Numerical Optimization

Lecture 10: Nonlinear Optimization (基础知识)

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

Gradients and Hessians

Definition (Gradient)

If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable (i.e. $f \in \mathcal{C}$), then the gradient of f is the function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ defined for all $x \in \mathbb{R}^n$ by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Definition (Hessian)

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable (i.e. $f \in \mathcal{C}^2$), then the Hessian of f is the function $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ defined for all $x \in \mathbb{R}^n$ by

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \quad \text{(a symmetric matrix)}.$$

Mean Value Theorem and Taylor's Theorem

Our primary tools in developing optimality conditions are the following.

Theorem (Mean Value Theorem)

Given $f \in \mathcal{C}$, $x \in \mathbb{R}^n$, and $d \in \mathbb{R}^n$, there exists $\alpha \in (0,1)$ such that

$$f(x+d) = f(x) + \nabla f(x+\alpha d)^T d.$$

The generalization of the Mean Value Theorem to higher order derivatives is often attributed to Taylor. For example, we have the following.

Theorem (Taylor's Theorem (Second Order))

Given $f \in \mathcal{C}^2$, $x \in \mathbb{R}^n$, and $d \in \mathbb{R}^n$, there exists $\alpha \in (0,1)$ such that

$$f(x+d) = f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x+\alpha d)^T d.$$

定义 2.4 (梯度利普希茨连续) 给定可微函数 f，若存在 $L>0$，对任意的 $x,y\in\mathbf{dom}\,f$ 有

$$\|\nabla f(x)-\nabla f(y)\|\leqslant L\|x-y\|, \tag{2.2.2}$$

则称 f 是**梯度利普希茨连续的**，相应利普希茨常数为 L. 有时也简记为**梯度 L-利普希茨连续**或 **L-光滑**.

引理 2.1 (二次上界) 设可微函数 $f(x)$ 的定义域 $\mathbf{dom}\,f=\mathbb{R}^n$，且为梯度 L-利普希茨连续的，则函数 $f(x)$ 有二次上界：

$$f(y)\leqslant f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{L}{2}\|y-x\|^2,\quad \forall\, x,y\in\mathbf{dom}\,f. \tag{2.2.3}$$

推论 2.1 设可微函数 $f(x)$ 的定义域为 $\mathbb{R}^n$ 且存在一个全局极小点 x^*，若 $f(x)$ 为梯度 L-利普希茨连续的，则对任意的 x 有

$$\frac{1}{2L}\|\nabla f(x)\|^2\leqslant f(x)-f(x^*). \tag{2.2.5}$$

Epigraphs and effective domains

(广义实值函数)

Consider $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ (the extended reals; i.e. $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$).

(Epigraph) 上方图

The epigraph of f is

$$\text{epi}(f) := \{(x, z) : x \in \mathcal{X},\ z \in \mathbb{R}, \text{ and } f(x) \leq z\}.$$

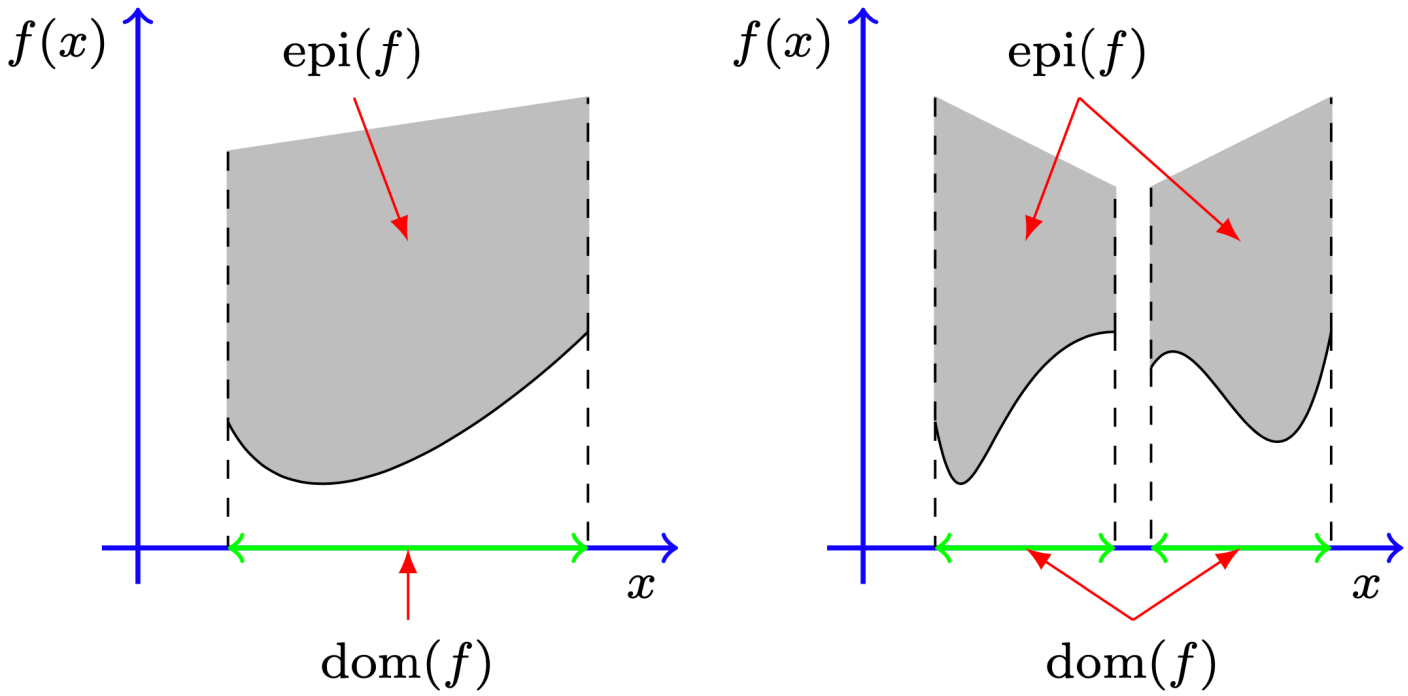
(Effective domain) 有效域

The effective domain of f is

$$\text{dom}(f) := \{x : x \in \mathcal{X} \text{ and } f(x) < \infty\}.$$

These definitions imply that

$$\text{dom}(f) = \{x : x \in \mathcal{X} \text{ and there exists } z \in \mathbb{R} \text{ such that } (x, z) \in \text{epi}(f)\}.$$



Epigraphs of a convex (left) and a nonconvex (right) function

(Proper function) 适当函数

A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper if

$$f(x) \begin{cases} < +\infty & \text{for some } x \in \mathcal{X} \\ > -\infty & \text{for all } x \in \mathcal{X}. \end{cases}$$

Otherwise, it is improper.

定义 2.8 (α-下水平集) 对于广义实值函数 $f: \mathbb{R}^n \to \overline{\mathbb{R}}$,

$$C_\alpha = \{x \,|\, f(x) \leqslant \alpha\}$$

称为 f 的**α-下水平集**.

定义 2.12 如果过集合 C 中任意两点的直线都在 C 内，则称 C 为**仿射集**，即

$$x_1, x_2 \in C \Longrightarrow \theta x_1 + (1-\theta) x_2 \in C, \forall \theta \in \mathbb{R}.$$

定义 2.13 如果连接集合 C 中任意两点的线段都在 C 内，则称 C 为**凸集**，即

$$x_1, x_2 \in C \Longrightarrow \theta x_1 + (1-\theta) x_2 \in C, \forall 0 \leqslant \theta \leqslant 1.$$

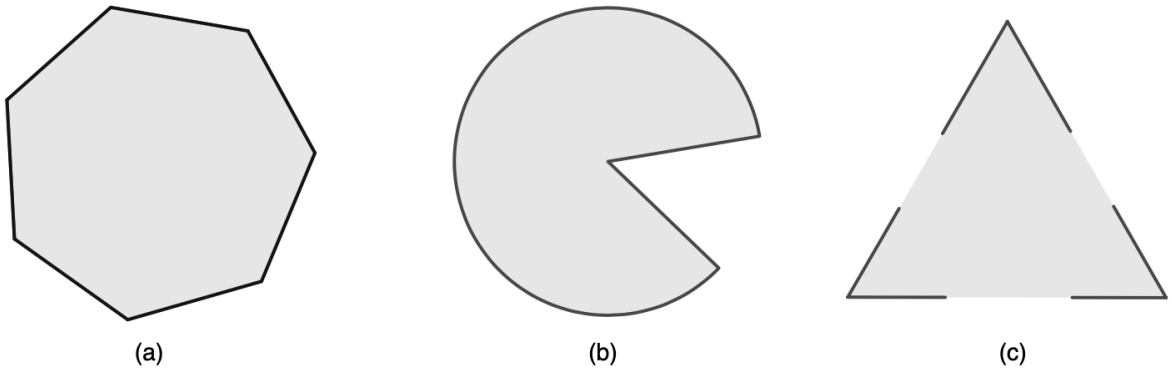
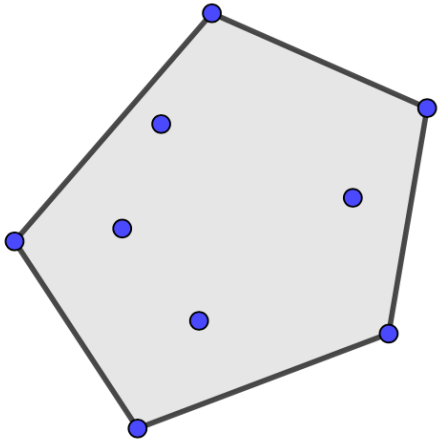


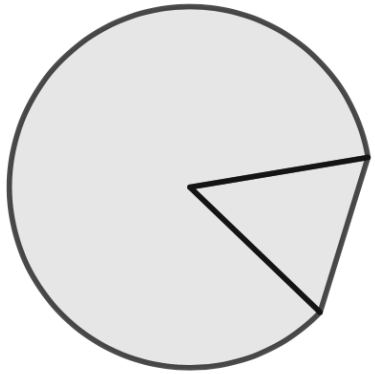
图 2.4 一个凸集和两个非凸集

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k,$$

$$1 = \theta_1 + \theta_2 + \cdots + \theta_k, \quad \theta_i \geqslant 0, i = 1,2,\cdots,k$$

的点称为 $x_1, x_2, \cdots, x_k$ 的**凸组合**. 集合 $\boldsymbol{S}$ 中点所有可能的凸组合构成的集合称作 S 的**凸包**，记作 $\mathbf{conv}\, S$. 实际上，$\mathbf{conv}\, \boldsymbol{S}$ 是包含 $\boldsymbol{S}$ 的最小的凸集. 如

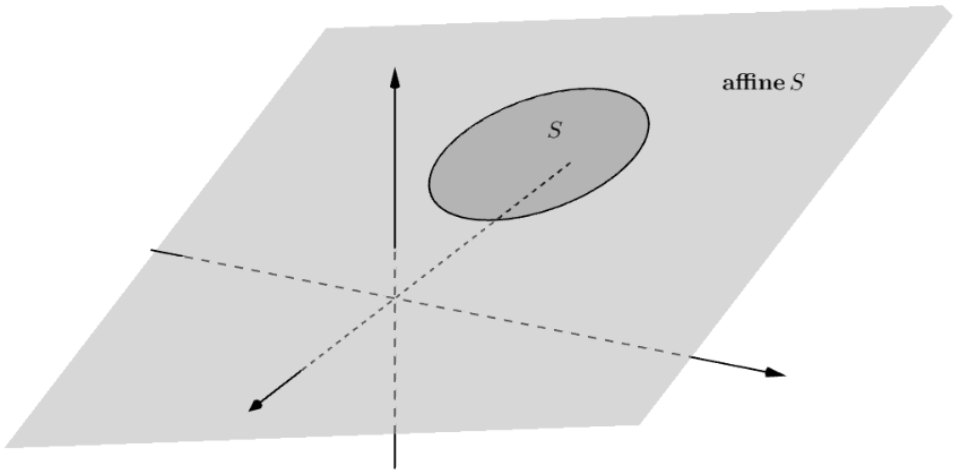




定义 2.14 (仿射包) 设 S 为 $\mathbb{R}^n$ 的子集，称如下集合为 S 的**仿射包**:

$$\{x \mid x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k, \quad x_1, x_2, \cdots, x_k \in S, \quad \theta_1 + \theta_2 + \cdots + \theta_k = 1\},$$

记为 $\mathbf{affine}\, S$.



Cones

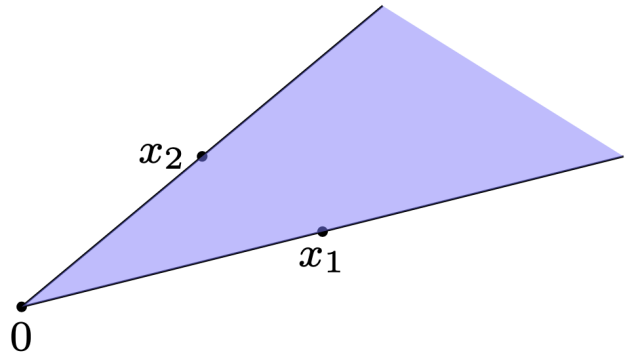
Cones, a particular class of sets, play significant roles in optimization.

- They arise in optimality conditions and constraint qualifications.
- They also arise in the area of conic optimization.

(Cone)

A set $\mathcal{X} \subseteq \mathbb{R}^n$ is a cone if

$$\alpha x \in \mathcal{X} \quad \text{for all} \quad x \in \mathcal{X} \quad \text{and} \quad \alpha \in [0, \infty)$$

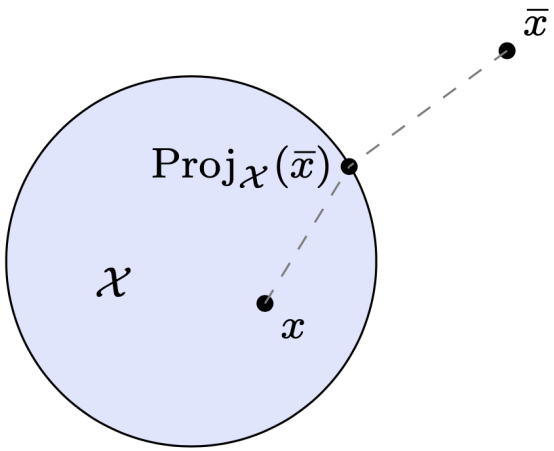


Not all cones are convex! But many important ones are convex.

Projections

(Projection)

If $\mathcal{X} \subseteq \mathbb{R}^n$ is nonempty, closed, and convex and $\bar{x} \in \mathbb{R}^n$, then a projection of $\bar{x}$ onto $\mathcal{X}$ is any minimizer of $\|x - \bar{x}\|_2$ over $x \in \mathcal{X}$ (a nonlinear optimization problem!).



Fundamental properties

(Projection Theorem)

Let $\mathcal{X}$ be a nonempty closed convex subset of $\mathbb{R}^n$ and consider $\bar{x} \in \mathbb{R}^n$. Then, the projection of $\bar{x}$ onto $\mathcal{X}$ is unique. Moreover, $\mathrm{Proj}_{\mathcal{X}}(\bar{x})$ is the projection of $\bar{x}$ onto $\mathcal{X}$ if and only if, for all $x \in \mathcal{X}$, we have

$$(\bar{x} - \mathrm{Proj}_{\mathcal{X}}(\bar{x}))^T (x - \mathrm{Proj}_{\mathcal{X}}(\bar{x})) \leq 0.$$

Theorem

If $\mathcal{X} \subseteq \mathbb{R}^n$ is nonempty, closed, and convex, then

$$\| \mathrm{Proj}_{\mathcal{X}}(x_1) - \mathrm{Proj}_{\mathcal{X}}(x_2) \|_2 \leq \|x_1 - x_2\|_2 \quad \textit{for all } \{x_1, x_2\} \subset \mathbb{R}^n.$$

Due to this theorem, we say that the projection operator is nonexpansive.

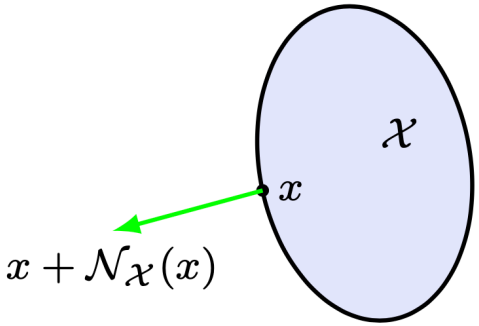
Normal cone

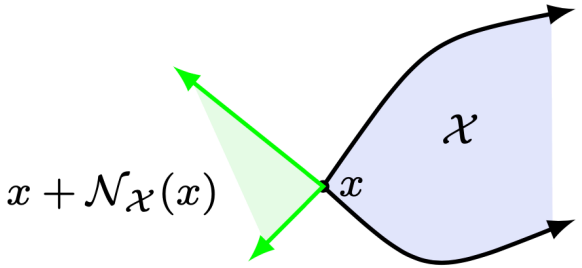
Definition (Normal Cone)

Given a nonempty convex $\mathcal{X} \subseteq \mathbb{R}^n$ and $x \in \mathcal{X}$, the normal cone of $\mathcal{X}$ at x is

$$\mathcal{N}_{\mathcal{X}}(x) := \{g : g^T(\bar{x} - x) \leq 0 \text{ for all } \bar{x} \in \mathcal{X}\}.$$

If $x \in \text{int}(\mathcal{X})$, then clearly $\mathcal{N}_{\mathcal{X}}(x) = \{0\}$, but for $x \notin \text{int}(\mathcal{X})$ the normal cone contains at least one halfline.





Separation

The following is a result of the Projection Theorem (2.1.6).

Theorem (Supporting Hyperplane Theorem)

Consider a nonempty convex $\mathcal{X} \subseteq \mathbb{R}^n$ and $\bar{x} \in \mathbb{R}^n$. If $\bar{x} \notin \operatorname{int}(\mathcal{X})$, then there exists a hyperplane passing through $\bar{x}$ that contains $\mathcal{X}$ in one of its closed halfspaces, i.e., there exists $a \neq 0$ such that

$$a^T \bar{x} \leq a^T x \quad \text{for all } x \in \mathcal{X}.$$

The previous result is easily extended to the case of separating two sets.

Theorem (Separating Hyperplane Theorem)

Consider nonempty convex sets $\mathcal{X}_1 \subseteq \mathbb{R}^n$ and $\mathcal{X}_2 \subseteq \mathbb{R}^n$. If $\mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing$, then there exists a hyperplane that separates them, i.e., there exists $a \neq 0$ such that

$$a^T x_1 \leq a^T x_2 \quad \text{for all } x_1 \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2.$$

Separation can also be strict, strong, proper, etc.

Geometry vs Algebraic

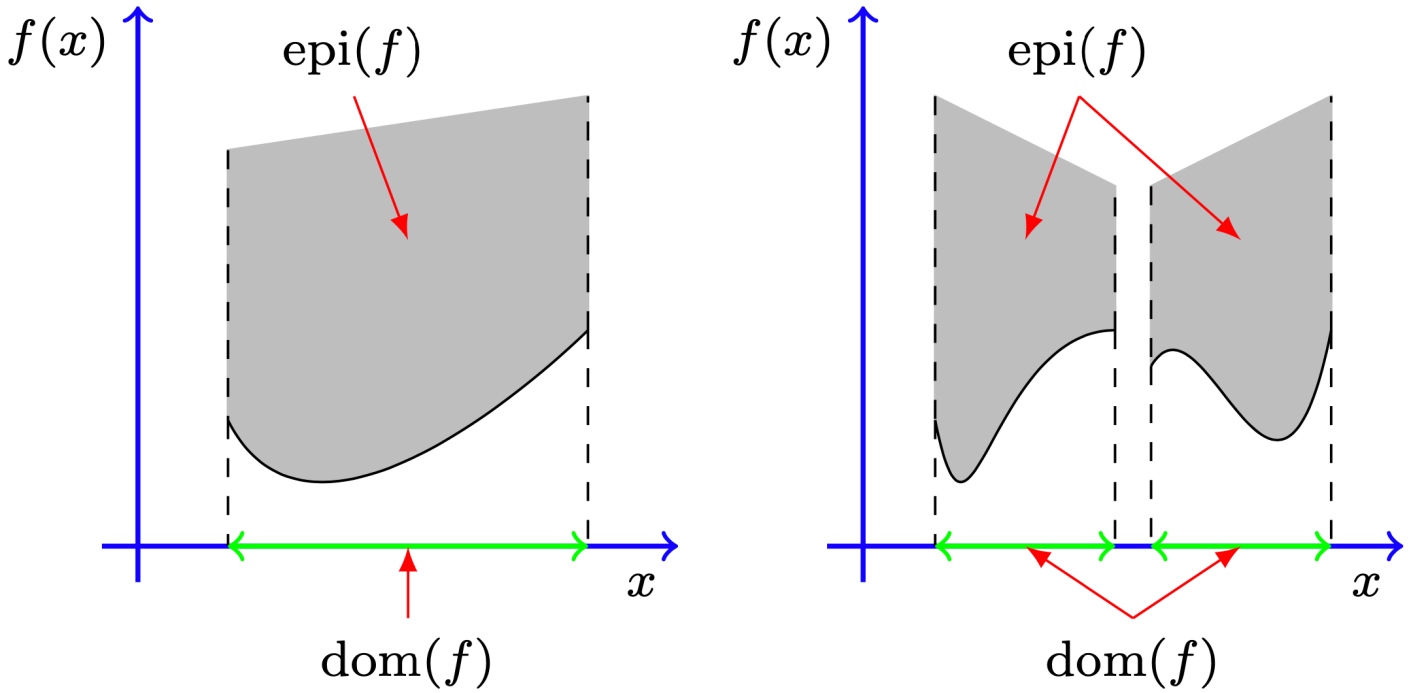
- Notice that the definition of the normal cone and the tangent cone of a set $\mathcal{X}$ at a point x had nothing to do with any algebraic description of the set $\mathcal{X}$.
- Rather, the definition is based purely on geometry.
- This is important since for any given set in $\mathbb{R}^n$, there are an infinite number of algebraic descriptions, and not all of them are equal in optimization theory and algorithms; e.g., consider the convex sets

$$\mathcal{X}_1 := \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 - 1 \leq 0\}$$
$$\text{and} \quad \mathcal{X}_2 := \{x \in \mathbb{R}^2 : (x_1^2 + x_2^2 - 1)^3 \leq 0\}.$$

Convex functions (revisited)

Definition (Convex function, extended real-valued)

A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if $\text{epi}(f) \subseteq \mathbb{R}^{n+1}$ is a convex set.



Epigraphs of a convex (left) and a nonconvex (right) function

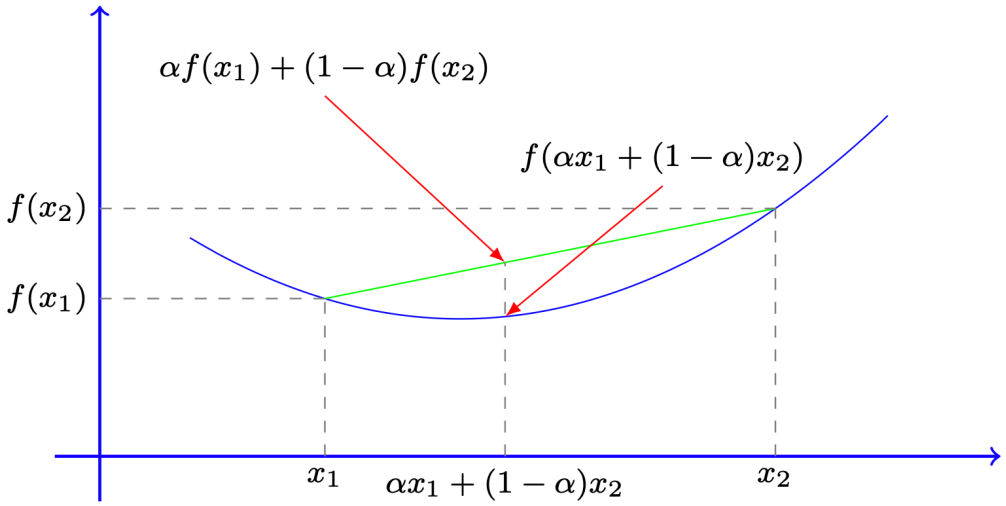
Convex/concave functions

Definition (Convex function, real-valued)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $\{x_1, x_2\} \subset \mathbb{R}^n$ and $\alpha \in [0, 1]$ we have

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2). \tag{1}$$

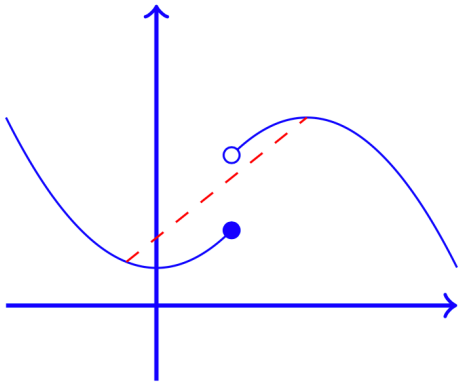
- More generally, convexity of a function presumes convexity of its domain.
- A function f is strictly convex if for $x_1 \neq x_2$ inequality (1) holds strictly.
- f is concave if $-f$ is convex.



Continuity

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then it is continuous.

Intuitively, a discontinuity violates the definition of a convex function.



命题 2.4 设 $f(x)$ 是凸函数，则 $f(x)$ 所有的 α-下水平集 C_α 为凸集.

Examples

Convex functions:

- Affine: $ax + b$ for $a, b \in \mathbb{R}$
- Powers: x^a for $x > 0$ and $a \notin (0, 1)$
- Powers of absolute values: $|x|^a$ for $a \geq 1$
- Exponential: e^{ax} for $a \in \mathbb{R}$
- Negative entropy: $x \log x$ for $x > 0$
- p-norms: $\|x\|_p := (\sum_i |x^i|^p)^{1/p}$ for $p \geq 1$

Concave functions:

- Affine: $ax + b$ for $a, b \in \mathbb{R}$
- Powers: x^a for $x > 0$ and $a \in [0, 1]$
- Logarithms: $\log x$ for $x > 0$

Operations preserving convexity

- Addition:
  If $f_1, \ldots, f_k : \mathbb{R}^n \to \overline{\mathbb{R}}$ are convex and $\alpha_1, \ldots, \alpha_k > 0$, then

$$f(x) = \sum_{i=1}^{k} \alpha_i f_i(x) \quad \text{is convex.}$$

- Maximization:
  If $f_1, \ldots, f_k : \mathbb{R}^n \to \overline{\mathbb{R}}$ are convex, then

$$f(x) = \max\{f_1(x), \ldots, f_k(x)\} \quad \text{is convex.}$$

- Composition:
  If $g : \mathbb{R} \to \overline{\mathbb{R}}$ is nondecreasing and convex, and $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex, then

$$f(x) = g(h(x)) \quad \text{is convex.}$$

Differentiability and convexity

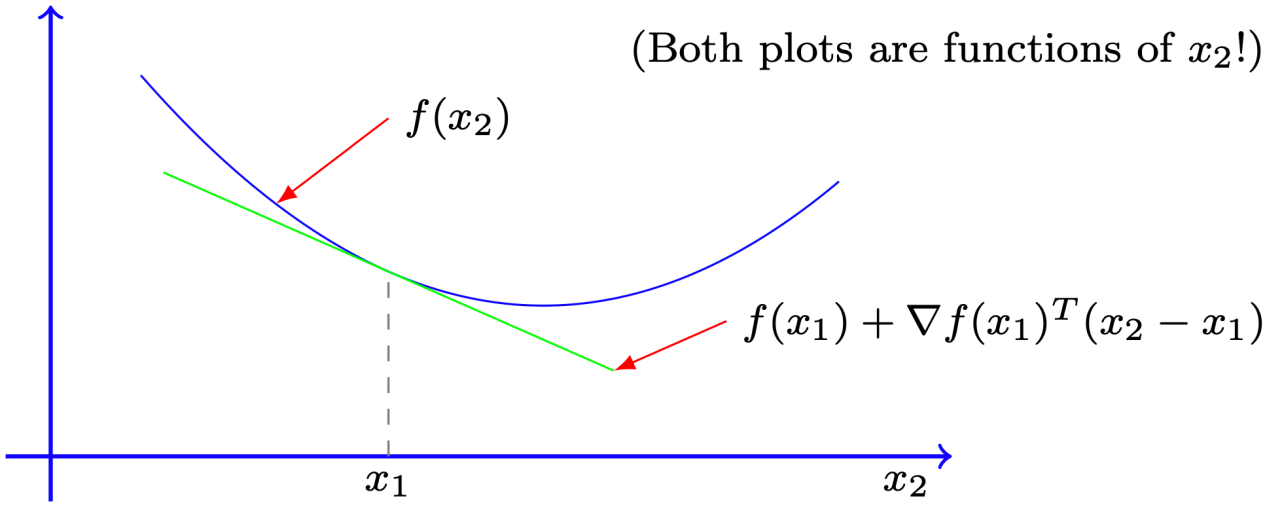
Theorem

Let $\mathcal{X}$ be a nonempty convex subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable over an open set containing $\mathcal{X}$. Then, the following hold:

(a) *f is convex over $\mathcal{X}$ if and only if, for all $\{x_1, x_2\} \subseteq \mathcal{X}$, we have*

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1). \tag{2}$$

(b) *f is strictly convex over $\mathcal{X}$ if and only if (2) is strict when $x_1 \neq x_2$.*



Twice differentiability and convexity

Theorem

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be nonempty and convex and let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open set containing $\mathcal{X}$. Then, the following hold:

(a) *If $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathcal{X}$, then f is convex over $\mathcal{X}$.*

(b) *If $\nabla^2 f(x)$ is positive definite for all $x \in \mathcal{X}$, then f is strictly convex over $\mathcal{X}$.*

(c) *If $\mathcal{X}$ is open and f is convex over $\mathcal{X}$, then $\nabla^2 f(x)$ is p.s.d. for all $x \in \mathcal{X}$.*

Overall, remember the following:

- Convexity implies certain properties of first and second derivatives.
- However, the first and second derivative of a function at a particular point only provides limited information about the function itself.
- For example, even if the Hessian of a function is positive semidefinite at all elements of an infinite sequence $\{x_k\}$, the function is not necessarily convex.

定义 2.17 (强凸函数) 若存在常数 $m > 0$，使得

$$g(x) = f(x) - \frac{m}{2}\|x\|^2$$

为凸函数，则称 $f(x)$ 为**强凸函数**，其中 m 为**强凸参数**. 为了方便我们也称 $f(x)$ 为 m-强凸函数.

引理 2.2 (二次下界) 设 $f(x)$ 是参数为 m 的可微强凸函数，则如下不等式成立：

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{m}{2}\|y - x\|^2, \quad \forall\, x, y \in \mathbf{dom}\, f. \tag{2.5.1}$$

推论 2.4 设 f 为可微强凸函数，则 f 的所有 α-下水平集有界.

Generalizing differentiability

For $f \in \mathcal{C}$, recall that the gradient $\nabla f(x)$ represents the direction along which the function increases at the fastest rate from x.

- The negative gradient $-\nabla f(x)$ represents the direction of fastest decrease.
- Hence, it is known as the steepest descent direction of f at x.

But what if $f \notin \mathcal{C}$?

- Can we generalize our notion of a gradient?
- Can we generalize our notion of a steepest ascent/descent direction?

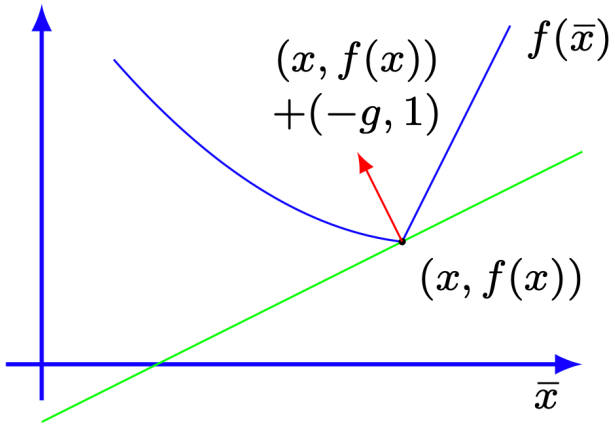
More simply, given f and $x \in \text{dom}(f)$, what can we say about the rate of change in f from x along a given direction d?

Subgradients

Definition (Subgradient)

A vector $g \in \mathbb{R}^n$ is a subgradient of a proper convex f at $x \in \mathrm{dom}(f)$ if

$$f(\overline{x}) \geq f(x) + g^T(\overline{x} - x) \quad \text{for all} \quad \overline{x} \in \mathbb{R}^n.$$



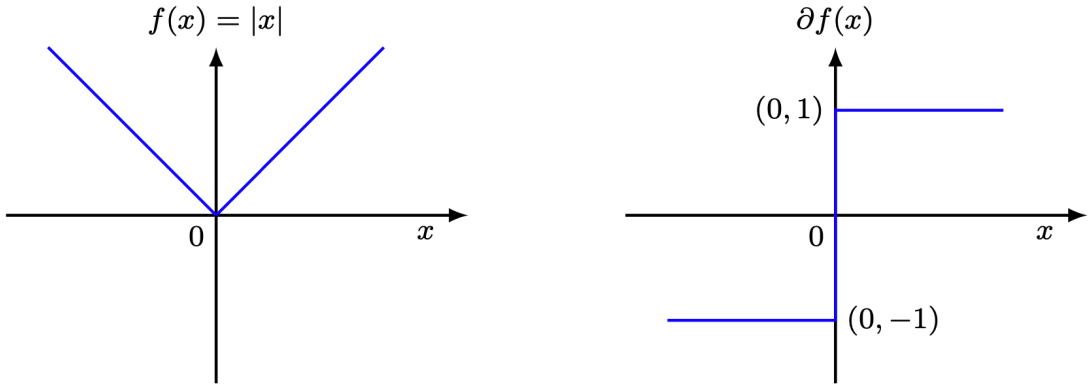
Theorem

If f is differentiable at $x \in \mathrm{int}(\mathrm{dom}(f))$, then $\nabla f(x)$ is its unique subgradient at x.

Subdifferentials

Definition (Subdifferential)

The set of all subgradients of f at x, denoted $\partial f(x)$, is the subdifferential of f at x.



Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued convex function and let $\mathcal{X} \subseteq \mathbb{R}^n$ be nonempty and compact. Then, the following hold.

(a) *The set $\cup_{x \in \mathcal{X}} \partial f(x)$ is nonempty and bounded.*

(b) *f is Lipschitz continuous over $\mathcal{X}$, i.e., there exists $L > 0$ such that*

$$|f(x) - f(\bar{x})| \leq L\|x - \bar{x}\|_2 \quad \text{for all} \quad \{x, \bar{x}\} \subseteq \mathcal{X}.$$

Directional derivatives

Definition (Directional derivative)

Given proper $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, a point $x \in \mathrm{dom}(f)$, and a direction $d \in \mathbb{R}^n$, the directional derivative of f at x in the direction d (if it exists) is

$$f'(d; x) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

- Note that this definition does not require f to be differentiable.
- If f is convex, then for every $x \in \mathrm{dom}(f)$ and $d \in \mathbb{R}^n$, the limit exists.
- If f is convex and $x \in \mathrm{int}(\mathrm{dom}(f))$, then $f'(d; x)$ is finite.
- If $f \in \mathcal{C}$, then $f'(d; x)$ exists and

$$f'(d; x) = \nabla f(x)^T d.$$

Subgradients and directional derivatives

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex.

- If $x \in \mathrm{dom}(f)$, then $g \in \partial f(x)$ if and only if

$$f'(d; x) \geq g^T d \quad \text{for all} \quad d \in \mathbb{R}^n.$$

- If $x \in \mathrm{int}(\mathrm{dom}(f))$, then $\partial f(x)$ is a nonempty, convex, and compact and

$$f'(d; x) = \max_{g \in \partial f(x)} g^T d \quad \text{for all} \quad d \in \mathbb{R}^n.$$

Descent directions

Consider the problem of minimizing a convex function $f : \mathbb{R}^n \to \mathbb{R}$.

- At a point $x \in \mathbb{R}^n$, a descent direction d is one for which we have

$$\sup_{g \in \partial f(x)} g^T d = f'(d; x) < 0.$$

- We can decrease f by moving (a small distance) along such a direction d.

Along which direction is f decreasing at the fastest rate?

- The steepest descent direction is the solution of the optimization problem

$$\min_{d \in \mathbb{R}^n} f'(d; x) \quad \text{s.t.} \quad \|d\|_2 \leq 1.$$

- For proper convex f and $x \in \text{int}(\text{dom}(f))$, we find that

$$\begin{aligned} \min_{\|d\|_2 \leq 1} f'(d; x) &= \min_{\|d\|_2 \leq 1} \max_{g \in \partial f(x)} g^T d \\ &= \max_{g \in \partial f(x)} \min_{\|d\|_2 \leq 1} g^T d = \max_{g \in \partial f(x)} (-\|g\|_2) = - \min_{g \in \partial f(x)} \|g\|_2, \end{aligned}$$

That is, in such a case, the steepest descent direction is $d = -g/\|g\|_2$ where g is the minimum norm element in $\partial f(x)$.