APM462: Nonlinear Optimization

Tianyu Du

September 21, 2019

Contents

1 Preliminaries

1.1 Mean Value Theorems and Taylor Approximations.

Definition 1.1. Let $f : S \subset \mathbb{R}^n \to \mathbb{R}$, the **gradient** of f at $x \in S$, if exists, is a vector $\nabla f(x) \in \mathbb{R}^n$ characterized by the property

$$\lim_{v \to 0} \frac{f(x+v) - f(x) - \nabla f(x) \cdot v}{||v||} = 0 \tag{1.1}$$

Theorem 1.1 (The First Order of Mean Value Theorem)**.** Let f be a C^1 real-valued function defined on $\mathbb{R}^n$, then for any $x, v \in \mathbb{R}^n$, there exists some $\theta \in (0,1)$ such that

$$f(x+v) = f(x) + \nabla f(x + \theta v) \cdot v \tag{1.2}$$

Proof. Let $x, v \in \mathbb{R}^n$, define $g(t) : \mathbb{R} \to \mathbb{R} := f(x+tv)$, which is C^1. By the mean value theorem on $\mathbb{R}^{\mathbb{R}}$, there exists $\theta \in (0,1)$ such that $g(0+1) = g(0) + g'(\theta)(1-0)$, that is, $f(x+v) = f(x) + g'(\theta)$. Note that $g'(\theta) = \nabla(x + \theta v) \cdot v$, what desired is immediate. ■

Proposition 1.1 (The First Order Taylor Approximation)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function, then

$$f(x+v) = f(x) + \nabla f(x) \cdot v + o(||v||) \tag{1.3}$$

that is

$$\lim_{||v|| \to 0} \frac{f(x+v) - f(x) - \nabla f(x) \cdot v}{||v||} = 0 \tag{1.4}$$

Proof. By the mean value theorem, $\exists \theta \in (0,1)$ such that $f(x+v) - f(x) = \nabla f(x+\theta v) \cdot v$. The limit becomes $\lim_{||v|| \to 0} \frac{[\nabla f(x+\theta v) - \nabla f(x)] \cdot v}{||v||} = \lim_{||v|| \to 0; x+\theta v \to x} \frac{[\nabla f(x+\theta v) - \nabla f(x)] \cdot v}{||v||}$. Since $f \in C^1$, $\lim_{x+\theta v \to x} \nabla f(x+\theta v) = \nabla f(x)$. And $\frac{v}{||v||}$ is a unit vector, and every component of it is bounded, as the result, the limit of inner product vanishes instead of explodes. ■

Theorem 1.2 (The Second Order Mean Value Theorem)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2 function, then for any $x, v \in \mathbb{R}^n$, there exists $\theta \in (0,1)$ satisfying

$$f(x+v) = f(x) + \nabla f(x) \cdot v + \frac{1}{2} v' H_f(x + \theta v)\ v \tag{1.5}$$

where H_f is the Hessian matrix of f, may also be written as $\nabla^2 f$.

Proposition 1.2 (The Second Order Taylor Approximation)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2 function, and $x, v \in \mathbb{R}^n$, then

$$f(x+v) = f(x) + \nabla f(x) \cdot v + \frac{1}{2} v' H_f(x)\ v + o(||v||^2) \tag{1.6}$$

that is

$$\lim_{||v|| \to 0} \frac{f(x+v) - f(x) - \nabla f(x) \cdot v - \frac{1}{2} v' H_f(x)\ v}{||v||^2} = 0 \tag{1.7}$$

Proof. By the second mean value theorem, there exists $\theta \in (0,1)$ such that the limit is equivalent to

$$\lim_{||v|| \to 0} \frac{1}{2} \left(\frac{v}{||v||} \right)' [H_f(x+\theta v) - H_f(x)] \frac{v}{||v||} \tag{1.8}$$

Since $f \in C^2$, the limit of $[H_f(x+\theta v) - H_f(x)]$ is in fact $\mathbf{0}_{n \times n}$. And every component of unit vector $\frac{v}{||v||}$ is bounded, the quadratic form converges to zero as an immediate result. ■

It is often noted that the gradient at a particular $x_0 \in dom(f) \subset \mathbb{R}^n$ gives the direction f increases most rapidly. Let $x_0 \in dom(f)$, and v be a unit vector representing a *feasible direction* of change. That is, there exists $\delta > 0$ such that $x_0 + tv \in dom(f)\ \forall t \in [0, \delta)$. Then the rate of change of f along feasible direction v can be written as

$$\left.\frac{d}{dt}\right|_{t=0} f(x_0 + tv) = \nabla f(x_0) \cdot v = ||\nabla f(x_0)||\ ||v|| \cos(\theta) \tag{1.9}$$

where $\theta = \angle(v, \nabla f(x_0)$. And the derivative is maximized when $\theta = 0$, that is, when v and ∇f point the same direction.

1.2 Implicit Function Theorem

Theorem 1.3 (Implicit Function Theorem)**.** Let $f : \mathbb{R}^{n+1} \to \mathbb{R}$ be a C^1 function, let $(a, b) \in \mathbb{R}^n \times \mathbb{R}$ such that $f(a, b) = 0$. If $\nabla f(a, b) \neq 0$, then $\{(x, y) \in \mathbb{R}^n \times \mathbb{R} : f(x, y) = 0\}$ is locally a graph of a function $g : \mathbb{R}^n \to \mathbb{R}$.

Remark 1.1. $\nabla f(x_0) \perp$ level set of f near x_0.

2 Convexity

2.1 Terminologies

Definition 2.1. Set $\Omega \subset \mathbb{R}^n$ is **convex** if and only if

$$\forall x_1, x_2 \in \Omega,\ \lambda \in [0, 1],\ \lambda x_1 + (1 - \lambda)x_2 \in \Omega \tag{2.1}$$

Definition 2.2. A function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is **convex** if and only if Ω is convex, and

$$\forall x_1, x_2 \in \Omega,\ \lambda \in [0, 1],\ f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \tag{2.2}$$

Definition 2.3. A function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** if and only if Ω is convex and

$$\forall x_1, x_2 \in \Omega,\ \lambda \in (0, 1),\ f(\lambda x_1 + (1 - \lambda)x_2) < \lambda f(x_1) + (1 - \lambda)f(x_2) \tag{2.3}$$

2.2 Basic Properties of Convex Functions

Definition 2.4. A function $f : \Omega \to \mathbb{R}$ is **concave** if and only if $-f$ is **convex**.

Proposition 2.1. (i) If f_1, f_2 are convex on Ω, so is $f_1 + f_2$;

(ii) If f is convex on Ω, then for any $a > 0$, af is also convex on Ω;

(iii) Any **sub-level/lower contour set** of a convex function f

$$SL(c) := \{x \in \mathbb{R}^n : f(x) \leq c\} \tag{2.4}$$

is convex.

Proof of (iii). Let $c \in \mathbb{R}$, and $x_1, x_2 \in SL(c)$. Let $s \in [0, 1]$. Since $x_1, x_2 \in SL(c)$, and $f(\cdot)$ is convex, $f(sx_1 + (1 - s)x_2) \leq sf(x_1) + (1 - s)f(x_2) \leq sc + (1 - s)c = c$. Which implies $sx_1 + (1 - s)x_2 \in SL(c)$. ∎

Example 2.1. $f(x) : \mathbb{R}^n \to \mathbb{R} := ||x||$ is convex.

Proof. Note that for any $u, v \in \mathbb{R}^n$, by triangle inequality, $||u-(-v)|| \leq ||u-0||+||0-(-v)|| = ||u||+||v||$. Consequently, let $u, v \in \mathbb{R}^n$ and $s \in [0,1]$, then $||su+(1-s)v|| \leq ||su||+||(1-s)v|| = s||u||+(1-s)||v||$. Therefore, $||\cdot||$ is convex. ∎

2.3 Characteristics of C^1 Convex Functions

Theorem 2.1 (C^1 criterions for convexity)**.** Let $f \in C^1$, then f is convex on a convex set Ω <u>if and only if</u>

$$\forall x, y \in \Omega,\ f(y) \geq f(x) + \nabla f(x) \cdot (y-x) \tag{2.5}$$

that is, *the linear approximation is never an overestimation of value of f.*

Proof. ($\implies$) Suppose f is convex on a convex set Ω. Then $f(sy+(1-s)x) \leq sf(y)+(1-s)f(x)$ for every $x, y \in \Omega$ and $s \in [0,1]$, which implies, for every $s \in (0,1]$:

$$\frac{f(sy+(1-s)x)-f(x)}{s} \leq f(y)-f(x) \tag{2.6}$$

By taking the limit of $s \to 0$,

$$\lim_{s\to 0} \frac{f(x+s(y-x))-f(x)}{s} \leq f(y)-f(x) \tag{2.7}$$

$$\implies \left.\frac{d}{ds}\right|_{s=0} f(x+s(y-x)) \leq f(y)-f(x) \tag{2.8}$$

$$\implies \nabla f(x)\cdot(y-x) \leq f(y)-f(x) \tag{2.9}$$

($\impliedby$) Let $x_0, x_1 \in \Omega$, let $s \in [0,1]$. Define $x^* := sx_0 + (1-s)x_1$, then

$$f(x_0) \geq f(x^*) + \nabla f(x^*)\cdot(x_0 - x^*) \tag{2.10}$$

$$\implies f(x_0) \geq f(x^*) + \nabla f(x^*)\cdot[(1-s)(x_0-x_1)] \tag{2.11}$$

Similarly,

$$f(x_1) \geq f(x^*) + \nabla f(x^*)\cdot(x_1 - x^*) \tag{2.12}$$

$$\implies f(x_1) \geq f(x^*) + \nabla f(x^*)\cdot[s(x_1-x_0)] \tag{2.13}$$

Therefore, $sf(x_0)+(1-s)f(x_1) \geq f(x^*)$. ∎

Theorem 2.2 (C^2 criterion for convexity)**.** $f \in C^2$ is a convex function on a convex set $\Omega \subset \mathbb{R}^n$ <u>if and only if</u> $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in \Omega$.

Remark 2.1. When f is defined on $\mathbb{R}$, the C^2 criterion becomes $f''(x) \geq 0$.

Proof. ($\impliedby$) Suppose $\nabla^2 f(x) \succcurlyeq 0$ for every $x \in \Omega$, let $x, y \in \Omega$. By the second order MVT,

$$f(y) = f(x) + \nabla f(x)\cdot(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x+s(y-x))(y-x) \text{ for some } s \in [0,1] \tag{2.14}$$

$$\implies f(y) \geq f(x) + \nabla f(x)\cdot(y-x) \tag{2.15}$$

So f is convex by the C^1 criterion of convexity.
($\implies$) Let $v \in \mathbb{R}^n$. Suppose, for contradiction, that for some $x \in \Omega$, $\nabla^2 f(x) \not\succcurlyeq 0$. If such $x \in \partial\Omega$, note that $v^T\nabla^2 f(\cdot)v$ is continuous because $f \in C^2$, then there exists $\varepsilon > 0$ such that $\forall x' \in V_\varepsilon(x) \cap \Omega^{int}$, $v^T \nabla^2 f(x')v <$

0. Hence, one may assume with loss of generality that such $x \in \Omega^{int}$. Because $x \in \Omega^{int}$, exists $\varepsilon' > 0$, such that $V_{\varepsilon'}(x) \subseteq \Omega^{int}$. Define $\hat{v} := \frac{v}{\sqrt{\varepsilon'}}$, then for every $s \in [0,1]$, $\hat{v}^T \nabla^2 f(x + s\hat{v})\hat{v} < 0$. Let $y = x + \hat{v}$, by the mean value theorem, $f(y) = f(x) + \nabla f(x) \cdot (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x + s(y-x))(y-x)$ for some $s \in [0,1]$. This implies $f(y) < f(x) + \nabla f(x) \cdot (y-x)$, which contradicts the C^1 criterion for convexity. ■

2.4 Minimum and Maximum of Convex Functions

Theorem 2.3. Let $\Omega \subset \mathbb{R}^n$ be a convex set, and $f : \Omega \to \mathbb{R}$ is a convex function. Let

$$\Gamma := \left\{ x \in \Omega : f(x) = \min_{x \in \Omega} f(x) \right\} \equiv \operatorname*{argmin}_{x \in \Omega} f(x) \tag{2.16}$$

If $\Gamma \neq \varnothing$, then

(i) Γ is convex;

(ii) any local minimum of f is the global minimum.

Proof (i). Let $x, y \in \Gamma$, $s \in [0,1]$, then $sx + (1-s)y \in \Omega$ because Ω is convex. Since f is convex, $f(sx + (1-s)y) \leq sf(x) + (1-s)f(y) = \min_{x \in \Omega} f(x)$. The inequality must be equality since it would contradicts the fact that $x, y \in \Gamma$. Therefore, $sx + (1-s)y \in \Gamma$. ■

Proof (ii). Let $x \in \Omega$ be a local minimizer for f, but assume, for contradiction, it is not a global minimizer. That is, there exists some other y such that $f(y) < f(x)$. Since f is convex,

$$f(x + t(y-x)) = f((1-t)x + ty) \leq (1-t)f(x) + tf(y) < f(x) \tag{2.17}$$

for every $t \in (0,1]$. Therefore, for every $\varepsilon > 0$, there exists $t^* \in (0,1]$ such that $x + t^*(y-x) \in V_\varepsilon(x)$ and $f(x + t^*(y-x)) < f(x)$, this contradicts the fact that x is a local minimum. ■

Theorem 2.4. Let $\Omega \subset \mathbb{R}^n$ be a convex set, and $f : \Omega \to \mathbb{R}$ is a convex function. Then

$$\max_{x \in \Omega} f(x) = \max_{x \in \partial\Omega} f(x) \tag{2.18}$$

Proof. As we assumed, Ω is closed, therefore $\partial\Omega \subseteq \Omega$. Hence, $\max_{x \in \Omega} f \geq \max_{x \in \partial\Omega} f$. Suppose $\max_{x \in \Omega} f > \max_{x \in \partial\Omega} f$, let $x^* := \operatorname{argmax}_{x \in \Omega} f \in \Omega^{int}$. Then we can construct a straight line through x^* and intersects $\partial\Omega$ at two points, $y_1, y_2 \in \partial\Omega$, such that $x^* = sy_1 + (1-s)y_2$ for some $s \in (0,1)$. Further, since f is convex, $\max_{x \in \Omega} f(x) = f(x^*) \leq sf(y_1) + (1-s)f(y_2) \leq s \max_{\partial\Omega} f + (1-s) \max_{\partial\Omega} f = \max_{\partial\Omega} f$, which leads to a contradiction. Therefore, $\max_{x \in \Omega} f = \max_{x \in \partial\Omega} f$. ■

Proposition 2.2. For p, g ¿ 1 and $\frac{1}{p} + \frac{1}{g} = 1$,

$$|ab| \leq \frac{1}{p}|a|^p + \frac{1}{g}|b|^g \tag{2.19}$$

Proof.

$$(-\log)|ab| = (-\log)|a| + (-\log)|b| \tag{2.20}$$

$$= \frac{1}{p}(-\log)|a|^p + \frac{1}{g}(-\log)|b|^p \tag{2.21}$$

$$(\because (-\log) \text{ is convex}) \geq (-\log)\left(\frac{1}{p}|a|^p + \frac{1}{g}|b|^p\right) \tag{2.22}$$

And since $(-\log)$ is monotonically decreasing,

$$|ab| \leq \frac{1}{p}|a|^p + \frac{1}{g}|b|^p \tag{2.23}$$

■

Corollary 2.1.

$$|ab| \leq \frac{|a|^2 + |b|^2}{2} \tag{2.24}$$

3 Finite Dimensional Optimization

3.1 Unconstraint Optimization

Theorem 3.1 (Extreme Value Theorem)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and $K \subset \mathbb{R}^n$ be a compact set, then the minimization problem $\min_{x \in K} f(x)$ has a solution.

Remark 3.1. $f : \Omega \to \mathbb{R}$ is convex does not imply f is continuous.

Proposition 3.1. A convex function f defined on a convex open set is continuous.

Proof. Let $f : \Omega \to \mathbb{R}$ be a convex function, where $\Omega \subset \mathbb{R}^n$ is open. TODO ■

Corollary 3.1. A convex function f defined on an open interval in $\mathbb{R}$ is continuous.

Proof. See homework 1, using squeeze theorem. ■

Proof of EVT.. Let $f : K \to \mathbb{R}$ be a continuous function defined on a compact set K.
WLOG, we only prove the existence of $\min f$, since the existence of max can be easily proven by applying the exact same argument on $-f$. Because K is compact, the continuity of f implies $f(K)$ is compact. By the completeness axiom of $\mathbb{R}$, $m := \inf_{x \in K} f(x)$ is well-defined. There exists a sequence $(x_i) \subset K$, such that $(f(x_i)) \to m$. Because K is compact, there exists a subsequence (x_{ik}) of (x_i) converges to some limit $x^* \in K$. Because f is continuous, $(f(x_{ik})) \to f(x^*)$, which is a subsequence of the convergent sequence $(f(x_i))$, and they must converge to the same limit. Hence, $f(x^*) = m$, and the infimum is attained at $x^* \in K$. ■

Theorem 3.2 (Heine–Borel)**.** Let $K \subset \mathbb{R}^n$, then K is compact (every open cover of K has a finite sub-cover) $\iff$ K is closed and bounded.

Proposition 3.2. Let $\{h_i\}$ and $\{g_i\}$ be sets of continuous functions on $\mathbb{R}^n$, the the set of all points in $\mathbb{R}^n$ that satisfy

$$\begin{cases} h_i(x) = 0 \ \forall i \\ g_j(x) \leq 0 \ \forall j \end{cases} \tag{3.1}$$

is a closed set (intersection of finitely many closed sets). Moreover, if the qualified set is also bounded, then it is compact.

Proof. For every equality constraint h_i, it can be represented as the conjunction of two inequality constraint, namely $h_i^\alpha(x) := -h_i(x) \leq 0 \wedge h_i^\beta(x) := h_i(x) \leq 0$. Then the constraint collection is equivalent to

$$\begin{cases} h_i^\alpha(x) \leq 0 \ \forall i \\ h_i^\beta(x) \leq 0 \ \forall i \\ g_j(x) \leq 0 \ \forall j \end{cases} \tag{3.2}$$

The subset of $\mathbb{R}^n$ qualified by each individual constraint is closed by the property of continuous functions (i.e. the continuous function's pre-image of closed set is closed). And the intersection of arbitrarily many closed sets is closed. ■

Example 3.1. The set $\{(x, y) \in \mathbb{R}^2 : x^2 - y^2 - 1 = 0\}$ is closed and bounded, therefore it is compact.

Remark 3.2. Computer algorithms for solving minimization problems try to construct a sequence of (x_i) such that $f(x_i)$ decreases to $\min f$ rapidly.

The optimization problems investigated in this section can be formulated as

$$\min_{x \in \Omega} f(x) \tag{3.3}$$

where $\Omega \subset \mathbb{R}^n$. Typically, for simplicity, Ω are often $\mathbb{R}^n$, an open subset of R^n, or the closure of some open subset of $\mathbb{R}^n$.

Everything above minimization discussed in this section is applicable to maximization as well using the proposition below.

Proposition 3.3. When $\Omega = \mathbb{R}^n$, the unconstrained minimization has the following properties

(i) $\operatorname{argmax} f = \operatorname{argmin}(-f)$;

(ii) $\max f = -\min(-f)$

Proof. Omitted. ■

Definition 3.1. A function $f : \Omega \to \mathbb{R}$ has **local minimum** at $x_0 \in \Omega$ if

$$\exists \varepsilon > 0 \ s.t. \ \forall x \in V_\varepsilon(x_0) \cap \Omega \ f(x_0) \leq f(x) \tag{3.4}$$

f attains **strictly local minimum** at x_0 if

$$\exists \varepsilon > 0 \ s.t. \ \forall x \in V_\varepsilon(x_0) \cap \Omega \backslash \{x_0\} \ f(x_0) < f(x) \tag{3.5}$$

f attains **global minimum** at x_0 if

$$\forall x \in \Omega \ f(x_0) \leq f(x) \tag{3.6}$$

f attains **strict global minimum** at x_0 if

$$\forall x \in \Omega \backslash \{x_0\} \ f(x_0) < f(x) \tag{3.7}$$

Note that strict global minimum is always unique.

Theorem 3.3 (Necessary Condition for Local Minimum)**.** Let $C^1 \ni f : \Omega \to \mathbb{R}$, let $x_0 \in \Omega$ be a local minimum of f then, for every *feasible direction* v at x_0,

$$\nabla f(x_0) \cdot v \geq 0 \tag{3.8}$$

Definition 3.2. For $x_0 \in \Omega \subset \mathbb{R}^n$, $v \in \mathbb{R}^n$ is a **feasible direction**at x_0 if

$$\exists \overline{s} > 0 \ s.t. \ \forall s \in [0, \overline{s}], x_0 + sv \in \Omega \tag{3.9}$$

Proof of Necessary Condition. The prove is almost immediate, if there exists a feasible direction v^* such that $\nabla f(x_0) \cdot v^* < 0$, for every $\varepsilon > 0$, one can construct $x' := x^* + sv^*$ with sufficiently small s so that $x' \in V_\varepsilon(x^*) \cap \Omega$ and $f(x') < f(x^*)$. ■

Corollary 3.2. When Ω is open, then x_0 is a local minimum $\implies \nabla f(x_0) = 0$.

Proof. Since Ω is open, any sufficiently small $v \neq 0$ such that both v and $-v$ are feasible directions at x_0, applying the necessary condition on both v and $-v$ provides the equality. ■

3.2 Equality Constraints

3.3 Inequality Constraints