Lecture 3: Solution Concepts; Taylor's Theorems

Yudong Chen

Consider the problem

$$\min_{x \in \mathcal{X}} f(x), \tag{P}$$

where $\mathcal{X} \subseteq \mathrm{dom}(f) \subseteq \mathbb{R}^n$ is a closed set.

1 A Taxonomy of Solutions to (P)

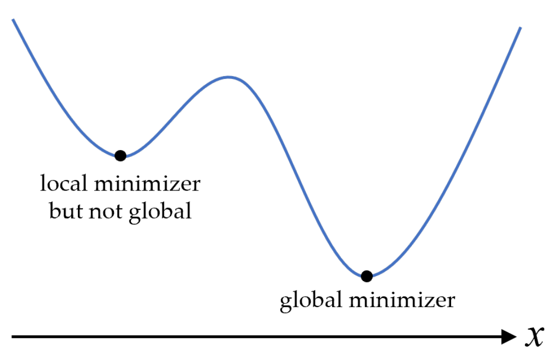
Will use "solution" and "minimizer" interchangeably.

Definition 1. We say that $x^* \in \mathcal{X} \subseteq \mathrm{dom}(f)$ is

1. a *local minimizer/solution* of (P) if there exists a neighborhood $\mathcal{N}_{x^*}$ of x^* such that for all $x \in \mathcal{N}_{x^*} \cap \mathcal{X}$ we have $f(x) \geq f(x^*)$;
2. a *global minimizer* of (P) if $\forall x \in \mathcal{X}$: $f(x) \geq f(x^*)$
3. a *strict local minimizer* of (P) if there exists a neighborhood $\mathcal{N}_{x^*}$ of x^* such that for all $x \in \mathcal{N}_{x^*} \cap \mathcal{X}$ and $x \neq x^*$ we have $f(x) > f(x^*)$; (i.e., satisfies part 1 with a strict inequality)
4. an *isolated local minimizer* of (P) if there exists a neighborhood $\mathcal{N}_{x^*}$ such that $\forall x \in \mathcal{N}_{x^*} \cap \mathcal{X}$: $f(x) \geq f(x^*)$ and $\mathcal{N}_{x^*}$ does not contain any other local minimizer.
5. a *unique minimizer* if it is the only global minimizer.

Example 1. A local minimizer that is not strict: consider a constant function

Example 2. A local minimizer that is not global: (picture)



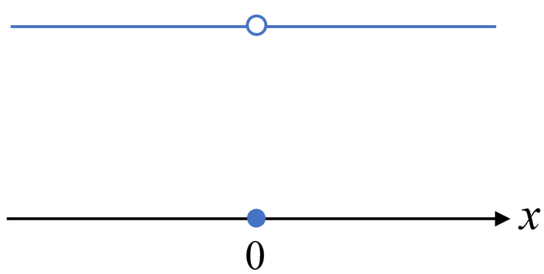
Exercise 1. Prove that every isolated local minimizer is strict.

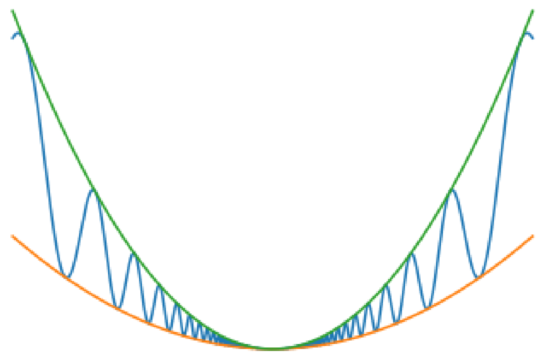
The converse of the above statement does *not* hold in general, as demonstrated by the example below.

Example 3. A strict minimizer that is not isolated:

- (not continuous) $f_1(x) = \begin{cases} 1 & x \neq 0 \\ 0 & x = 0 \end{cases}$ and $x^* = 0$.
- (continuous) $f_2(x) = \begin{cases} x^2\left(1+\sin^2(\frac{1}{x})\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$ and $x^* = 0$.

Illustration: Left f_1. Right: f_2.





We want to determine whether a particular point is a local or global minimizer. A powerful tool is Taylor's theorem.

2 Taylor's Theorem

For this part and until explicitly stated otherwise, we will be assuming that f is at least once continuously differentiable (i.e., gradient exists everywhere and is continuous).

Recall: Taylor's Theorem for 1D functions from calculus: Let $f : \mathbb{R} \to \mathbb{R}$ be a k-times continuously differentiable function. Then

$$\forall x, y \in \mathbb{R} : f(y) = f(x) + \frac{1}{1!}f'(x)(y-x) + \frac{1}{2!}f''(x)(y-x)^2 + \cdots + \frac{1}{k!}f^{(k)}(x)(y-x)^k + \underbrace{R_k(y)}_{\text{remainder}}.$$

Typical forms of $R_k(y)$ (assume that f is $k+1$ times continuously differentiable):

- Lagrange (mean-value) remainder:
$$R_k(y) = \frac{1}{(k+1)!}f^{(k+1)}\left(x + \gamma(y-x)\right) \cdot (y-x)^{k+1}$$
for some $\gamma \in (0,1)$;
- Integral remainder:
$$R_k(y) = \frac{1}{k!}\int_0^1 (1-t)^k f^{(k+1)}\left(x + t(y-x)\right)(y-x)^{k+1} \mathrm{d}t.$$

Below is the multivariate version.

Theorem 1 (Taylor's Theorem; Thm 2.1 in Wright-Recht)**.** *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a continuously differentiable function. Then, for all $x, y \in \operatorname{dom}(f)$ such that $\{(1-\alpha)x + \alpha y : \alpha \in (0,1)\} \subseteq \operatorname{dom}(f)$, we have*

1. $f(y) = f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y - x\rangle \,\mathrm{d}t$
2. $f(y) = f(x) + \langle \nabla f(x + \gamma(y-x)), y - x\rangle$ *for some* $\gamma \in (0,1)$ *(a.k.a. Mean Value Thm).*

If f is twice continuously differentiable:

3. $\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + t(y-x))(y-x)\mathrm{d}t$. *Here*

$$\nabla^2 f(x) = \begin{bmatrix} & \cdots & \\ \vdots & \frac{\partial^2 f}{\partial x_i \partial x_j}(x) & \vdots \\ & \cdots & \end{bmatrix} \in \mathbb{R}^{d\times d}$$

denotes the Hessian matrix ("second-order derivative") of f at x.

4. $\exists \gamma \in (0,1)$:

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}\left\langle \nabla^2 f(x + \gamma(y-x))(y-x), y-x\right\rangle \\ &= f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}(y-x)^\top \nabla^2 f(x+\gamma(y-x))(y-x). \end{aligned}$$

Remark 1. A common mistake is to write down the following "Mean-Value Thm" for the gradient:

$$\exists \gamma \in (0,1) : \nabla f(y) = \nabla f(x) + \nabla^2 f(x + \gamma(y-x))(y-x)? \longleftarrow \textbf{This is wrong!}$$

2.1 Digression: order notation

Two sequences: $\{a_k\}_{k\geq 1}, \{b_k\}_{k\geq 1}$, for all k: $a_k, b_k \geq 0$.

Big-Oh notation: $a_k = O(b_k) \iff$

$$(\exists M > 0)(\exists K < \infty)(\forall k \geq K) : a_k \leq M b_k.$$

e.g. $k = O(\frac{1}{10}k^2)$, $k = O(\frac{1}{10!}k)$

If $a_k = O(b_k)$ and $b_k = O(a_k)$, we write $a_k = \Theta(b_k)$.

Little-oh notation:

$$a_k = o(b_k) \iff \lim_{k\to\infty} \frac{a_k}{b_k} = 0.$$

So $a_k = o(1)$ means $a_k \to 0$.

Using the notations above, we can show that for f continuously differentiable at x, we have

$$f(x+p) = f(x) + \nabla f(x)^\top p + o(\|p\|).$$

Explicitly, this means

$$\lim_{\|p\|\to 0} \frac{\left|f(x+p) - f(x) + \nabla f(x)^\top p\right|}{\|p\|} = 0.$$

Proof. By part 2 of Theorem 1 (Taylor's), we have

$$\begin{aligned} f(x+p) &= f(x) + \nabla f(x+\gamma p)^\top p \\ &= f(x) + \nabla f(x)^\top p + \left(\nabla f(x+\gamma p) - \nabla f(x)\right)^\top p \\ &= f(x) + \nabla f(x)^\top p + O\left(\|\nabla f(x+\gamma p) - \nabla f(x)\|_2 \cdot \|p\|_2\right) \qquad \text{Cauchy-Schwarz} \\ &= f(x) + \nabla f(x)^\top p + o\left(\|p\|_2\right), \end{aligned}$$

where the step follows from continuity of ∇f: $\|\nabla f(x+\gamma p) - \nabla f(x)\|_2 \to 0$ as $p \to 0$. $\square$