Numerical Optimization

Lecture 11: Unconstrained Optimization Theory

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

Outline

Problem and Definitions

First-Order Conditions

Second-Order Conditions

Outline

Problem and Definitions

First-Order Conditions

Second-Order Conditions

Unconstrained optimization

Consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x).$$

In these notes, we discuss situations when $f : \mathbb{R}^n \to \mathbb{R}$ is

- continuously differentiable (i.e., $f \in \mathcal{C}$),
- twice continuously differentiable (i.e., $f \in \mathcal{C}^2$), or, instead,
- convex (with, perhaps, $f \notin \mathcal{C}$).

Global and local minima

Ideal minima are those that minimize a function globally over its domain.

Definition 3.1.1 (Global minimum)

A vector x_* is a global minimum of f if

$$f(x_*) \leq f(x) \quad \text{for all} \quad x \in \mathbb{R}^n.$$

Commonly, however, we are satisfied with a weaker form of minimum.

Definition 3.1.2 (Local minimum)

A vector x_* is a local minimum of f if there exists $\epsilon > 0$ such that

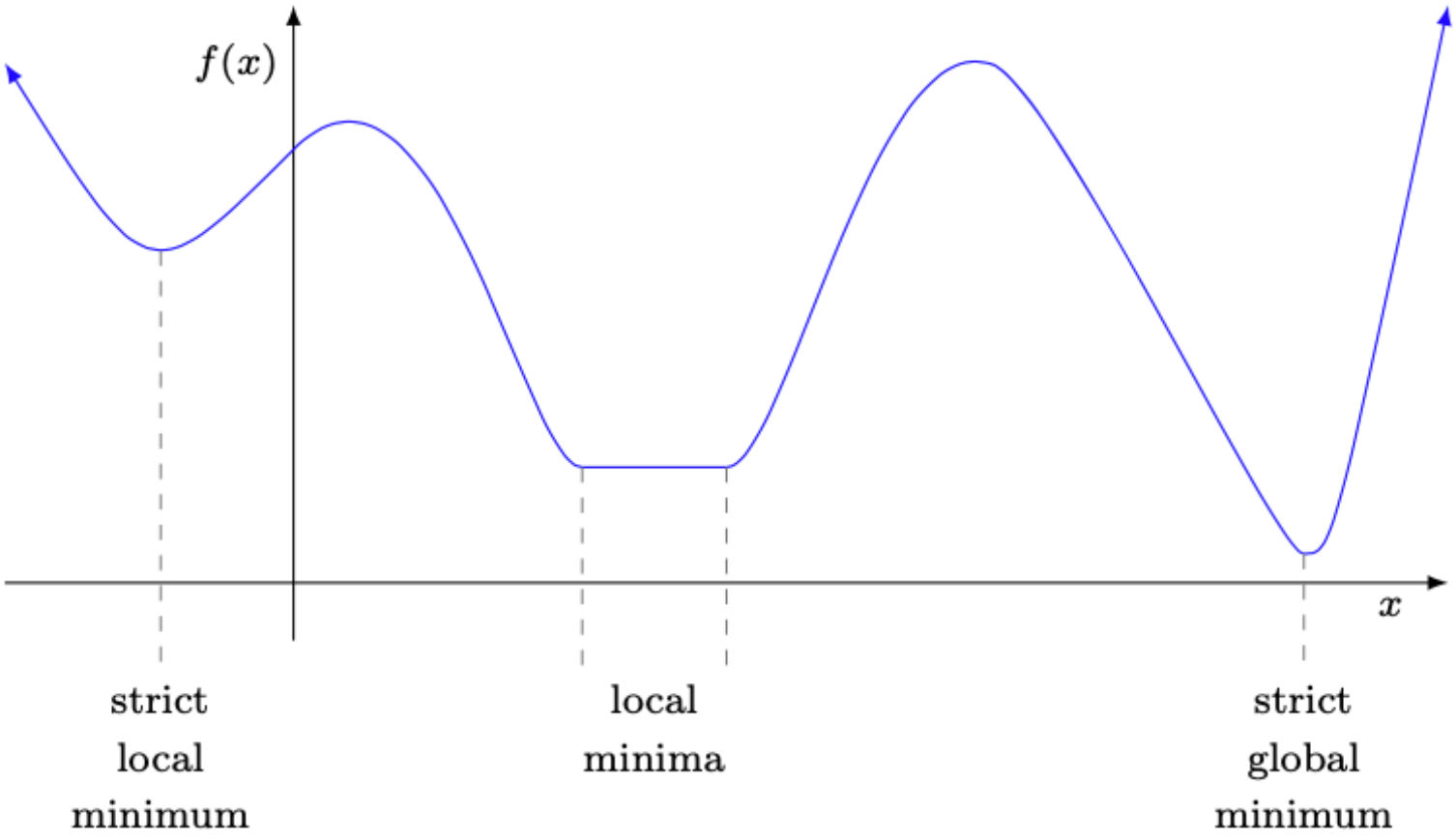
$$f(x_*) \leq f(x) \quad \text{for all} \quad x \in \mathbb{B}(x_*, \epsilon) := \{x \in \mathbb{R}^n : \|x - x_*\|_2 \leq \epsilon\}.$$

We also characterize certain types of global and/or local minima:

- x_* is a strict global/local minimizer if the inequality holds strictly for $x \neq x_*$.
- x_* is an isolated global/local minimizer if, for some $\epsilon' > 0$, it is the only local minimizer in the neighborhood $\mathbb{B}(x_*, \epsilon')$.

An isolated minimum is a strict minimum, but (typically only for some pathological examples) the reverse is not always true.

Illustration



Local ⇒ global minimum in convex optimization

A special fact in convex optimization is that all local minima are global minima.

Theorem 3.1.3

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then a local minimum of f is a global minimum of f. If f is strictly convex, then there exists at most one global minimum of f.

Proof.

To derive a contradiction, suppose that x_* is a local minimum of f that is not a global minimum. Then, there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < f(x_*)$. By convexity of f, we have for all $\alpha \in (0, 1)$ that

$$f(\alpha x_* + (1-\alpha)\bar{x}) \leq \alpha f(x_*) + (1-\alpha) f(\bar{x}) < f(x_*).$$

This means that f has a value strictly lower than $f(x_*)$ at every point on the line segment $(x_*, \bar{x}]$, which violates the local minimality of x_*. (The statement about strictly convex f can be proved in a similar manner.)

Global vs. local minimization

- Unfortunately, for nonconvex optimization, the conditions in the definitions of global and local minima are not entirely useful.
- Unless we can verify strict quasiconvexity, we rarely have global information about f, and so have no way to verify if a point is a global minimizer.
- Thus, in nonconvex optimization, we often focus on finding a local minimizer.
- Using calculus, we can derive local optimality conditions that aid in determining if a point is a local minimizer.
- In this manner, we rarely (if ever) use the aforementioned definitions directly.

Mean Value Theorem and Taylor's Theorem

Our primary tools in developing optimality conditions are the following.

Theorem 3.1.4 (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 3.1.5 (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

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

Outline

Problem and Definitions

First-Order Conditions

Second-Order Conditions

First-order necessary condition

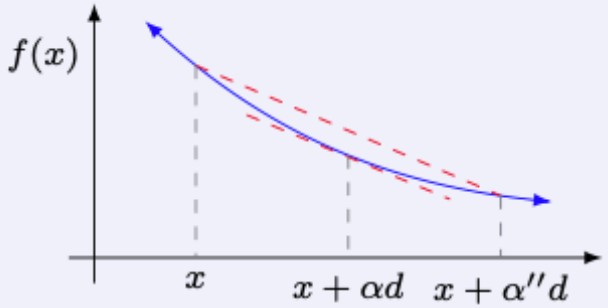
Theorem 3.2.1 (First-order necessary condition)

If $f \in \mathcal{C}$ and x_* is a local minimizer of f, then $\nabla f(x_*) = 0$.

Proof.

For $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$, let $d = -\nabla f(x)$ (with $\nabla f(x)^T d = -\|\nabla f(x)\|_2^2 < 0$). Since ∇f is continuous, there exists $\alpha' > 0$ such that $d^T \nabla f(x + \alpha d) < 0$ for all $\alpha \in [0, \alpha']$, i.e., the directional derivative remains negative some way along d. By the Mean Value Theorem (3.1.4), for any $\alpha'' \in (0, \alpha']$ we have

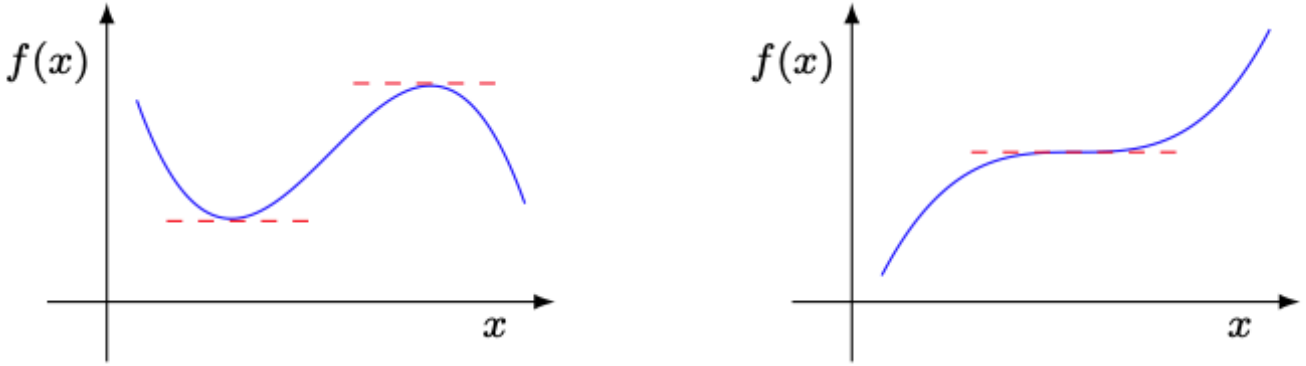
$$f(x + \alpha'' d) = f(x) + \alpha'' d^T \nabla f(x + \alpha d) \quad \text{for some} \quad \alpha \in (0, \alpha'').$$



Thus, $f(x + \alpha'' d) < f(x)$ for all $\alpha'' \in (0, \alpha']$.

Stationary points

- We can limit our search to points where $\nabla f(x_*) = 0$.
- However, $\nabla f(x_*) = 0$ does not imply that we have a local minimizer!



- At least we know that if $\nabla f(x) \neq 0$, then x is not a local minimizer.

Definition 3.2.2 (Stationary point)

A point $x \in \mathbb{R}^n$ is a stationary point for $f \in \mathcal{C}$ if $\nabla f(x) = 0$.

Convex optimization

If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex (but not necessarily real-valued or differentiable), then we have the following stronger result.

Theorem 3.2.3 (First-order necessary and sufficient condition)

If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex and $0 \in \partial f(x_*)$, then x_* is a global minimizer of f.

In fact, we can say more to characterize the solution set of a convex problem...

Conjugate functions

Consider an extended real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$.

Definition 3.2.4 (Conjugate function)

The conjugate of f is the function $f^\star : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$f^\star(y) = \sup_{x \in \mathbb{R}^n} (y^T x - f(x)).$$

Regardless of the structure of f, we have the following properties.

- $f^\star$ is a closed convex function since it is the pointwise supremum of a collection of affine functions.
- $f^\star$ need not be proper, even if f is proper.

Visualizing the conjugate

- The crossing height on the vertical axis with the hyperplane with normal $(-y, 1)$ that passes through $(\bar{x}, f(\bar{x}))$ is

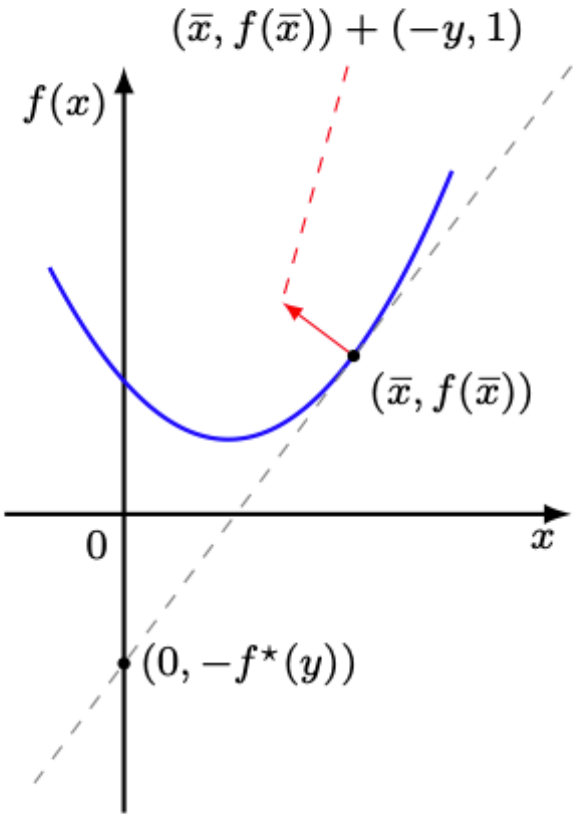
$$f(\bar{x}) - y^T \bar{x}.$$

- Thus, the crossing point corresponding to the hyperplane that supports $\text{epi}(f)$ is

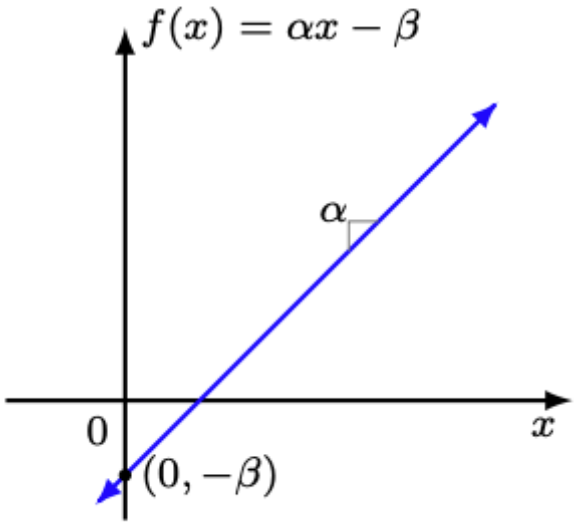
$$\inf_{x \in \mathbb{R}^n} (f(x) - y^T x) = -f^\star(y).$$

- It may help to consider the vector form

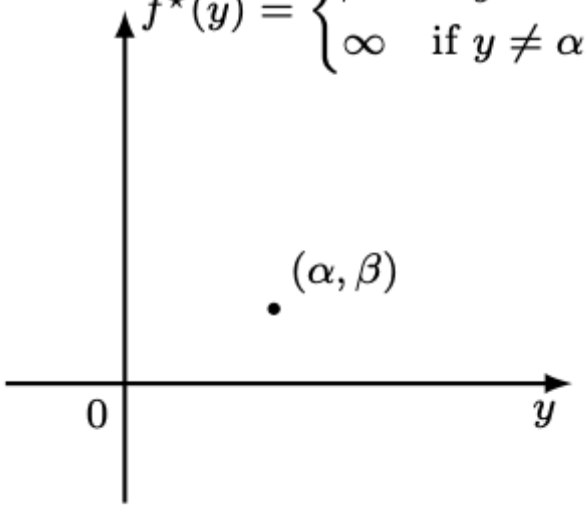
$$-f^\star(y) = \inf_{x \in \mathbb{R}^n} \left(\begin{bmatrix} x \\ f(x) \end{bmatrix}^T \begin{bmatrix} -y \\ 1 \end{bmatrix} \right).$$



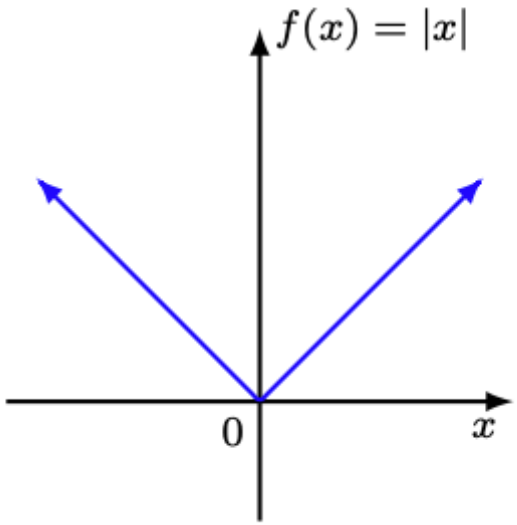
Example: $f^\star(y) = \sup_{x \in \mathbb{R}^n}(y^T x - f(x))$

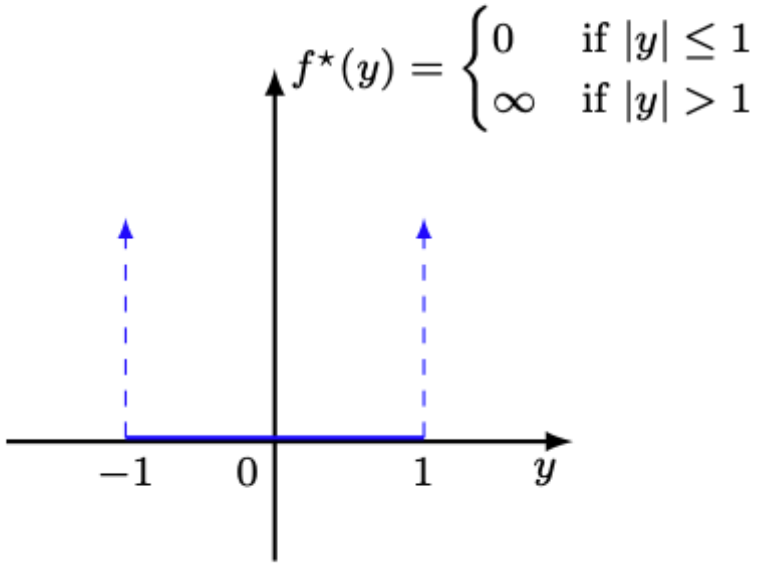


$$f^\star(y) = \begin{cases} \beta & \text{if } y = \alpha \\ \infty & \text{if } y \neq \alpha \end{cases}$$

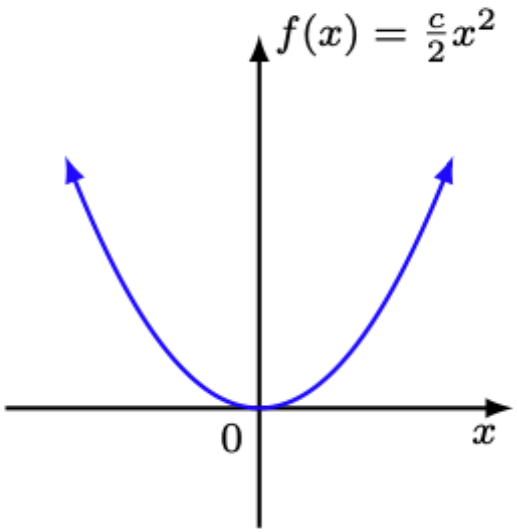


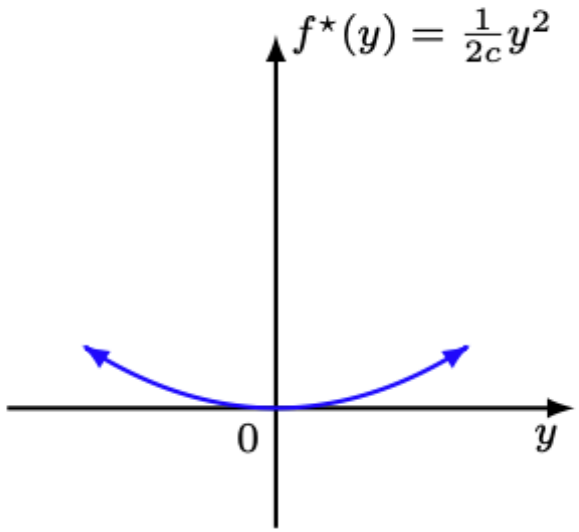
Example: $f^\star(y) = \sup_{x \in \mathbb{R}^n} (y^T x - f(x))$





Example: $f^\star(y) = \sup_{x \in \mathbb{R}^n} (y^T x - f(x))$





Minima of convex functions

The following result summarizes facts about the set of minima of convex f.

Theorem 3.2.5

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function with conjugate $f^\star : \mathbb{R}^n \to \overline{\mathbb{R}}$. Then, with $\mathcal{X}_*$ defined as the set of minima of f over $\mathbb{R}^n$, the following hold true.

(a) $x_* \in \mathcal{X}_*$ *if and only if* $0 \in \partial f(x_*)$.

(b) $\mathcal{X}_* = \partial f^\star(0)$.

(c) $\mathcal{X}_*$ *is nonempty if* $0 \in \mathrm{ri}(\mathrm{dom}(f^\star))$.

(d) $\mathcal{X}_*$ *is nonempty and compact if and only if* $0 \in \mathrm{int}(\mathrm{dom}(f^\star))$.

Outline

Problem and Definitions

First-Order Conditions

Second-Order Conditions

Second-order necessary condition

Theorem 3.3.1 (Second-order necessary condition)

If $f \in \mathcal{C}^2$ and x_* is a local minimizer of f, then $\nabla^2 f(x_*) \succeq 0$.

Proof.

For $x \in \mathbb{R}^n$ with $\nabla f(x) = 0$ but $\nabla f^2(x) \not\succeq 0$, let $d \in \mathbb{R}^n$ satisfy $d^T \nabla f(x) d < 0$. (We call such a d a direction of negative curvature.) Since $\nabla^2 f$ is continuous, there exists $\alpha' > 0$ such that

$$d^T \nabla^2 f(x + \alpha d) d < 0 \quad \text{for all} \quad \alpha \in [0, \alpha'],$$

i.e., the curvature remains negative some way along d. By Taylor's Theorem (3.1.5), for all $\alpha'' \in (0, \alpha']$ and some $\alpha \in (0, \alpha'')$ we have

$$\begin{aligned} f(x + \alpha'' d) &= f(x) + \alpha'' \nabla f(x)^T d + \tfrac{1}{2} \alpha''^2 d^T \nabla^2 f(x + \alpha d) d \\ &= f(x) + \tfrac{1}{2} \alpha''^2 d^T \nabla^2 f(x + \alpha d) d \\ &< f(x). \end{aligned}$$

Thus, x cannot be a minimizer.

Discussion

- Thus, at a local minimizer x_*, the Hessian of f is positive semidefinite.
- We already know that at a minimizer x_*, we have $\nabla f(x_*) = 0$, so together

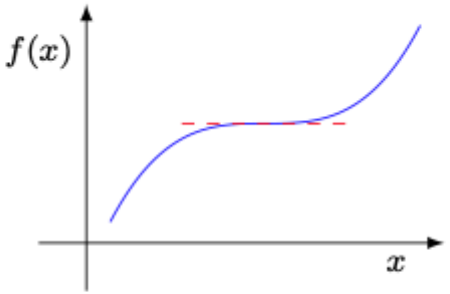
$$\nabla f(x_*) = 0 \quad \text{and} \quad \nabla^2 f(x_*) \succeq 0$$

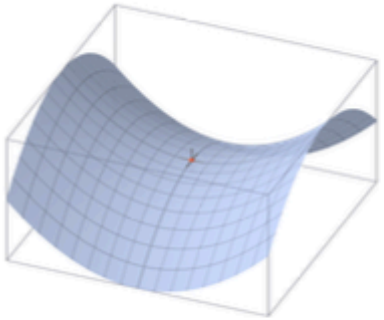
must be true at any local minimizer x_* of f.

- We can limit our search to points with zero gradient, then throw out any points where the Hessian is not positive semidefinite.

Necessary, but not sufficient

The fact that we may have $\nabla^2 f(x_*)d = 0$ for some d makes these insufficient.





- $f(x) = 1 + (x-4)^3$ has

$$\nabla f(x)|_{x=4} = 3(x-4)^2\big|_{x=4} = 0 \quad \text{and} \quad \nabla^2 f(x)\big|_{x=4} = 6(x-4)|_{x=4} = 0,$$

so the second order necessary conditions are satisfied at $x = 4$!

- $f(x) = x_1^4 - x_2^4$ has

$$\nabla f(x)|_{x=0} = \begin{bmatrix} 4x_1^3 \\ -4x_2^3 \end{bmatrix}\Bigg|_{x=0} = 0 \quad \text{and} \quad \nabla^2 f(x)\big|_{x=0} = \begin{bmatrix} 12x_1^2 & 0 \\ 0 & -12x_2^2 \end{bmatrix}\Bigg|_{x=0} = 0$$

so the second order necessary conditions are satisfied at $x = 0$.

- Note: the second order necessary conditions can be satisfied at a maximizer!

Second-order sufficient conditions

Theorem 3.3.2 (Second-order sufficient conditions)

If $f \in \mathcal{C}^2$, $\nabla f(x_*) = 0$, and $\nabla^2 f(x_*) \succ 0$, then x_* is a strict local minimizer.

Proof sketch.

Since $\nabla^2 f$ is continuous, it remains positive definite near x_*. Taylor's Theorem (3.1.5) and $\nabla f(x_*) = 0$ then imply that, for some $\alpha \in (0, 1)$,

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

Hence, f must take larger values at other points near x_*. (See textbook.)

- A nice fact, when we can actually use it!
- By designing algorithms that find a sequence of points with decreasing function values, one hopes that maximizers and saddle points are avoided, i.e., one often focuses on finding a point with zero gradient. That being said, one can search over negative curvature directions to find a point satisfying the second-order necessary conditions, but, in general, a point satisfying the second-order sufficient conditions may not exist.