COMPACTNESS & EXTREMA

12

§1. Extrema

$$\vec{a} \in \underset{\substack{\cap \ \text{subset} \\ \mathbb{R}^n}}{S} \xrightarrow{f} \mathbb{R}$$

function of n variables $(x_1, \dots, x_n)$

Definition: (i) f has a (global/absolute) $\overset{\text{minimum}}{\text{maximum}}$ at $\vec{a}$ if

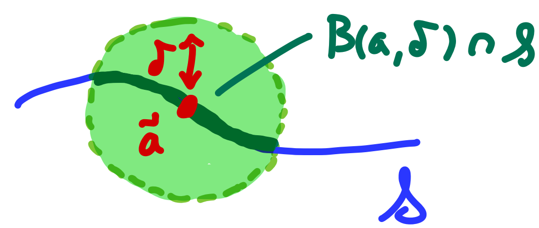
$$f(\vec{a}) \overset{\leq}{\geq} f(\vec{x}) \quad (\forall \vec{x} \in S).$$

(Then $f(\vec{a})$ is called the $\overset{\text{minimum}}{\text{maximum}}$ value)

(ii) f has a relative/local $\overset{\text{minimum}}{\text{maximum}}$ at $\vec{a}$ if $f|_{B(\vec{a},\delta)\cap S}$ has a $\overset{\text{minimum}}{\text{maximum}}$ at $\vec{a}$ for $\delta > 0$ suff. small.

(iii) an extremum (local or global) of f means a maximum or minimum (local or global) of f.

restriction of f to the intersection of S with a small ball



§1. Extrema

$$\vec{a} \in S \xrightarrow{f} \mathbb{R}$$

$S \subset \mathbb{R}^n$ (subset)

function of n variables $(x_1, \dots, x_n)$

Definition: (i) f has a global/absolute) maximum (minimum) at $\vec{a}$ if

$$f(\vec{a}) \geq f(\vec{x}) \quad (\leq) \quad (\forall \vec{x} \in S).$$

(Then $f(\vec{a})$ is called the maximum (minimum) value)

(ii) f has a relative/local maximum (minimum) at $\vec{a}$ if $f\big|_{B(\vec{a},\delta)\cap S}$ has a maximum (minimum) at $\vec{a}$ for $\delta > 0$ suff. small.

(iii) an extremum (local or global) of f means a maximum or minimum (local or global) of f.

Theorem 1: Assume $\vec{a} \in \operatorname{int}(S)$. If f is differentiable at $\vec{a}$ and has a local extremum at $\vec{a}$, then $\vec{\nabla} f(\vec{a}) = \vec{0}$, i.e. $\vec{a}$ is a stationary point of f.

Shifrin calls these critical points, but that terminology should be more inclusive. Maybe "differentiable critical points"?

§1. Extrema

$$\vec{a} \in \underset{\substack{\cap \\ \mathbb{R}^n}}{S} \xrightarrow{f} \mathbb{R}$$

$S \subset \mathbb{R}^n$ subset

function of n variables $(x_1, \dots, x_n)$

Definition: (i) f has a global/absolute) maximum (minimum) at $\vec{a}$ if

$$f(\vec{a}) \overset{\leq}{\geq} f(\vec{x}) \quad (\forall \vec{x} \in S).$$

(Then $f(\vec{a})$ is called the maximum (minimum) value)

(ii) f has a relative/local maximum (minimum) at $\vec{a}$ if $f|_{B(\vec{a},\delta)\cap S}$ has a maximum (minimum) at $\vec{a}$ for $\delta > 0$ suff. small.

(iii) an extremum (local or global) of f means a maximum or minimum (local or global) of f.

Theorem 1: Assume $\vec{a} \in \operatorname{int}(S)$. If f is differentiable at $\vec{a}$ and has a local extremum (minimum) at $\vec{a}$, then $\vec{\nabla} f(\vec{a}) = \vec{0}$, i.e. $\vec{a}$ is a stationary point of f.

Proof (surprisingly easy): Consider the directional derivative $D_{\vec{v}} f(\vec{a})$

$$= \lim_{t \to 0^+} \frac{f(\vec{a} + t\vec{v}) - f(\vec{a})}{t} \geq 0$$

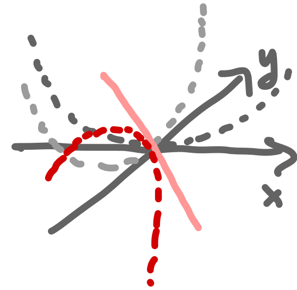
$$= \lim_{t \to 0^-} \frac{f(\vec{a} + t\vec{v}) - f(\vec{a})}{t} \leq 0 \quad \ldots$$

which is zero! So $Df(\vec{a})\,\vec{v} = 0$ for any $\vec{v}$, hence $Df(\vec{a}) = \vec{0}$. ☐

N.B. (1) The tangent plane to the graph of f at a stationary point is horizontal.

(2) The converse of Thm. 1 is FALSE: e.g. $f(x,y) = x^2 - y^2$ at $\vec{0}$. A stationary point that isn't an extremum is a saddle point.

Ex 1/ $f\begin{pmatrix}x\\y\end{pmatrix} = x^2+4xy+y^2$, $\vec{a}=\begin{pmatrix}0\\0\end{pmatrix}$



- xz-cross-section: $f(x,0)=x^2$ ⇝ concave up
- yz-cross-section: $f(0,y)=y^2$ ⇝ also concave up
- also $\underbrace{Df(\vec{0})}_{\text{i.e. } {}^t\vec{\nabla} f(\vec{0})} = (2x+4y,\ 2y+4x)(\vec{0}) = (0,0)$ ⇒ $\vec{a}$ is a stationary point of f

⇝ suggests local (or global) minimum at $\vec{0}$?

Just to be sure, let's differentiate twice in the "SE-direction" $\vec{v}=\begin{pmatrix}1\\-1\end{pmatrix}$:

$$D^2_{\vec{v}} f = D_{\vec{v}}\left((2x+4y,\ 2y+4x)\begin{pmatrix}1\\-1\end{pmatrix}\right)$$
$$= D_{\vec{v}}\{2y-2x\}$$
$$= (-2,\ 2)\begin{pmatrix}1\\-1\end{pmatrix} = -4.$$

⇒ $\vec{a}$ is a stationary point which is neither a local max nor local min, i.e. a saddle point. //

Theorem 1: Assume $\vec{a}\in \operatorname{int}(\mathcal{S})$. If f is differentiable at $\vec{a}$ and has a local extremum (minimum) at $\vec{a}$, then $\vec{\nabla} f(\vec{a})=\vec{0}$, i.e. $\vec{a}$ is a stationary point of f.

Proof (surprisingly easy): Consider the directional derivative $D_{\vec{v}} f(\vec{a})$

$$= \lim_{t\to 0^+} \frac{f(\vec{a}+t\vec{v})-f(\vec{a})}{t} \geq 0$$

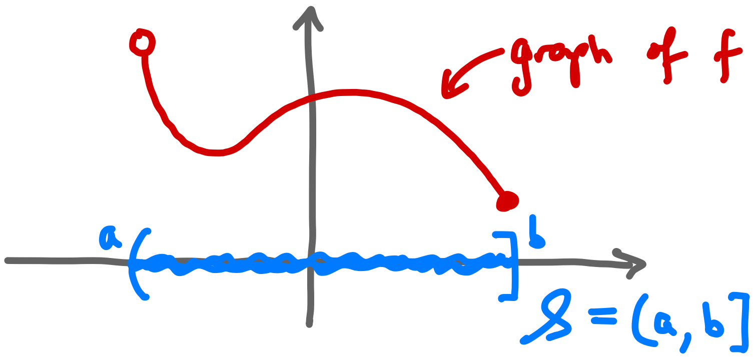
$$= \lim_{t\to 0^-} \frac{f(\vec{a}+t\vec{v})-f(\vec{a})}{t} \leq 0 \ \ldots$$

which is zero! So $Df(\vec{a})\vec{v}=0$ for any $\vec{v}$, hence $Df(\vec{a})=\vec{0}$. □

N.B. (1) The tangent plane to the graph of f at a stationary point is horizontal.

(2) The converse of Thm. 1 is FALSE: e.g. $f(x,y)=x^2-y^2$ at $\vec{0}$. A stationary point that isn't an extremum is a saddle point.

But do any extrema actually occur?



Theorem 1: Assume $\vec{a} \in \operatorname{int}(\mathcal{S})$. If f is differentiable at $\vec{a}$ and has a local extremum at $\vec{a}$, then $\vec{\nabla} f(\vec{a}) = \vec{0}$, i.e. $\vec{a}$ is a stationary point of f.

Definition: A critical point of f is any of the following:

(i) a stationary point of f

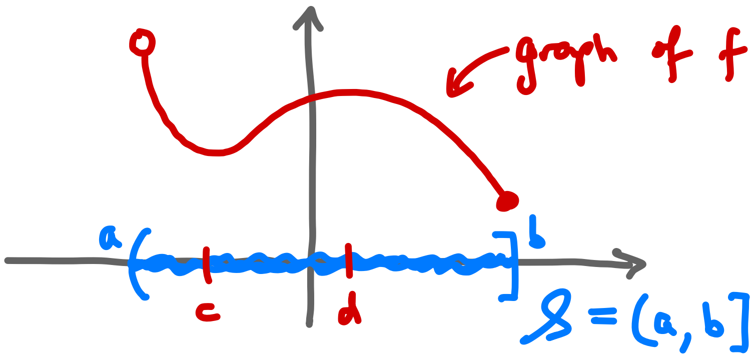
(ii) a point at which f is not differentiable

(iii) a point of the boundary $\partial\mathcal{S}$.

Corollary: Extrema can only occur at critical points.

This follows immediately from Theorem 1, since $\partial\mathcal{S} = \mathcal{S} \setminus \operatorname{int}(\mathcal{S})$ after all.
So if we can identify all the points of types (i), (ii), and (iii), we have all the candidates for extrema.

But do any extrema actually occur?



The function $f : S \to \mathbb{R}$ shown has a local minimum at c, local maximum at d (both stationary pts.), and global minimum at the boundary point b. It has NO global maximum: there is no point $s \in S$ for which $f(s) \geq f(x)$ for all other $x \in S$. The problem is that S isn't closed: no matter how close we take s to a, there is another $x \in S$ which is closer.

On the other hand, if we take $S = \mathbb{R}$ the whole real line, that is closed. But $f(x) = x^2$ has no global maximum value (because for each value there is a bigger one); this time unboundedness of S takes the blame.

Clearly, to have any hope of being guaranteed global maxima & minima, we need to assume that S is closed and bounded.

[N.B.: Another annoying thing is that a manifold like a curve or surface in $\mathbb{R}^3$ is "all boundary": $\partial S = S$. To fix this, one should parametrize S by a subset of $\mathbb{R}^1$ or $\mathbb{R}^2$, as the IFT says we can (at least locally).]

§2. Compactness

Let $S \subset \mathbb{R}^n$ be a subset.

We'll say that it is bounded if it is contained in some ball $B(\vec{0}, R)$.

Definition: S is compact if it is closed and bounded.

Here I recall that a set S is closed if its complement is open, or (equivalently) if it contains its boundary, or (also equivalently) if every convergent sequence of points in S converges to a point of S.

By the induction hypothesis, $\tilde{x}_{k_j}$ then has a convergent subsequence $\tilde{x}_{k_{j_i}} \to \tilde{c} \in [a_1,b_1] \times \cdots \times [a_{n-1}, b_{n-1}]$.

Since $y_{k_{j_i}} \to c_n$, now $\vec{x}_{k_{j_i}} \to \vec{c} = \begin{pmatrix} \tilde{c} \\ c_n \end{pmatrix} \in \mathcal{S}$.

Finally, if $\mathcal{S} \subset \mathbb{R}^n$ is a general compact set, it is bounded hence contained in some compact rectangle $R = [a_1,b_1] \times \cdots \times [a_n, b_n]$. Any sequence $\{\vec{x}_k\} \subset \mathcal{S}$ lies in R, hence has a convergent subsequence $\{\vec{x}_{k_j}\}$ which converges to a point of R. But since $\mathcal{S}$ is closed, the limit of this convergent sequence must lie in $\mathcal{S}$. $\square$

Theorem 2: Any sequence of points in a compact set $\mathcal{S}$ has a convergent subsequence, with limit in $\mathcal{S}$.

Proof: By an exercise from HW #2, we know that this is true for closed intervals $[a,b] \subset \mathbb{R}$. (Subdivide the interval repeatedly.)

For $\mathcal{S}$ a closed rectangle $[a_1,b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$, it follows from the 1-dim'l. case by induction: suppose we have proved it for $n-1$, and let $\{\vec{x}_k\}_{k\geq 1} \subset \mathcal{S}$ be a sequence. Writing $\vec{x}_k = \begin{pmatrix} \tilde{x}_k \\ y_k \end{pmatrix}$ $\begin{matrix} \leftarrow \text{in } [a_1,b_1] \times \cdots \times [a_{n-1}, b_{n-1}] \\ \leftarrow \text{in } [a_n, b_n] \end{matrix}$

we know (by 1-dim'l. case) that $\{y_k\}$ has a convergent subsequence $y_{k_j} \xrightarrow{j \to \infty} c_n \in [a_n, b_n]$.

Theorem 2: Any sequence of points in a compact set S has a convergent subsequence, with limit in S.

Theorem 3: Any continuous function f on a compact set $S \subset \mathbb{R}^n$ has a global maximum and minimum (on S).

Upshot: we have justified the following strategy for maximizing/minimizing a C^0 function f on a compact set $S \subset \mathbb{R}^n$:

$\boxed{\text{STEP 1}}$ maximize/minimize f on ∂S (which is a problem in one less dimension)

$\boxed{\text{STEP 2}}$ evaluate f at stationary points (and any non-differentiable points)

$\boxed{\text{STEP 3}}$ pick the biggest/smallest values from Steps 1 & 2.

Proof of Thm. 3: I claim f is bounded: $\exists\, B>0$ s.t. $|f(\vec{x})| \le B \;\; \forall \vec{x} \in \mathcal{S}$. Otherwise, $\exists$ sequence $\{\vec{x}_k\} \subset \mathcal{S}$ with $|f(\vec{x}_k)| > k$. By Thm. 2, this has a convergent subsequence $\vec{x}_{k_j} \to \vec{a} \in \mathcal{S}$. Since f is continuous, $f(\vec{a}) = \lim_{j\to\infty} f(\vec{x}_{k_j})$; but this limit diverges, a contradiction.

Now $f(\mathcal{S})$ is bounded above, so $M := \sup f(\mathcal{S})$ is defined. Since this is the least upper bound, for every $k \in \mathbb{N}$ there is an $\vec{x}_k \in \mathcal{S}$ such that $f(\vec{x}_k) > M - \frac{1}{k}$. As $\mathcal{S}$ is compact, there is a convergent subsequence $\vec{x}_{k_j} \to \vec{a} \in \mathcal{S}$; and since f is C^0, $f(\vec{a}) = \lim_{j\to\infty} f(\vec{x}_{k_j}) = M$. So f attains a global maximum at $\vec{a}$, and we can argue similarly for a minimum. $\square$

Theorem 2: Any sequence of points in a compact set $\mathcal{S}$ has a convergent subsequence, with limit in $\mathcal{S}$.

Theorem 3: Any continuous function f on a compact set $\mathcal{S} \subset \mathbb{R}^n$ has a global maximum and minimum (on $\mathcal{S}$).

Theorem 4: A continuous function f on a compact set $\mathcal{S}$ is uniformly continuous: for each $\epsilon > 0$, $\exists\, \delta > 0$ s.t. $\|x-y\| < \delta \Rightarrow \|f(x) - f(y)\| < \epsilon$.
$(x, y \in \mathcal{S})$

Please read the proof of this one in Shifrin (Thm. 3.1.4).

§3. Optimization — i.e. finding global max/min

Ex 2/ Optimize

$$f\begin{pmatrix}x\\y\end{pmatrix} = 2x^2+y^2-4x-2y+5$$

on the compact set $S = \left\{ \begin{pmatrix}x\\y\end{pmatrix} \,\middle|\, x^2+\frac{y^2}{2} \le 1 \right\}$.

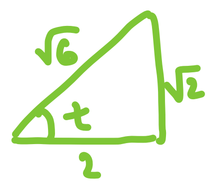
- ∂S = ellipse, parametrized by $\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}\cos t\\ \sqrt{2}\sin t\end{pmatrix}$. Writing $F(t) = f\begin{pmatrix}x(t)\\y(t)\end{pmatrix}$,

$$F'(t) = f_x\begin{pmatrix}x(t)\\y(t)\end{pmatrix}x'(t) + f_y\begin{pmatrix}x(t)\\y(t)\end{pmatrix}y'(t)$$

$$= (4\cos t-4)(-\sin t) + (2\sqrt{2}\sin t-2)(\sqrt{2}\cos t)$$

$$= 4\sin t - 2\sqrt{2}\cos t.$$

Setting $0 = F'(t)$ gives $\tan(t) = \frac{\sqrt{2}}{2}$



$$\Rightarrow x(t) = \cos\left(\arctan\left(\tfrac{\sqrt{2}}{2}\right) + \pi\right) = -\frac{2}{\sqrt{6}}$$

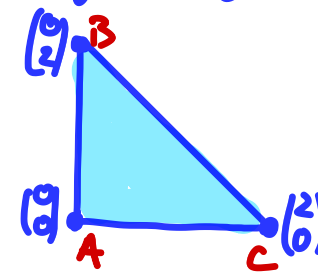
$$y(t) = \sqrt{2}\sin\left(\arctan\left(\tfrac{\sqrt{2}}{2}\right) + \pi\right) = -\frac{2}{\sqrt{6}}$$

$$\Rightarrow f\begin{pmatrix}2/\sqrt{6}\\2/\sqrt{6}\end{pmatrix} = 7-2\sqrt{6}, \quad f\begin{pmatrix}-2/\sqrt{6}\\-2/\sqrt{6}\end{pmatrix} = 7+2\sqrt{6}$$

- Interior: $\vec{0} = \vec{\nabla} f = (4x-4, 2y-2) \Rightarrow$ $\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix}$, and $f\begin{pmatrix}1\\1\end{pmatrix} = 2 < 7-2\sqrt{6}$ so ...

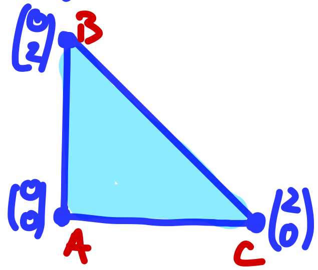
WAIT! $\begin{pmatrix}1\\1\end{pmatrix}$ is not in S! So there is no interior critical point, and the max/min values are $7 \pm 2\sqrt{6}$. //

PROBLEM Optimize $f(x,y) = 8xy - x - y$ on the compact triangular region



PROBLEM Find the minimum distance between the origin $\vec{0}$ and the surface $M = \left\{ \begin{pmatrix}x\\y\\z\end{pmatrix} \in \mathbb{R}^3 \,\middle|\, z^2 = x^2y+4 \right\}$ in $\mathbb{R}^3$. (Warning: not compact!)

PROBLEM Optimize $f(x,y) = 8xy - x - y$ on the compact triangular region



- $\overline{AC}$: $y=0$. $f(x,0) = -x$ < max at $A : 0$; min at $C : -2$
- $\overline{AB}$: $x=0$. $f(0,y) = -y$ < max at $A : 0$; min at $B : -2$
- $\overline{BC}$: $f(2-t,t) = 16t - 8t^2 - 2$

  $\Rightarrow 0 = \frac{d}{dt} f(2-t,t) = 16 - 16t$ gives $t = 1$, and $f(1,1) = 6$.
- interior : $\vec{0} = \vec{\nabla} f = (8y-1, 8x-1)$

  $\rightarrow (x,y) = \left(\frac{1}{8}, \frac{1}{8}\right)$, at which $f = \frac{-1}{8}$.

  (it's a saddle point, but we don't care!)

Conclude that overall < max $= 6$, at $\begin{pmatrix}1\\1\end{pmatrix}$; min $= -2$, at B & C.

PROBLEM Find the minimum distance between the origin $\vec{O}$ and the surface $M = \left\{ \begin{pmatrix}x\\y\\z\end{pmatrix} \in \mathbb{R}^3 \,\middle|\, z^2 = x^2y + 4 \right\}$ in $\mathbb{R}^3$.

(Warning: not compact!)

Let $\vec{P} = \begin{pmatrix}x\\y\\z\end{pmatrix}$ be any point on M.

$\|\vec{P} - \vec{O}\|^2 = x^2 + y^2 + z^2$ is easier to work with.

So consider $f\begin{pmatrix}x\\y\end{pmatrix} = x^2 + y^2 + (x^2y + 4)$.

Start by finding the critical points:

$0 = \vec{\nabla} f = (2x + 2xy, 2y + x^2) \Rightarrow y = -\frac{x^2}{2}$

$\Rightarrow 0 = 2x + 2xy = 2x - x^3 \Rightarrow x = 0, \pm\sqrt{2}$

$\rightarrow \begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sqrt{2}\\-1\end{pmatrix}, \begin{pmatrix}-\sqrt{2}\\-1\end{pmatrix}$.

Since x & y don't live in a bounded set, take S to be a large disk about $\begin{pmatrix}0\\0\end{pmatrix}$ and note that f is at least the radius of this disk on the boundary. So the minimum must occur on the interior. Evaluating $f\begin{pmatrix}0\\0\end{pmatrix} = 4$, $f\begin{pmatrix}\sqrt{2}\\-1\end{pmatrix} = 5$, and $f\begin{pmatrix}-\sqrt{2}\\-1\end{pmatrix} = 5$ we see that the minimum distance is $\sqrt{4} = 2$.

§4. An application to matrices

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map, with $m \times n$ matrix A.

The $(n-1)$-sphere $S^{n-1} := \{\hat{x} \in \mathbb{R}^n \mid \|\hat{x}\| = 1\}$ is compact, and so the following definition makes sense:

$$\|A\| \ (\text{or } \|T\|) := \max_{\|\hat{x}\|=1} \|A\hat{x}\|.$$

For $\vec{x} \neq \vec{0}$ not necessarily of unit length,

$$\frac{1}{\|\vec{x}\|}\|A\vec{x}\| = \left\|A\left(\frac{\vec{x}}{\|\vec{x}\|}\right)\right\| \leq \|A\| \text{ gives}$$

$$\|A\vec{x}\| \leq \|A\|\,\|\vec{x}\|.$$

Since $\|(A+B)\hat{x}\| \leq \|A\hat{x}\| + \|B\hat{x}\|$ by the Δ-inequality, taking $\max_{\|\hat{x}\|=1}$ on both sides yields

$$\|A+B\| \leq \|A\| + \|B\|.$$

Similarly, if B is a $n \times p$ matrix,

$$\max_{\|\hat{x}\|=1}\|AB\hat{x}\| = \max_{\|\hat{x}\|=1}\left\|A\frac{B\hat{x}}{\|B\hat{x}\|}\right\|\,\|B\hat{x}\|$$

$$\leq \left(\max_{\|\hat{y}\|=1}\|A\hat{y}\|\right)\left(\max_{\|\hat{x}\|=1}\|B\hat{x}\|\right) \text{ yields}$$

$$\|AB\| \leq \|A\|\,\|B\|.$$

(We clearly also have $\|cA\| = |c|\,\|A\|$.)

This is an example of a matrix norm called the <u>operator norm</u>, since it tells how much A as an "operator" can dilate the length of a vector.

For instance, writing $\vec{a}_j$ for columns of A,

$$\|A\|_{(2)}^2 = \sum_{i,j} a_{ij}^2 = \sum_j \|\vec{a}_j\|^2 = \sum_j \|A\hat{e}_j\|^2$$

$$\le \sum_j \|A\|^2 \underbrace{\|\hat{e}_j\|^2}_{1} = n\|A\|^2$$

by $\|A\vec{x}\| \le \|A\|\|\vec{x}\|$

yields

$$\|A\|_{(2)} \le \sqrt{n}\,\|A\|,$$

while (writing $\vec{A}_i$ for the rows of A)

$$\|A\vec{x}\|^2 = \left\|\begin{matrix}\vec{A}_1\cdot\vec{x}\\ \vdots \\ \vec{A}_m\cdot\vec{x}\end{matrix}\right\|^2 = \sum_i (\vec{A}_i\cdot\vec{x})^2$$

Cauchy-Schwarz

$$\le \sum_i \|\vec{A}_i\|^2\|\vec{x}\|^2 = \left(\sum_{i,j} a_{ij}^2\right)\|\vec{x}\|^2$$

yields $\|A\vec{x}\| \le \|A\|_{(2)}\|\vec{x}\|$ hence (by taking max over $\|\vec{x}\|=1$)

$$\|A\| \le \|A\|_{(2)}.$$

This is used later in the book.

$$\|A+B\| \le \|A\| + \|B\|.$$

Similarly, if B is a $n\times p$ matrix,

$$\max_{\|\hat{x}\|=1}\|AB\hat{x}\| = \max_{\|\hat{x}\|=1}\left\|A\frac{B\hat{x}}{\|B\hat{x}\|}\right\|\|B\hat{x}\|$$

$$\le\left(\max_{\|\hat{y}\|=1}\|A\hat{y}\|\right)\left(\max_{\|\hat{x}\|=1}\|B\hat{x}\|\right)$$ yields

$$\|AB\| \le \|A\|\|B\|.$$

(We clearly also have $\|cA\| = |c|\|A\|$.)

This is an example of a matrix norm called the operator norm, since it tells how much A as an "operator" can dilate the length of a vector.

Another such example is

$$\|A\|_{(2)} := \sqrt{\sum_{i,j} a_{ij}^2}$$

and it is of interest to compare them.