Unconstrained Optimization

Yuqian Zhang

Rutgers University

yqz.zhang@rutgers.edu

February 18, 2021

Overview

1. Unconstrained Optimization
2. Existence of minimizers
3. Local is global
4. Optimality conditions

Unconstrained Optimization

$$\text{minimize} \quad f(x), \quad f \text{ is convex}$$

- conditions under which a minimizer exists
- if x^* is a local minimizer, then it's a global
- when f differentiable, then x^* is a minimizer if and only if the derivative is equal to zero

$$x^* \text{ is a global minimizer} \iff \nabla f(x^\star) = 0$$

Unconstrained Optimization

Minimum does not necessarily have to be achieved for any x^*

$$f(x) = e^x$$

- optimal value $p^* = 0$
- no optimal solution

$$\lim_{x \to -\infty} f(x) = 0$$

Compact sublevel set

Existence of minimizer

there exist at least one global minimizer if the sublevel sets are compact (closed and bounded)

$$s(f, a) = \{x \mid f(x) \leq a\}$$

Proof: choose a such that $s(f, a)$ is non-empty, then

$$\underset{x \in s(f,a)}{\text{minimize}} \quad f(x)$$

has a minimizer, which corresponds to a minimizer of f

Local is global

Theorem

Let $f(x)$ be convex function on $\mathbb{R}^n$, and suppose x^* is a local minimizer of f in that there exists an $\epsilon > 0$ such that

$$f(x^*) \le f(x) \quad \forall \|x - x^*\|_2 \le \epsilon$$

Then x^* is also a global minimizer: $f(x^*) \le f(x)$ for all $x \in \mathbb{R}^N$.

Unique minimizer

Theorem

Let f be strictly convex on $\mathbb{R}^n$. If f has a global minimizer, then it is unique.

- Let x^* be a global minimizer, and suppose that $x \neq x^*$ with $f(x) = f(x^*)$
- choose $0 < \alpha < 1$, then

$$\begin{aligned} f(\alpha x + (1-\alpha)x^*) &< \alpha f(x) + (1-\alpha) f(x^*) \\ &= f(x^*) \end{aligned}$$

- contradicts the assumption that x^* is the global minimizer

Continuous, differentiable and smooth function

- Continuous function

$$\lim_{x \to c} f(x) = f(c)$$

- Differentiable function: derivative exists

$$f'(a) = \lim_{\delta \to 0} \frac{f(a+\delta) - f(a)}{\delta}$$

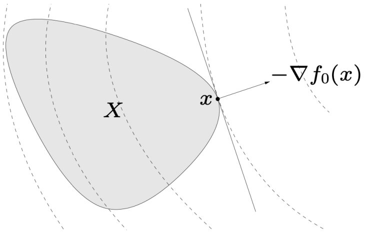
- Higher-order differentiable function: higher-order derivative exists
- Smooth function: infinitely differentiable function

Optimality conditions

Let f be a convex and differentiable function on $\mathbb{R}^n$. Then x^* solves

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)$$

if and only if $\nabla f(x^*) = 0$.



$$\nabla f_0(x)^T(y - x) \geq 0, \quad \forall \text{ feasible } y$$

Optimality conditions

- Let f be a function on $\mathbb{R}^n$ that is differentiable at x, and let $d \in \mathbb{R}^n$ be a vector obeying $\langle d, \nabla f(x) \rangle < 0$. Then for small enough $t > 0$

$$f(x + td) < f(x)$$

  We call such a d a descent direction from x.
- Similarly, if $\langle d, \nabla f(x) \rangle > 0$, then for small enough $t > 0$, $f(x + td) > f(x)$. We call such a d an ascent direction from x.

Optimality conditions – Proof

- For any $\boldsymbol{u} \in \mathbb{R}^n$

$$f(\boldsymbol{x} + \boldsymbol{u}) = f(\boldsymbol{x}) + \langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle + h(\boldsymbol{u}) \|\boldsymbol{u}\|_2$$

  where $h(\boldsymbol{u}) : \mathbb{R}^n \to R$ is some function satisfying $h(\boldsymbol{u}) \to 0$ as $\boldsymbol{u} \to \mathbf{0}$.

- take $\boldsymbol{u} = t\boldsymbol{d}$, we have

$$f(\boldsymbol{x} + \boldsymbol{u}) = f(\boldsymbol{x}) + t\left(\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle + h(t\boldsymbol{d}) \|\boldsymbol{d}\|_2\right)$$

- For $t > 0$ small, we can make $|\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle| > |h(t\boldsymbol{d})| \|\boldsymbol{d}\|_2$

Optimality conditions – Proof

- At a particular point x^*, the only way to make $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle \geq 0$ for all choice of $\boldsymbol{d}$ is $\nabla f(\boldsymbol{x}^*) = 0$

$$x^* \text{ is a minimizer } \implies \nabla f(\boldsymbol{x}^*) = 0$$

- On the other hand, if f is convex, then

$$f(\boldsymbol{x}^* + t\boldsymbol{d}) \geq f(\boldsymbol{x}) + t \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}^*) \rangle$$

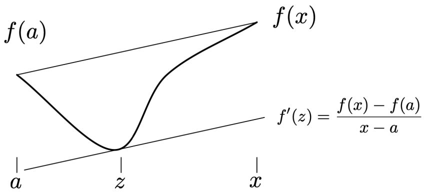
for any t and $\boldsymbol{d} \in \mathbb{R}^n$, hence

$$\nabla f(\boldsymbol{x}^*) = 0 \implies x^* \text{ is a minimizer}$$

Taylor's Theorem

If $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function on the interval $[a, x]$, then there is a point inside this interval where the derivative of f matches the line drawn between $f(a)$ and $f(x)$, there exists $z \in [a, x]$ such that

$$\begin{aligned} f'(z) &= \frac{f(x) - f(a)}{x - a} \\ \implies f(x) &= f(a) + f'(z)(x - a) \end{aligned}$$



Taylor's Theorem

If f is twice differentiable on $[a, x]$, and that the first derivative f' is continuous.

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(z)(x-a)^2$$

In general, if f is $k+1$ times differentiable, and the first k derivatives are continuous, then there is a point z between a and x such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(k+1)}(z)}{(k+1)!}(x-a)^{k+1}$$

Taylor's Theorem

To quantify accuracy of the Taylor approximation around a point

- If f is differentiable

$$f(x) = f(a) + f'(a)(x-a) + h_1(x)(x-a)$$

  where $h_1(x) \to 0$ as x approaches a

- If f is twice differentiable

$$h_1(x) = \frac{f''(z)}{2}(x-a)$$

Taylor's Theorem

In multidimensional case $f : \mathbb{R}^n \to \mathbb{R}$

- If f is differentiable, then

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \langle \nabla f(\boldsymbol{a}), \boldsymbol{x} - \boldsymbol{a} \rangle + h_1(\boldsymbol{x}) \left\| \boldsymbol{x} - \boldsymbol{a} \right\|_2$$

  where $h_1(\boldsymbol{x}) \to 0$ as $\boldsymbol{x}$ approaches $\boldsymbol{a}$ from any direction

- If f is twice differentiable on $[\boldsymbol{a}, \boldsymbol{x}]$, and that the first derivative is continuous

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \langle \nabla f(\boldsymbol{a}), \boldsymbol{x} - \boldsymbol{a} \rangle + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a})^T \nabla^2 f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{a})$$

General descent algorithm

- choose a starting point $x^{(0)}$
- determine a descent direction $d^{(k)}$
- choose a step size $t \geq 0$
- update $x^{(k)}$ as $x^{(k-1)} + td^{(k)}$
- jump to step 2 until $\|\nabla f(x)\|_2 \leq \epsilon$

Gradient descent algorithm

$$d^{(k)} = -\nabla f(x^{(k-1)})$$

- This is the direction of steepest descent

$$\left\langle d^{(k)}, \nabla f(x^{(k-1)}) \right\rangle = -\left\| \nabla f(x^{(k-1)}) \right\|_2^2$$

- Gradient descent iterations are cheap, but typically many iterations are required for convergence.

Newton's method

$$d^{(k)} = -(\nabla^2 f(x^{(k-1)}))^{-1} \nabla f(x^{(k-1)})$$

and

$$\left\langle d^{(k)}, \nabla f(x^{(k-1)}) \right\rangle = -\nabla f(x^{(k-1)})^T (\nabla^2 f(x^{(k-1)}))^{-1} \nabla f(x^{(k-1)})$$

- Idea: use a second-order approximation to function.

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

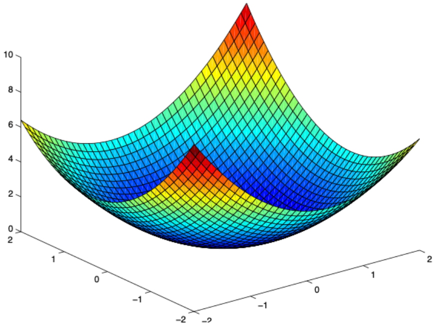
- tend to be expensive (as they require a system solve), but they typically converge in far fewer iterations than gradient descent

Second order convexity conditions

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. Then f is convex if and only if

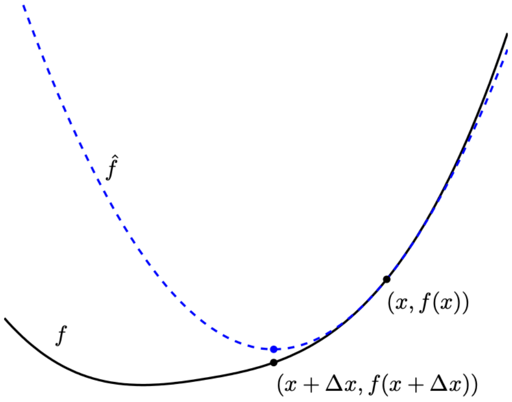
$$\nabla^2 f(x) \succeq 0$$

for all $x \in \mathrm{dom}(f)$.



Newton's method

$\hat{f}$ is 2nd order approximation of f



More on algorithms

- What's the convergence rate?
- How to choose the step size?
- ...