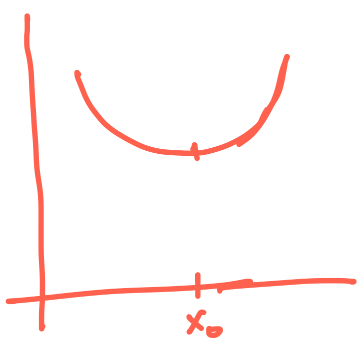
f has a global minimum at x_0 if

$$\forall x: \ f(x) \geq f(x_0)$$



How can we identify which type of point we have?

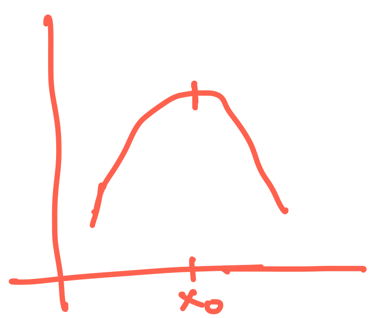
Intuition in $\mathbb{R}$:



local min.

$f'(x) = 0$

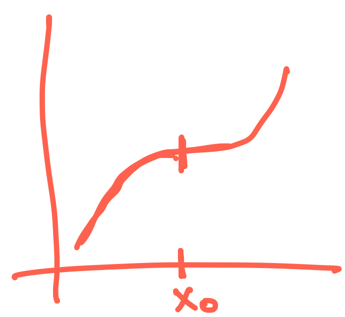
$f''(x) > 0$



local max.

$f'(x) = 0$

$f''(x) < 0$



saddle point

$f'(x) = 0$

$f''(x) = 0$

Theorem: $f: \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(\mathbb{R}^n)$. Assume that x_0 is a critical point, i.e. $\nabla f(x_0) = 0$.

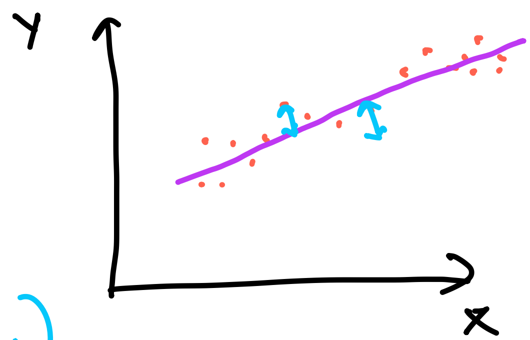
Then:

(i) If x_0 is a local minimum (maximum), then then the Hessian $Hf(x_0)$ is positive semi definite (negative semi definite).

(ii) If $Hf(x_0)$ is positive definite (negative definite), then x_0 is a strict local min (max). If $Hf(x_0)$ is indefinite then x_0 is a saddle point.

Matrix / Vector Calculus

Example: Linear least squares



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

pred $\hat{y}(w) = Aw$ ← weight vector (params we want to find)

prediction, input data

$$f(w) = \|y - \hat{y}(w)\|_2^2 = \|y - Aw\|_2^2$$

how good pred. is with param w.

Want to minimize $f(w)$. Need to look at $\nabla f: \mathbb{R}^n \to \mathbb{R}^n$.

Compute gradient:

$$f\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \sum_{j=1}^{n} \Big(y_j - \overbrace{\sum_{k=1}^{n} a_{jk} w_k}^{(Aw)_j} \Big)^2$$

$$\frac{\partial f}{\partial w_i} = \sum_{j=1}^{n} (-a_{ji}) \cdot 2 \cdot \Big(y_j - \underbrace{\sum_{k=1}^{n} a_{jk} \cdot w_k}_{(Aw)_j} \Big)$$

$$\underbrace{-2 \cdot \sum_{j=1}^{n} a_{ij} \cdot (y - Aw)_j}_{(A^T(y-Aw))_i}$$

$$\nabla f(w) = -2A^T(y - Aw)$$

Intuition: "syntax" close to 1-dim case:

$$f(w) = (y - aw)^2$$

$$f'(w) = -a(y-aw)\cdot 2 = -2a(y-aw)$$

Matrix-vector calculus: lookup table ("matrix cookbook") for gradients of many important functions:

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

- $f(x) = a^T x \quad (a \in \mathbb{R}^n)$
  $= \langle a, x \rangle$
  $$\frac{\partial f}{\partial x} = a \in \mathbb{R}^n$$

- $f(x) = x^T A x \Rightarrow \dfrac{\partial f}{\partial x} = (A + A^T)x \in \mathbb{R}^n$

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

- $f(X) = \underbrace{a^T}_{1\times n} \overbrace{X}^{\mathbb{R}^{n\times m}} \underbrace{b}_{m\times 1} \Rightarrow \dfrac{\partial f}{\partial X} = \underbrace{\underbrace{a}_{n\times 1} \cdot \underbrace{b^T}_{1\times m}}_{n\times m} \in \mathbb{R}^{n\times m}$

- $$f(X) = \underbrace{a^T}_{1\times m} \underbrace{X^T}_{m\times n} \underbrace{C}_{n\times n} \underbrace{X}_{n\times m} \underbrace{b}_{m\times 1}$$

$$\frac{\partial f}{\partial X} = C^T X a b + C X b a^T$$

- $$f(X) = tr(X) \rightarrow \text{Trace}$$

$$\frac{\partial f}{\partial X} = I$$

- $$f(x) = tr(AX) \Rightarrow \frac{\partial f}{\partial x} = A$$

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

- $$f(x) = det(X) \rightarrow \text{Determinant}$$

$$\frac{\partial f}{\partial x} = det(x) \cdot (x^T)^{-1}$$

$$\frac{\partial \, det}{\partial x_{sr}} = det(x) \cdot (x^{-1})_{rs}$$

$$\boxed{f : \mathbb{R}^{n \times m} \longrightarrow \mathbb{R}^{n \times m}}$$ inverse.

$$f(A) = A^{-1}, \qquad f_{ij} := (A^{-1})_{ij}.$$

$$\frac{\partial f_{ij}}{\partial a_{uv}} = -(a_{iu})^{-1} (a_{vj})^{-1}$$