1. Review of Multi-variable Calculus

Throughout this course we will be working with the vector space $\mathbb{R}^n$. For this reason we begin with a brief review of its metric space properties

Definition 1.1 (Vector Norm)**.** A function $\nu : \mathbb{R}^n \to \mathbb{R}$ is a *vector norm* on $\mathbb{R}^n$ if the following three properties hold.

i. **(Positivity)**: $\nu(x) \geq 0 \ \forall \ x \in \mathbb{R}^n$ with equality iff $x = 0$.
ii. **(Homogeneity)**: $\nu(\alpha x) = |\alpha|\nu(x) \ \forall \ x \in \mathbb{R}^n \ \alpha \in \mathbb{R}$
iii. **(Triangle inequality)**: $\nu(x + y) \leq \nu(x) + \nu(y) \ \forall \ x, y \in \mathbb{R}^n$

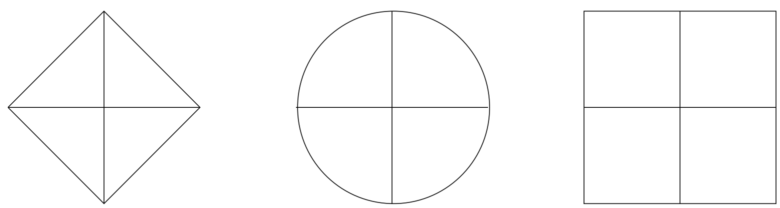
We usually denote $\nu(x)$ by $\|x\|$. Norms are convex functions.

Example: l_p norms

$$\begin{array}{rcl} \|x\|_p & := & (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}, \quad 1 \leq p < \infty \\ \|x\|_\infty & = & \max_{i=1,\ldots,n} |x_i| \end{array}$$

– $P = 1, 2, \infty$ are most important cases

$\|x\|_1 = 1$ $\qquad$ $\|x\|_2 = 1$ $\qquad$ $\|x\|_\infty = 1$



– The unit ball of a norm is a convex set.

1.1. Equivalence of Norms.

All norms on $\mathbb{R}^n$ are comparable, meaning that for any norms $\|\cdot\|_p$ and $\|\cdot\|_q$, there exist constants $\alpha_{p,q}$ and $\beta_{p,q}$ satisfying

$$\alpha_{p,q}\|x\|_q \leq \|x\|_p \leq \beta_{p,q}\|x\|_q \qquad \text{for all } x \in \mathbb{R}^n.$$

Here are some values of the constants $\alpha_{p,q}$ and $\beta_{p,q}$.

$\alpha_{p,q}$

$p \backslash q$	1	2	3
1	1	1	1
2	$n^{-\frac{1}{2}}$	1	1
3	n^{-1}	$n^{-\frac{1}{2}}$	1

$\beta_{p,q}$

$p \backslash q$	1	2	3
1	1	$n^{\frac{1}{2}}$	n
2	1	1	$n^{\frac{1}{2}}$
3	1	1	1

1.2. **Continuity and the Weierstrass Theorem.**

– A mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *continuous at the point* $\overline{x}$ if

$$\lim_{\|x-\overline{x}\|\to 0} \|F(x) - F(\overline{x})\| = 0.$$

F is said to be *continuous on a set* $D \subset \mathbb{R}^n$ if F is continuous at every point of D.

– A subset $D \subset \mathbb{R}^n$ is said to be *open* if for every $x \in D$ there exists $\epsilon > 0$ such that $\mathbb{B}_\epsilon(x) \subset D$ where

$$\mathbb{B}_\epsilon(x) = \{y \in \mathbb{R}^n : \|y - x\| < \epsilon\}.$$

– A subset $D \subset \mathbb{R}^n$ is said to be *closed* if every point x satisfying

$$\mathbb{B}_\epsilon(x) \cap D \neq \emptyset$$

for all $\epsilon > 0$, must be a point in D.

– A subset $D \subset \mathbb{R}^n$ is said to be *bounded* if there exists $m > 0$ such that

$$\|x\| \leq m \text{ for all } x \in D.$$

(Notice: the choice of the norm is irrelevant in the definition.)

– A subset $D \subset \mathbb{R}^n$ is said to be *compact*, if it is closed and bounded.

– A point $x \in \mathbb{R}^n$ is said to be a *cluster point* of the set $D \subset \mathbb{R}^n$ if

$$(\mathbb{B}_\epsilon(x) \setminus \{x\}) \cap D \neq \emptyset$$

for every $\epsilon > 0$.

For example, for the set $D := (0, 1] \cup \{2\}$, the set of cluster points is the set $[0, 1]$.

Theorem 1.1 (Weierstrass Compactness Theorem)**.** *A set $D \subset \mathbb{R}^n$ is compact if and only if every infinite subset of D has a cluster point in D.*

Next, we recall the notions of the supremum and infimum of a function. To this end, consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and a set $D \subset \mathbb{R}^n$. Define the set of upper bounds

$$U = \{r \in \mathbb{R} : f(x) \leq r \text{ for all } x \in D\}.$$

One can prove that U is a closed subinterval of the real line, namely we may write $U = [\alpha, +\infty)$ for some α. (Note α can be finite or infinite.) The value α is called the *supremum of f on D*. Intuitively this quantity is the "least upper bound" of f on D. Note that for any $r > \alpha$, there cannot exist a point $x \in D$ satisfying $r = f(x)$ (Why?). On the other hand, when there exists some point $\bar{x}$ in D satisfying $\alpha = f(\bar{x})$, we call α the *maximal value* of f on D, and we say that the *maximum of f on D is attained* at $\bar{x}$. Moreover, this point $\bar{x}$ is called a *maximizer of f on D*.

The definition of the *infimum of f on D* as the "greatest lower bound" is entirely analogous. Namely the set of lower bounds

$$L = \{r \in \mathbb{R} : f(x) \geq r \text{ for all } x \in D\}$$

can be shown to be an interval $(-\infty, \beta]$ for some β. This value β is called the *infimum of f on D*. Minimal values, minimizers, and attainment of the minimum are defined analogously. The following theorem, which we will use extensively, establishes a connection between continuous functions on compact sets and attainment of the minimum and the maximum.

Theorem 1.2 (Weierstrass Extreme Value Theorem)**.** *Every continuous function on a compact set attains its extreme values (maximum and minimum) on that set.*

1.3. **Dual Norms.** Let $\|\cdot\|$ be a given norm on $\mathbb{R}^n$ with associated closed unit ball $\mathbb{B}$. For each $x \in \mathbb{R}^n$ define

$$\|x\|_* := \max_{y\in\mathbb{R}^n} \{x^T y : \|y\| \le 1\}.$$

Since the transformation $y \mapsto x^T y$ is continuous (in fact, linear) and $\mathbb{B}$ is compact (can you prove this?), Weierstrass's Theorem says that the maximum in the definition of $\|x\|_*$ is attained. Thus, in particular, the function $x \to \|x\|_*$ is well defined and finite-valued. Indeed, the mapping defines a norm on $\mathbb{R}^n$. This norm $\|\cdot\|_*$ is said to be the *norm dual* to the norm $\|\cdot\|$. Thus, every norm has a norm dual to it.

We now show that the mapping $x \mapsto \|x\|_*$ is a norm.

(a) It is easily seen that $\|x\|_* = 0$ if $x = 0$. On the other hand, if $x \neq 0$, then

$$\|x\|_* = \max\{x^T y : \|y\| \le 1\} \ge x^T \left(\frac{x}{\|x\|}\right) = \frac{\|x\|_2^2}{\|x\|} > 0.$$

(b) From part (a), we have $\|0 \cdot x\|_* = 0 = 0 \cdot \|x\|_*$. Next suppose $\alpha \in \mathbb{R}$ with $\alpha \neq 0$. Then

$$\begin{aligned} \|\alpha x\|_* &= \max\{x^T(\alpha y) : \|y\| \le 1\}, \quad (\text{set } z := \alpha y) \\ &= \max\left\{x^T z : 1 \ge \left\|\tfrac{z}{\alpha}\right\| = \tfrac{1}{|\alpha|}\|z\| = \left\|\tfrac{z}{|\alpha|}\right\|\right\}, \quad \left(\text{set } w := \tfrac{z}{|\alpha|}\right) \\ &= \max\{x^T(|\alpha| w) : 1 \ge \|w\|\} \\ &= |\alpha|\, \|x\|_*. \end{aligned}$$

In order to establish the triangle inequality, we make use of the following elementary, but very useful, fact.

FACT: For a function $f : \mathbb{R}^n \to \mathbb{R}$ and sets $C \subset D \subset \mathbb{R}^n$, it holds:

$$\sup_{x\in C} f(x) \le \sup_{x\in D} f(x).$$

That is, the supremum over a larger set must be larger. Similarly, the infimum over a larger set must be smaller.

(c)
$$\begin{aligned} \|x+z\|_* &= \max\{x^T y + z^T y : \|y\| \le 1\} \\ &= \max\left\{x^T y_1 + z^T y_2 : \begin{array}{l} \|y_1\| \le 1 \\ \|y_2\| \le 1 \end{array}, y_1 = y_2\right\} \end{aligned}$$
(max over a larger set)
$$\begin{aligned} &= \le \max\{x^T y_1 + z^T y_2 : \|y_1\| \le 1, \|y_2\| \le 1\} \\ &= \|x\|_* + \|z\|_* \end{aligned}$$

FACTS:

(i) $x^T y \le \|x\|\, \|y\|_*$ (apply definition)

(ii) $(\|x\|_p)_* = \|x\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$, $1 \le p \le \infty$

(iii) Hölder's Inequality: $|x^T y| \le \|x\|_p \|y\|_q$

$$\frac{1}{p} + \frac{1}{q} = 1$$

(iv) Cauchy-Schwartz Inequality:

$$|x^T y| \le \|x\|_2 \|y\|_2$$

1.4. **Operator Norms.** For the a matrix $A \in \mathbb{R}^{m\times n}$, the *p-operator norm* is given by

$$\|A\|_p := \max\{\|Ax\|_p : \|x\|_p \le 1\}$$

EXAMPLE:
$$\begin{aligned} \|A\|_2 &= \max\{\|Ax\|_2 : \|x\|_2 \le 1\} \\ \|A\|_\infty &= \max\{\|Ax\|_\infty : \|x\|_\infty \le 1\} \\ &= \max_{1\le i\le m} \textstyle\sum_{j=1}^n |a_{ij}|, \text{max row form} \\ \|A\|_1 &= \max\{\|Ax\|_1 : \|x\|_1 \le 1\} \\ &= \max_{1\le j\le n} \textstyle\sum_{i=1}^m |a_{ij}|, \text{max column sum} \end{aligned}$$

FACT: $\|Ax\|_p \le \|A\|_p \|x\|_p$.

(a) $\|A\| \ge 0$ with equality iff $A \equiv 0$.

(b)
$$\begin{aligned} \|\alpha A\| &= \max\{\|\alpha Ax\| : \|x\| \le 1\} \\ &= \max\{|\alpha|\,\|Ax\| : \|x\| \le 1\} = |\alpha|\,\|A\| \end{aligned}$$

(c)
$$\begin{aligned} \|A+B\| &= \max\{\|Ax+Bx\| : \|x\| \le 1\} \le \max\{\|Ax\| + \|Bx\| : \|x\| \le 1\} \\ &= \max\{\|Ax_1\| + \|Bx_2\| : x_1 = x_2, \|x_1\| \le 1, \|x_2\| \le 1\} \\ &\le \max\{\|Ax_1\| + \|Bx_2\| : \|x_1\| \le 1, \|x_2\| \le 1\} \\ &= \|A\| + \|B\| \end{aligned}$$

1.4.1. *Condition number.* The *condition number* of a matrix $A \in \mathbb{R}^{n\times n}$ is defined by

$$\kappa(A) := \begin{cases} \|A\|\,\|A^{-1}\| & \text{if } A^{-1} \text{ exists} \\ \infty & \text{otherwise} \end{cases}$$

FACT: [Error estimates in the solution of linear equations] If $Ax_1 = b$ and $Ax_2 = b + e$, then

$$\frac{\|x_1 - x_2\|}{\|x_1\|} \le \kappa(A)\frac{\|e\|}{\|b\|}$$

Proof. $\|b\| = \|Ax_1\| \le \|A\|\,\|x_1\| \Rightarrow \frac{1}{\|x_1\|} \le \frac{\|A\|}{\|b\|}$, so

$$\frac{\|x_1 - x_2\|}{\|x_1\|} \le \frac{\|A\|}{\|b\|}\|A^{-1}(A(x_1 - x_2)\| \le \|A\|\,\|A^{-1}\|\frac{1}{\|b\|}\|Ax_1 - Ax_2\|$$

□

1.5. **The Frobenius Norm.** There is one further norm for matrices that is very useful. It is called the Frobenius norm.

Observe that we can identify $\mathbb{R}^{m\times n}$ with $\mathbb{R}^{(mn)}$ by simply stacking the columns of a matrix one on top of the other to create a very long vector in $\mathbb{R}^{(mn)}$. The Frobenius norm is then the 2-norm of this vector. It can be verified that

$$\|A\|_F^2 = \operatorname{tr} A^2.$$

1.6. **Review of Differentiation.**

1) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and let $x, d \in \mathbb{R}^n$. If the limit

$$\lim_{t\downarrow 0} \frac{F(x+td) - F(x)}{t} =: F'(x; d)$$

exists, it is called the *directional derivative* of F at x in the direction d. If this limit exists for all $d \in \mathbb{R}^n$ and is linear in the d argument, meaning

$$F'(x; \alpha d_1 + \beta d_2) = \alpha F'(x; d_1) + \beta F'(x; d_2),$$

then F is said to be *Gâteaux differentiable* at x.

2) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and let $x \in \mathbb{R}^n$. If there exists a matrix $J \in \mathbb{R}^{m\times n}$ such that

$$\lim_{\|y-x\|\to 0} \frac{\|F(y) - (F(x) + J(y-x))\|}{\|y-x\|} = 0,$$

then F is said to be *Fréchet differentiable* at x and J is said to be its *Fréchet derivative*. We denote J by $J = F'(x)$.

FACTS:

(i) If $F'(x)$ exists, it is unique.

(ii) If $F'(x)$ exists, then $F'(x; d)$ exists for all d and

$$F'(x; d) = F'(x)d.$$

(iii) If $F'(x)$ exists, then F is continuous at x.

(iv) (Matrix Representation)
Suppose $F'(x)$ exists for all x near $\overline{x}$ and that the mapping $x \mapsto F'(x)$ is continuous at $\overline{x}$, meaning as usual

$$\lim_{\|x-\overline{x}\|\to 0} \|F'(x) - F'(\overline{x})\| = 0,$$

then the partial derivatives $\partial F_i/\partial x_j$ exist for each $i = 1, \dots, m$, $j = 1, \dots, n$ and with respect to the standard basis the linear operator $F'(\overline{x})$ has the representation

$$\nabla F(\overline{x}) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial F_n}{\partial x_1} & \cdots & \cdots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}^T = \left[\frac{\partial F_i}{\partial x_j}\right]^T$$

where each partial derivative is evaluated at $\overline{x} = (\overline{x}_1, \dots, \overline{x}_n)^T$. This matrix is called the *Jacobian matrix* for F at $\overline{x}$.

Notation: For a function $f : \mathbb{R}^n \to \mathbb{R}$ and the vector $f'(x) := \left[\frac{\partial f_1}{\partial x_1}, \dots, \frac{\partial f_*}{\partial x_n}\right]$ we write $\nabla f(x) = f'(x)^T$.

(v) If $F : \mathbb{R}^n \to \mathbb{R}^m$ has continuous partials $\partial F_i / \partial x_i$ on an open set $D \subset \mathbb{R}^n$, then F is differentiable on D. Moreover, in the standard basis the matrix representation for $F'(x)$ is the Jacobian of F at x.

(vi) (Chain Rule) Let $F : A \subset \mathbb{R}^m \to \mathbb{R}^k$ be differentiable on the open set A and let $G : B \subset \mathbb{R}^k \to \mathbb{R}^n$ be differentiable on the open set B. If $F(A) \subset B$, then the composite function $G \circ F$ is differentiable on A and

$$(G \circ F)'(x_0) = G'(F(x_0)) \circ F'(x_0).$$

Remarks: Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable. If $L(\mathbb{R}^n, \mathbb{R}^m)$ denotes the set of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$, then

$$F' : \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m).$$

(v) <u>The Mean Value Theorem</u>:

(a) If $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then for every $x, y \in \mathbb{R}$ there exists z between x and y such that

$$f(y) = f(x) + f'(z)(y - x).$$

(b) If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then for every $x, y \in \mathbb{R}$ there is a $z \in [x, y]$ such that

$$f(y) = f(x) + \nabla f(z)^T (y - x).$$

(c) If $F : \mathbb{R}^n \to \mathbb{R}^m$ continuously differentiable, then for every $x, y \in \mathbb{R}$

$$\|F(y) - F(x)\| \leq [\sup_{z \in [x,y]} \|F'(z)\|] \|x - y\|.$$

Proof of (b): Set $\varphi(t) = f(x + t(y - x))$. Then, by the chain rule, $\varphi'(t) = \nabla f(x + t(y - x))^T (y - x)$ so that φ is differentiable. Moreover, $\varphi : \mathbb{R} \to \mathbb{R}$. Thus, by (a), there exists $\bar{t} \in (0, 1)$ such that

$$\varphi(1) = \varphi(0) + \varphi'(\bar{t})(1 - 0),$$

or equivalently,

$$f(y) = f(x) + \nabla f(z)^T (y - x)$$

where $z = x + \bar{t}(y - x)$. ∎

1.6.1. *The Implicit Function Theorem.* Let $F : \mathbb{R}^{n+m} \to \mathbb{R}^n$ be continuously differentiable on an open set $E \subset \mathbb{R}^{n+m}$. Further suppose that there is a point $(\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ at which $F(\bar{x}, \bar{y}) = 0$. If $\nabla_x F(\bar{x}, \bar{y})$ is invertable, then there exist open sets $U \subset \mathbb{R}^{n+m}$ and $W \subset \mathbb{R}^m$, with $(\bar{x}, \bar{y}) \in U$ and $\bar{y} \in W$, having the following property:
To every $y \in W$ corresponds a unique $x \in \mathbb{R}^n$ such that

$$(x, y) \in U \qquad \text{and} \qquad F(x, y) = 0\,.$$

Moreover, if x is defined to be $G(y)$, then G is a continuously differentiable mapping of W into $\mathbb{R}^n$ satisfying

$$G(\bar{y}) = \bar{x}, \quad F(G(y), y) = 0 \;\forall\; y \in W, \quad \text{and} \quad G'(\bar{y}) = -(\nabla_x F(\bar{x}, \bar{y}))^{-1} \nabla_y F(\bar{x}, \bar{y})\ .$$

1.6.2. *Some facts about the Second Derivative.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Then ∇f is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. The second derivative of f is by definition the first derivative of the gradient mapping $x \mapsto \nabla f(x)$, if it exists, that is the second derivative of f at x is the mapping $\nabla^2 f(x) := \nabla[\nabla f](x)$.

(i) If $\nabla^2 f(x)$ exists and is continuous at x, then with respect to the standard basis, it is given as the matrix of second partial derivatives:

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

Moreover, $\frac{\partial f}{\partial x_i \partial x_j} = \frac{\partial f}{\partial x_j \partial x_i}$ for all $i, j = 1, \ldots, n$. The matrix $\nabla^2 f(x_2)$ is called the *Hessian* of f at x. It is a symmetric matrix.

(ii) Second-Order Taylor Theorem:
If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable on an open set containing the interval $[x, y]$, then there is a point $z \in [x, y]$ such that

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

We also obtain

$$\|f(y) - (f(x) + \nabla f(x)(y - x))\| \leq \frac{1}{2}\|x - y\|^2 \sup_{z \in [x,y]} \|\nabla^2 f(z)\|.$$

1.6.3. *Integration.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and set $\varphi(t) := f(x + t(y - x))$ so that $\varphi : \mathbb{R} \to \mathbb{R}$. Then

$$\begin{aligned} f(y) - f(x) &= \varphi(1) - \varphi(0) = \textstyle\int_0^1 \varphi'(t)\, dt \\ &= \textstyle\int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt \end{aligned}$$

Similarly, for a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$, we have

$$\begin{aligned} F(y) - F(x) &= \begin{bmatrix} \int_0^1 \nabla F_1(x + t(y - x))^T (y - x) dt \\ \vdots \\ \int_0^1 \nabla F_m(x + t(y - x))^T (y - x) dt \end{bmatrix} \\ &= \textstyle\int_0^1 \nabla F(x + t(y - x))(y - x) dt \end{aligned}$$

1.6.4. *More Facts about Continuity.* Let $F : \mathbb{R}^n \to \mathbb{R}^m$.

– We say that F is *continuous relative to a set* $D \subset \mathbb{R}^n$ if for every $x \in D$ and $\epsilon > 0$ there exists a $\delta(x, \epsilon) > 0$ such that

$$\|F(y) - F(x)\| \leq \epsilon \text{ whenever } \|y - x\| \leq \delta(x, \epsilon) \quad \text{and} \quad y \in D.$$

– We say that F is *uniformly continuous* on $D \subset \mathbb{R}^n$ if for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that

$$\|F(y) - F(x)\| \leq \epsilon \text{ whenever } \|y - x\| \leq \delta(\epsilon) \quad \text{and} \quad x, y \in D.$$

FACT: If F is continuous on a compact set $D \subset \mathbb{R}^n$, then F is uniformly continuous on D.

– We say that F is *Lipschitz continuous on a set* $D \subset \mathbb{R}^n$ if there exists a constant $K \geq 0$ such that

$$\|F(x) - F(y)\| \leq K\|x - y\|$$

for all $x, y \in D$.

FACT: Lipschitz continuity implies uniform continuity.

Proof. Set $\delta = \epsilon/K$. □

EXAMPLES:

(1) $4(x) = x^{-1}$ is continuous on $(0, 1)$, but it is not uniformly continuous on $(0, 1)$.
(2) $f(x) = \sqrt{x}$ is uniformly continuous on $[0, 1]$, but it is not Lipschitz continuous on $[0, 1]$.

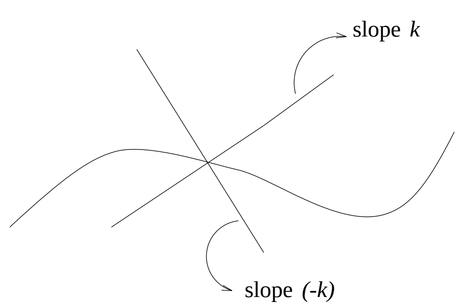
FACT: If ∇F exists and is continuous on a compact convex set $D \subset \mathbb{R}^m$, then F is Lipschitz continuous on D.

Proof. Mean value Theorem:

$$\|F(x) - F(y)\| \leq (\sup_{z \in [x,y]} \|\nabla F(z)\|)\|x - y\|.$$

Apply Weierstrass Compactness Theorem to ∇F. □

Lipschitz continuity is almost but not quite a differentiability hypothesis. The Lipschitz constant provides bounds on rate of change.



1.6.5. *Quadratic Bound Lemma.* Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be such that ∇F is Lipschitz continuous on the convex set $D \subset \mathbb{R}^n$. Then

$$\|F(y) - (F(x) + \nabla F(x)(y - x))\| \leq \frac{K}{2}\|y - x\|^2$$

for all $x, y \in D$ where K is a Lipschitz constant for ∇F on D.

Proof.

$$\begin{aligned}
F(y) - F(x) - \nabla F(x)(y-x) &= \int_0^1 \nabla F(x+t(y-x))(y-x)dt - \nabla F(x)(y-x) \\
&= \int_0^1 [\nabla F(x+t(y-x)) - \nabla F(x)](y-x)dt \\
\|F(y) - (F(x) + \nabla F(x)(y-x))\| &= \| \int_0^1 [\nabla F(x+t(y-x)) - \nabla F(x)](y-x)dt\| \\
&\leq \int_0^1 \|(\nabla F(x+t(y-x) - \nabla F(x))(y-x)\|dt \\
&\leq \int_0^1 \|\nabla F(x+t(y-x)) - \nabla F(x)\| \, \|y-x\|dt \\
&\leq \int_0^1 Kt\|y-x\|^2 dt \\
&= \frac{K}{2}\|y-x\|^2.
\end{aligned}$$

□

1.6.6. *Some Facts about Symmetric Matrices.* Let $H \in \mathbb{R}^{n\times n}$ be symmetric, i.e. $H^T = H$

(1) There exists an orthonormal basis of eigen-vectors for H, i.e. if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the n eigenvalues of H (not necessarily distinct), then there exist vectors $q_1, \ldots, q_n$ such that $\lambda_i q_i = H q_i$ $i = 1, \ldots, n$ with $q_i^T q_j = \delta_{ij}$. Equivalently, there exists an orthogonal transformation $Q = [q_1, \ldots, q_n]$ $(Q^T Q = I)$ such that

$$H = Q\Lambda Q^T$$

where $\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_n]$.

(2) $H \in \mathbb{R}^{n\times n}$ is positive semi-definite, i.e.

$$x^T H x \geq 0 \text{ for all } x \in \mathbb{R}^n,$$

if and only if all the eigenvalues of H are nonnegative.