

Lecture 7: Preliminaries III. Optimization



School of Mathematical Sciences, Xiamen University

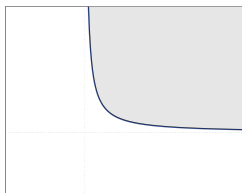
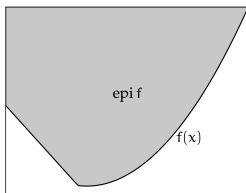
1. Basic definitions

- The *effective domain* of $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is defined as

$$\text{dom}(f) := \{\mathbf{x} \mid f(\mathbf{x}) < +\infty\}.$$

- A function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is called *proper* if there exists at least one $\mathbf{x} \in \mathbb{R}^n$ such that $f(\mathbf{x}) < +\infty$, meaning that $\text{dom}(f) \neq \emptyset$.
- The *epigraph* of $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is defined by

$$\text{epi}(f) = \{(\mathbf{x}, y) : f(\mathbf{x}) \leq y, \mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R}\}.$$



- A function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is called *closed* if $\text{epi}(f)$ is closed.

2. Solutions of $\min_{\mathbf{x}} f(\mathbf{x})$

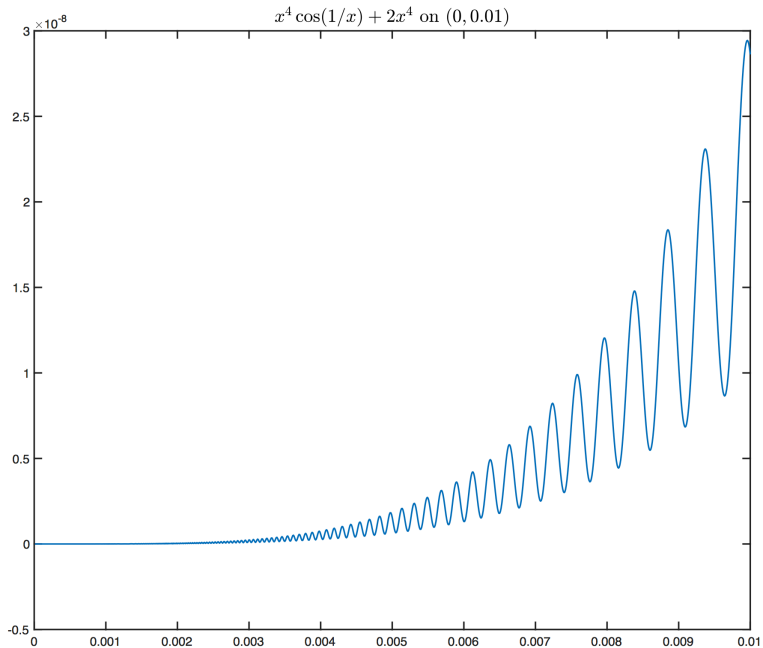
- \mathbf{x}_\star is a *local minimizer* of f if there is a neighborhood \mathcal{N} of \mathbf{x}_\star such that $f(\mathbf{x}) \geq f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathcal{N}$.
- \mathbf{x}_\star is a *strict local minimizer* if it is a local minimizer on some neighborhood \mathcal{N} and in addition $f(\mathbf{x}) > f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathcal{N}$ with $\mathbf{x} \neq \mathbf{x}_\star$.
- \mathbf{x}_\star is an *isolated local minimizer* if there is a neighborhood \mathcal{N} of \mathbf{x}_\star such that $f(\mathbf{x}) \geq f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathcal{N}$ and in addition, \mathcal{N} contains no local minimizers other than \mathbf{x}_\star .

Strict local minimizers are not always isolated: for example,

$$f(x) = x^4 \cos(1/x) + 2x^4, \quad f(0) = 0.$$

All isolated local minimizers are strict.

- \mathbf{x}_\star is a *global minimizer* of f if $f(\mathbf{x}) \geq f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathbb{R}^n$.



3. Convexity

- A set $\Omega \subseteq \mathbb{R}^n$ is called *convex* if it has the property that

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \Rightarrow (1 - \alpha)\mathbf{x} + \alpha\mathbf{y} \in \Omega \quad \forall \alpha \in [0, 1].$$

We usually deal with closed convex sets.

- For a convex set $\Omega \subseteq \mathbb{R}^n$ we define the *indicator function* I_Ω as follows

$$I_\Omega(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in \Omega \\ +\infty & \text{otherwise.} \end{cases}$$

The constrained optimization problem

$$\min_{\mathbf{x} \in \Omega} f(\mathbf{x})$$

can be restated equivalently as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + I_\Omega(\mathbf{x}).$$

- A convex function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ has the following defining property: $\text{dom}(f)$ is convex, and $\forall \mathbf{x}, \mathbf{y} \in \text{dom}(f), \forall \alpha \in [0, 1]$,

$$f((1 - \alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1 - \alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Theorem 1 (First-order convexity condition)

Differentiable f is convex if and only if $\text{dom}(f)$ is convex and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f).$$

Theorem 2 (Second-order convexity conditions)

Assume f is twice continuously differentiable. Then f is convex if and only if $\text{dom}(f)$ is convex and

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}, \quad \forall \mathbf{x} \in \text{dom}(f)$$

that is, $\nabla^2 f(\mathbf{x})$ is positive semidefinite.

- Important properties for convex objective functions:
 - ★ Any local minimizer is also a global minimizer (see Theorem 12).
 - ★ The set of global minimizers is a convex set. (easy to prove)
- If there exists a value $\gamma > 0$ such that

$$f((1 - \alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1 - \alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\gamma}{2}\alpha(1 - \alpha)\|\mathbf{x} - \mathbf{y}\|_2^2$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, we say that f is *strongly convex with modulus of convexity* γ .

- For differentiable f : **Equivalent** definition of *strongly convex with modulus of convexity* γ

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\gamma}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

4. Subgradient and subdifferential

- Definition: A vector $\mathbf{v} \in \mathbb{R}^n$ is a *subgradient* of f at a point \mathbf{x} if for all $\mathbf{d} \in \mathbb{R}^n$, it holds

$$f(\mathbf{x} + \mathbf{d}) \geq f(\mathbf{x}) + \mathbf{v}^\top \mathbf{d}.$$

The *subdifferential*, denoted $\partial f(\mathbf{x})$, is the set of all subgradients of f at \mathbf{x} . (see FOMO for concrete examples)

Lemma 3 (Monotonicity of subdifferentials of convex functions)

\forall convex f , if $\mathbf{a} \in \partial f(\mathbf{x})$ and $\mathbf{b} \in \partial f(\mathbf{y})$, we have $(\mathbf{a} - \mathbf{b})^\top (\mathbf{x} - \mathbf{y}) \geq 0$.

Proof. By the definition of subgradient, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{a}^\top (\mathbf{y} - \mathbf{x}) \quad \text{and} \quad f(\mathbf{x}) \geq f(\mathbf{y}) + \mathbf{b}^\top (\mathbf{x} - \mathbf{y}).$$

Adding these two inequalities yields the statement. □

Theorem 4 (Fermat's lemma: generalization in convex functions)

Let $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Then the point \mathbf{x}_\star is a minimizer of $f(\mathbf{x})$ if and only if

$$\mathbf{0} \in \partial f(\mathbf{x}_\star).$$

Proof. “ \Leftarrow ”: Suppose that $\mathbf{0} \in \partial f(\mathbf{x}_\star)$. We have

$$f(\mathbf{x}_\star + \mathbf{d}) \geq f(\mathbf{x}_\star) \quad \forall \mathbf{d} \in \mathbb{R}^n,$$

which implies that \mathbf{x}_\star is a minimizer of f . “ \Rightarrow ”: by definition. □

Theorem 5

Let $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ be proper convex, and let $\mathbf{x} \in \text{int}(\text{dom}(f))$.

- If f is differentiable at \mathbf{x} , then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.
- If $\partial f(\mathbf{x})$ is a singleton (a set containing a single vector), then f is differentiable at \mathbf{x} with gradient equal to the unique subgradient.

5. Taylor's theorem

- Taylor's theorem shows how smooth functions can be locally approximated by low-order (e.g., linear or quadratic) functions.

定理 12.3.1 (Taylor 公式) 设 $f(x, y)$ 在点 (x_0, y_0) 的邻域 $U = O((x_0, y_0), r)$ 上具有 $k+1$ 阶连续偏导数, 那么对于 U 内每一点 $(x_0 + \Delta x, y_0 + \Delta y)$ 都成立

$$\begin{aligned} & f(x_0 + \Delta x, y_0 + \Delta y) \\ &= f(x_0, y_0) + \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right) f(x_0, y_0) \\ & \quad + \frac{1}{2!} \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right)^2 f(x_0, y_0) + \cdots \\ & \quad + \frac{1}{k!} \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right)^k f(x_0, y_0) + R_k, \end{aligned}$$

其中 $R_k = \frac{1}{(k+1)!} \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right)^{k+1} f(x_0 + \theta \Delta x, y_0 + \theta \Delta y) (0 < \theta < 1)$
称为 **Lagrange 余项**.

Theorem 6 (Taylor's theorem)

Given a continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$, and given $\mathbf{x}, \mathbf{p} \in \mathbb{R}^n$, we have that

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + \int_0^1 \nabla f(\mathbf{x} + t\mathbf{p})^\top \mathbf{p} dt,$$

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + \xi\mathbf{p})^\top \mathbf{p}, \text{ for some } \xi \in (0, 1).$$

If f is twice continuously differentiable, we have

$$\nabla f(\mathbf{x} + \mathbf{p}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{p}) \mathbf{p} dt,$$

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{p} + \frac{1}{2} \mathbf{p}^\top \nabla^2 f(\mathbf{x} + \xi\mathbf{p}) \mathbf{p},$$

for some $\xi \in (0, 1)$.

- If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable and convex, then

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$$

and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

- Lipschitz continuously differentiable with constant L :

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

If f is Lipschitz continuously differentiable with constant L , then by Taylor's theorem, we have

$$f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

- For differentiable f : **Equivalent** definition of *strongly convex with modulus of convexity* γ

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\gamma}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Lemma 7

Suppose f is strongly convex with modulus of convexity γ and ∇f is uniformly Lipschitz continuous with constant L . We have $\forall \mathbf{x}, \mathbf{y}$ that

$$\frac{\gamma}{2} \|\mathbf{y} - \mathbf{x}\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

- Condition number $\kappa := L/\gamma$. (e.g., strictly convex, quadratic f)
- When f is twice continuously differentiable, the inequalities in Lemma 7 is **equivalent** to

$$\gamma \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L \mathbf{I}, \text{ for all } \mathbf{x}.$$

Theorem 8

Let f be differentiable and strongly convex with modulus of convexity $\gamma > 0$. Then the minimizer \mathbf{x}_\star of f exists and is unique.

Proof. (i) Compactness of level set: Show that for any point \mathbf{x}^0 , the level set

$$\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$$

is closed and bounded, and hence compact.

(ii) Existence: Since f is continuous, it attains its minimum on the compact level set, which is also the solution of $\min_{\mathbf{x}} f(\mathbf{x})$.

(iii) Uniqueness: Suppose for contradiction that the minimizer is not unique, so that we have two points \mathbf{x}_\star^1 and \mathbf{x}_\star^2 that minimize f . By using the strongly convex property, we can prove

$$f\left(\frac{\mathbf{x}_\star^1 + \mathbf{x}_\star^2}{2}\right) < f(\mathbf{x}_\star^1) = f(\mathbf{x}_\star^2).$$

This is a contradiction. □

6. Optimality conditions for smooth functions

Theorem 9 (First-order necessary condition)

If \mathbf{x}_\star is a local minimizer of f and f is continuously differentiable in an open neighborhood of \mathbf{x}_\star , then $\nabla f(\mathbf{x}_\star) = \mathbf{0}$.

Proof. Suppose for contradiction that $\nabla f(\mathbf{x}_\star) \neq \mathbf{0}$. Define the vector $\mathbf{p} = -\nabla f(\mathbf{x}_\star)$ and note that $\mathbf{p}^\top \nabla f(\mathbf{x}_\star) = -\|\nabla f(\mathbf{x}_\star)\|_2^2 < 0$. Because ∇f is continuous near \mathbf{x}_\star , there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla f(\mathbf{x}_\star + t\mathbf{p}) < 0, \quad \text{for all } t \in [0, T].$$

For any $s \in (0, T]$, we have by Taylor's theorem that

$$f(\mathbf{x}_\star + s\mathbf{p}) = f(\mathbf{x}_\star) + s\mathbf{p}^\top \nabla f(\mathbf{x}_\star + \xi s\mathbf{p}) \quad \text{for some } \xi \in (0, 1).$$

Therefore, $f(\mathbf{x}_\star + s\mathbf{p}) < f(\mathbf{x}_\star)$ for all $s \in (0, T]$. We have found a direction leading away from \mathbf{x}_\star along which f decreases, so \mathbf{x}_\star is not a local minimizer, and we have a contradiction. \square

Theorem 10 (Second-order necessary conditions)

If \mathbf{x}_\star is a local minimizer of f and $\nabla^2 f$ is continuous in an open neighborhood of \mathbf{x}_\star , then $\nabla f(\mathbf{x}_\star) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}_\star)$ is positive semidefinite.

Proof. We know from Theorem 9 that $\nabla f(\mathbf{x}_\star) = \mathbf{0}$. Assume that $\nabla^2 f(\mathbf{x}_\star)$ is not positive semidefinite. Then we can choose a vector \mathbf{p} such that $\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star) \mathbf{p} < 0$, and because $\nabla^2 f$ is continuous near \mathbf{x}_\star , there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + t\mathbf{p}) \mathbf{p} < 0, \quad \text{for all } t \in [0, T].$$

By doing a Taylor series expansion around \mathbf{x}_\star , we have for all $s \in (0, T]$ and some $\xi \in (0, 1)$ that

$$f(\mathbf{x}_\star + s\mathbf{p}) = f(\mathbf{x}_\star) + s\mathbf{p}^\top \nabla f(\mathbf{x}_\star) + \frac{1}{2}s^2 \mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi s\mathbf{p}) \mathbf{p} < f(\mathbf{x}_\star).$$

As in Theorem 9, we have found a direction from \mathbf{x}_\star along which f is decreasing, and so again, \mathbf{x}_\star is not a local minimizer. \square

Theorem 11 (Second-order sufficient conditions)

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of \mathbf{x}_\star and that $\nabla f(\mathbf{x}_\star) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}_\star)$ is positive definite. Then \mathbf{x}_\star is a strict local minimizer of f .

Proof. Because the Hessian $\nabla^2 f$ is continuous and positive definite at \mathbf{x}_\star , we can choose a radius $r > 0$ so that $\nabla^2 f(\mathbf{x})$ remains positive definite for all \mathbf{x} in the open ball $\mathcal{B} = \{\mathbf{z} \mid \|\mathbf{z} - \mathbf{x}_\star\| < r\}$. Taking any nonzero vector \mathbf{p} with $\|\mathbf{p}\| < r$, we have $\mathbf{x}_\star + \mathbf{p} \in \mathcal{B}$ and

$$\begin{aligned} f(\mathbf{x}_\star + \mathbf{p}) &= f(\mathbf{x}_\star) + \mathbf{p}^\top \nabla f(\mathbf{x}_\star) + \frac{1}{2} \mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi \mathbf{p}) \mathbf{p} \\ &= f(\mathbf{x}_\star) + \frac{1}{2} \mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi \mathbf{p}) \mathbf{p}, \end{aligned}$$

for some $\xi \in (0, 1)$. Since $\mathbf{x}_\star + \xi \mathbf{p} \in \mathcal{B}$, we have $\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi \mathbf{p}) \mathbf{p} > 0$, and therefore $f(\mathbf{x}_\star + \mathbf{p}) > f(\mathbf{x}_\star)$, giving the result. \square

- A point \mathbf{x} is called a *stationary point* if

$$\nabla f(\mathbf{x}) = \mathbf{0}.$$

- A stationary point \mathbf{x} is called a *saddle point* if there exist \mathbf{u} and \mathbf{v} such that

$$f(\mathbf{x} + \alpha \mathbf{u}) < f(\mathbf{x}) \quad \text{and} \quad f(\mathbf{x} + \alpha \mathbf{v}) > f(\mathbf{x})$$

for all sufficiently small $\alpha > 0$.

- Stationary points are not necessarily local minimizers. Stationary points can be *local maximizers* or *saddle points*.
- If $\nabla f(\mathbf{x}) = \mathbf{0}$, and $\nabla^2 f(\mathbf{x})$ has both strictly positive and strictly negative eigenvalues, then \mathbf{x} is a saddle point.
- If $\nabla^2 f(\mathbf{x})$ is positive semidefinite or negative semidefinite, then $\nabla^2 f(\mathbf{x})$ alone is insufficient to classify \mathbf{x} .

Theorem 12

- (i) \forall convex f , any local minimizer \mathbf{x}_\star is a global minimizer of f .
- (ii) If f is convex and differentiable, then any stationary point \mathbf{x}_\star is a global minimizer of f .

Proof. (i) Suppose that \mathbf{x}_\star is a local but not a global minimizer. Then we can find a point $\mathbf{z} \in \mathbb{R}^n$ with $f(\mathbf{z}) < f(\mathbf{x}_\star)$. Consider the line segment that joins \mathbf{x}_\star to \mathbf{z} , that is,

$$\mathbf{x} = \lambda \mathbf{z} + (1 - \lambda) \mathbf{x}_\star, \quad \text{for some } \lambda \in (0, 1].$$

By the convexity property for f , we have

$$f(\mathbf{x}) \leq \lambda f(\mathbf{z}) + (1 - \lambda) f(\mathbf{x}_\star) < f(\mathbf{x}_\star).$$

Any neighborhood \mathcal{N} of \mathbf{x}_\star contains a piece of the line segment, so there will always be points $\mathbf{x} \in \mathcal{N}$ at which the last inequality is satisfied. Hence, \mathbf{x}_\star is not a local minimizer.

(ii) Suppose that \mathbf{x}_\star is not a global minimizer and choose \mathbf{z} as above. Then, from convexity, we have

$$\begin{aligned}\nabla f(\mathbf{x}_\star)^\top (\mathbf{z} - \mathbf{x}_\star) &= \left. \frac{d}{d\lambda} f(\mathbf{x}_\star + \lambda(\mathbf{z} - \mathbf{x}_\star)) \right|_{\lambda=0} \\ &= \lim_{\lambda \rightarrow 0^+} \frac{f(\mathbf{x}_\star + \lambda(\mathbf{z} - \mathbf{x}_\star)) - f(\mathbf{x}_\star)}{\lambda} \\ &\leq \lim_{\lambda \rightarrow 0^+} \frac{\lambda f(\mathbf{z}) + (1 - \lambda)f(\mathbf{x}_\star) - f(\mathbf{x}_\star)}{\lambda} \\ &= f(\mathbf{z}) - f(\mathbf{x}_\star) < 0.\end{aligned}$$

Therefore, $\nabla f(\mathbf{x}_\star) \neq \mathbf{0}$, and so \mathbf{x}_\star is not a stationary point. □

- *Remark:* Theorems 9-12 provide the foundations for unconstrained optimization algorithms.
- Numerical algorithms try to seek a point where ∇f vanishes.

7. Karush–Kuhn–Tucker conditions

Theorem 13 (KKT conditions)

Consider the minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1 : m,$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and all $g_i : \mathbb{R}^n \mapsto \mathbb{R}$ are convex functions.

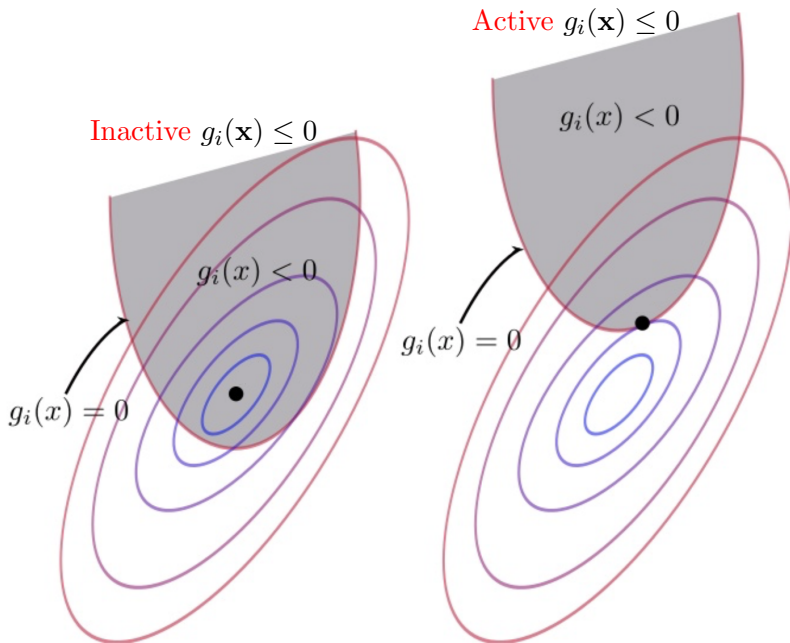
- Let \mathbf{x}_\star be an optimal solution and assume Slater's condition

$$\exists \mathbf{x} \in \mathbb{R}^n, \quad \text{s.t.} \quad g_i(\mathbf{x}) < 0, \quad i = 1 : m,$$

hold. Then there exist $\lambda_1, \dots, \lambda_m \geq 0$ satisfying

$$\mathbf{0} \in \partial f(\mathbf{x}_\star) + \sum_{i=1}^m \lambda_i \partial g_i(\mathbf{x}_\star), \quad \lambda_i g_i(\mathbf{x}_\star) = 0, \quad i = 1 : m.$$

- If \mathbf{x}_\star satisfies the above *conditions*, called **KKT conditions**, then it is an optimal solution of the optimization problem.



8. Moreau envelope and proximal operator

- For a closed proper convex function $h : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ and a positive scalar λ , the *Moreau envelope* of (λ, h) is

$$\begin{aligned} M_{\lambda, h}(\mathbf{x}) &:= \inf_{\mathbf{u}} \left\{ h(\mathbf{u}) + \frac{1}{2\lambda} \|\mathbf{u} - \mathbf{x}\|^2 \right\} \\ &= \frac{1}{\lambda} \inf_{\mathbf{u}} \left\{ \lambda h(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right\}. \end{aligned}$$

- For a closed proper convex function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$, the *proximal operator* of f is

$$\text{prox}_f(\mathbf{x}) := \underset{\mathbf{u}}{\text{argmin}} \left\{ f(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right\}.$$

From the optimality condition (see Theorem 4), we have

$$\mathbf{0} \in \partial f(\text{prox}_f(\mathbf{x})) + (\text{prox}_f(\mathbf{x}) - \mathbf{x}).$$

- For a closed proper convex function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$, the point \mathbf{x}_\star is a minimizer of f if and only if

$$\mathbf{x}_\star = \text{prox}_f(\mathbf{x}_\star).$$

- For a closed proper convex function $h : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ and a positive scalar λ , the Moreau envelope

$$M_{\lambda,h}(\mathbf{x}) = h(\text{prox}_{\lambda h}(\mathbf{x})) + \frac{1}{2\lambda} \|\text{prox}_{\lambda h}(\mathbf{x}) - \mathbf{x}\|^2,$$

can be viewed as a kind of smoothing or regularization of the function h . We have for all $\mathbf{x} \in \mathbb{R}^n$,

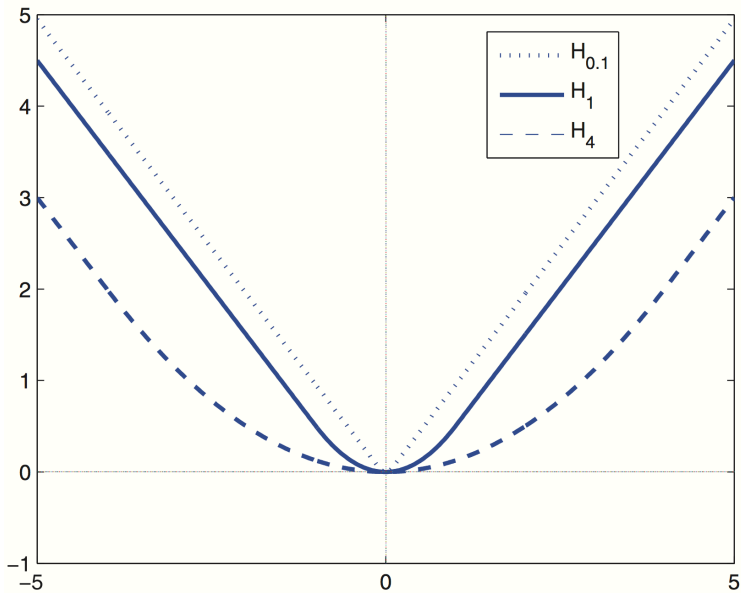
$$-\infty < M_{\lambda,h}(\mathbf{x}) < +\infty$$

even when $h(\mathbf{x}) = +\infty$ for some $\mathbf{x} \in \mathbb{R}^n$. Moreover, $M_{\lambda,h}(\mathbf{x})$ is [convex and differentiable](#) everywhere with gradient

$$\nabla M_{\lambda,h}(\mathbf{x}) = \frac{1}{\lambda}(\mathbf{x} - \text{prox}_{\lambda h}(\mathbf{x})).$$

Therefore, \mathbf{x}_\star is a minimizer of $h \Leftrightarrow \mathbf{x}_\star$ is a minimizer of $M_{\lambda,h}(\mathbf{x})$.

- Example: $h(x) = |x|$ and $H_\lambda(x) := M_{\lambda,h}(x)$.



Lemma 14 (Nonexpansivity of proximal operator)

For a closed proper convex function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$, we have for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|.$$

Proof. From the optimality conditions at two points \mathbf{x} and \mathbf{y} , we have

$$\mathbf{x} - \text{prox}_f(\mathbf{x}) \in \partial f(\text{prox}_f(\mathbf{x})) \quad \text{and} \quad \mathbf{y} - \text{prox}_f(\mathbf{y}) \in \partial f(\text{prox}_f(\mathbf{y})).$$

By applying monotonicity (see Lemma 3), we have

$$((\mathbf{x} - \text{prox}_f(\mathbf{x})) - (\mathbf{y} - \text{prox}_f(\mathbf{y})))^\top (\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})) \geq 0.$$

Rearranging this and applying the Cauchy-Schwartz inequality yields

$$\begin{aligned} \|\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})\|^2 &\leq (\mathbf{x} - \mathbf{y})^\top (\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})) \\ &\leq \|\mathbf{x} - \mathbf{y}\| \|\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})\|. \quad \square \end{aligned}$$

- Examples of several proximal operators

(1) $f(\mathbf{x}) = 0$:

$$\text{prox}_f(\mathbf{x}) = \mathbf{x}.$$

(2) $f(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$ with $\lambda > 0$:

$$\begin{aligned} [\text{prox}_{\lambda \|\cdot\|_1}(\mathbf{x})]_i &= \underset{u \in \mathbb{R}}{\operatorname{argmin}} \left\{ \lambda |u| + \frac{1}{2} (u - x_i)^2 \right\} \\ &= \begin{cases} x_i - \lambda & \text{if } x_i > \lambda, \\ 0 & \text{if } x_i \in [-\lambda, \lambda], \\ x_i + \lambda & \text{if } x_i < -\lambda, \end{cases} \end{aligned}$$

which is known as *soft-thresholding*.

(3) $f(\mathbf{x}) = \lambda \|\mathbf{x}\|_0$: the number of nonzero components, **non-convex**

$$[\text{prox}_{\lambda \|\cdot\|_0}(\mathbf{x})]_i = \begin{cases} x_i & \text{if } |x_i| > \sqrt{2\lambda}, \\ \{0, x_i\} & \text{if } |x_i| = \sqrt{2\lambda}, \\ 0 & \text{if } |x_i| < \sqrt{2\lambda}, \end{cases}$$

which is known as *hard thresholding*.

(4) $f(\mathbf{x}) = I_\Omega(\mathbf{x})$:

$$\text{prox}_{I_\Omega}(\mathbf{x}) = \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ I_\Omega(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right\} = \underset{\mathbf{u} \in \Omega}{\operatorname{argmin}} \|\mathbf{u} - \mathbf{x}\|,$$

which is simply the projection of \mathbf{x} onto the set Ω .