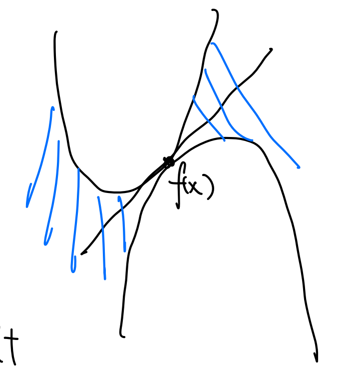
1 Properties of smooth functions

上下



Recall: f is called L-smooth w.r.t. $\|\cdot\|$ if

$$\forall x, y \in \mathrm{dom}(f) : \|\nabla f(x) - \nabla f(y)\|_* \le L \|x - y\| .$$

Lemma 1. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be an L-smooth function w.r.t. $\|\cdot\|$ .Then, $\forall x, y \in \mathrm{dom}(f)$:*

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad ①$$
$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|y - x\|^2. \quad ②$$

Pf: By Taylor's thm. $f(y) = f(x) + \int_0^1 \langle \nabla f(x+t(y-x)), y-x\rangle dt$

① $\int_0^1 \langle \nabla f(x+t(y-x)), y-x\rangle dt - \int_0^1 \langle \nabla f(x), y-x\rangle dt$

$= \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle dt \le \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\| \, \|y-x\| \, dt$

$\le \int_0^1 L\|y-x\|^2 t \, dt = \frac{L}{2}\|y-x\|^2.$

Ex. ② $-\int_0^1 \langle \nabla f(x) - \nabla f(x+t(y-x)), y-x\rangle dt$

$\ge -\int_0^1 L\|-t(y-x)\| \, \|y-x\| \, dt = -L\|y-x\|^2 \int_0^1 t \, dt = -\frac{L}{2}\|y-x\|^2.$ Plug back.

Remark 1. In fact, the condition in Lemma 1 is *equivalent* to L-smoothness; see Lemma 3.

Recall the Lowner order: For *symmetric* matrices A and B,

$$A \succcurlyeq B \iff A - B \succcurlyeq 0 \iff A - B \text{ is p.s.d.}$$

In particular,

$$aI \preccurlyeq A \preccurlyeq bI \iff a \le \lambda_i(A) \le b, \forall i$$

where $\lambda_1(A) \le \cdots \le \lambda_d(A)$ are the eigenvalues of A.

Lemma 2. *Suppose that $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is twice continuously differentiable on* $\mathrm{dom}(f)$*. Then f is L-smooth w.r.t. $\|\cdot\|_2$ if and only if*

$$-LI \preccurlyeq \nabla^2 f(x) \preccurlyeq LI, \qquad \forall x \in \mathrm{dom}(f).$$

To give the proof, we use the matrix operator norm:

$$\|A\|_2 := \sup_{x:\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \overset{\text{for symmetric } A}{=} \max_i |\lambda_i(A)| .$$

Then by definition:

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2 . \tag{1}$$

Pf: ($\Rightarrow$) f is smooth. Show that $\nabla^2 f(x) \preceq LI$.

let $x, x+\alpha p \in \mathrm{dom}(f)$. $\alpha > 0$. By Taylor's thm, (part 4)

$\exists \gamma \in (0,1)$

$$f(x+\alpha p) = f(x) + \langle \nabla f(x), \alpha p \rangle + \frac{1}{2} (\alpha p)^T \nabla^2 f(x+\gamma \alpha p)\, \alpha p$$

$$= f(x) + \langle \nabla f(x), \alpha p \rangle + \frac{\alpha^2}{2} p^T \nabla^2 f(x+\gamma\alpha p)\, p$$

Lemma I (part I) $\le f(x) + \langle \nabla f(x), \alpha p \rangle + \frac{L}{2} \|\alpha p\|^2$

$\Rightarrow \frac{\alpha^2}{2} p^T \nabla^2 f(x+\gamma\alpha p)\, p \le \frac{\alpha^2}{2} L \|p\|_2^2$ let $\alpha \to 0$ $\Rightarrow \forall p,\ p^T \nabla^2 f(x) p - L\|p\|_2^2 \le 0$

$\Leftrightarrow \forall p,\ p^T(\nabla^2 f(x) - LI)p \le 0 \Rightarrow \nabla^2 f(x) \preceq LI$.

To prove $\nabla^2 f(x) \succeq -LI$. plug in Lemma I, part 2.

($\Leftarrow$) $\forall x,\ -LI \preceq \nabla^2 f(x) \preceq LI$. By Taylor's Thm,

$$\|\nabla f(y) - \nabla f(x)\|_2 = \left\| \int_0^1 \nabla^2 f(x+t(y-x))(y-x)\, dt \right\|$$

$$\le \int_0^1 \|\nabla^2 f(x+t(y-x))(y-x)\|\, dt$$

$$\le \int_0^1 L \cdot \|y-x\|\, dt = L\|y-x\|_2$$

2 Characterizing minima of smooth functions

Where the fxxk the smoothness in this section?

In this part, we consider *unconstrained* optimization, that is, $\mathcal{X} = \mathbb{R}^d$ in the problem

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

2.1 Necessary conditions for optimality

Theorem 1.

1. *(First-order necessary condition) Suppose that $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is continuously differentiable. If x^* is a local minimizer of f, then $\nabla f(x^*) = 0$.*
2. *(Second-order necessary condition) Suppose that $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is twice continuously differentiable. Then in additional to 1), $\nabla^2 f(x^*) \succcurlyeq 0$.*

Remark 2. A point x satisfying $\nabla f(x) = 0$ is called a (first-order) stationary point of f. A point x satisfying $\nabla f(x) = 0$ and $\nabla^2 f(x) \succcurlyeq 0$ is called a *second-order stationary point* (SOSP). Theorem 1 says a local minimizer must be a stationary point if f is continuously differentiable, and it must be a SOSP if f is twice continuously differentiable.

pf: Part I: Suppose for contra $\nabla f(x^*) \neq 0$.

Let $y = x^* - \alpha \nabla f(x^*)$ and $x = x^*$, apply Taylor's thm. $(\alpha > 0)$

$$\exists \nu \in (0,1), \quad f(x^* - \alpha \nabla f(x^*)) - f(x^*) = \langle \nabla f(x^* + \nu \cdot (-\alpha \nabla f(x^*))), -\alpha \nabla f(x^*) \rangle$$
$$= -\alpha \langle \nabla f(x^* - \nu\alpha \nabla f(x^*)), \nabla f(x^*) \rangle$$

For all sufficiently small $\alpha \geq 0$,

$$-\langle \nabla f(x^* - \nu\alpha \nabla f(x^*)), \nabla f(x^*) \rangle \leq -\frac{1}{2} \|\nabla f(x^*)\|_2^2$$

As $\alpha \to 0$, $\langle \nabla f(x^* - \nu\alpha \nabla f(x^*)), \nabla f(x^*) \rangle \to \|\nabla f(x^*)\|_2^2$
$\geq \frac{1}{2} \|\nabla f(x^*)\|_2^2$

Hence $f(x^* - \alpha \nabla f(x^*)) \leq f(x^*) - \frac{\alpha}{2} \|\nabla f(x^*)\|_2^2 < f(x^*)$. Contradicts with x^* is local min.

$\Rightarrow \nabla f(x^*) = 0$.

Part 2: Suppose for contra, $\nabla^2 f(x)$ has neg eivenvalue $-\lambda$ $(\lambda > 0)$.

Fix $\theta \in \mathbb{R}^d$, $\|\theta\|_2 = 1$. $\theta^T \nabla^2 f(x^*) \theta = -\lambda$. $x = x^*$, $y = x + \alpha\theta$. $\alpha > 0$.

Apply Taylor, $\exists \nu \in (0,1)$

$$f(x^* + \alpha\theta) = f(x^*) + \langle \underbrace{\nabla f(x^*)}_{=0}, \alpha\theta \rangle + \frac{\alpha^2}{2} \theta^T \nabla^2 f(x + \nu\alpha\theta) \theta$$

As $\alpha \to 0$, $\theta^T \nabla^2 f(x + \nu\alpha\theta) \theta \to \lambda \|\theta\|_2^2 = \lambda$
$> \frac{\lambda}{2}$

For sufficiently small α, $\theta^T \nabla^2 f(x + \nu\alpha\theta) \theta \leq -\frac{\lambda}{2}$.

$\Rightarrow f(x^* + \alpha\theta) \leq f(x^*) - \frac{\lambda\alpha^2}{4} < f(x^*)$. Contradicts with x^* is a minimizer

2.1.1 An alternative proof

From calculus, we have the derivative tests for characterizing critical points of **1D** functions. Taking these 1D results as given, we can use them to prove the multivariate results in Theorem 1.

Part 1: Define the 1-D function $\phi(\alpha) = f(x^* - \alpha \nabla f(x^*))$. If x^* is a local minimizer of f, then 0 is a local minimizer of ϕ, then $\phi'(0) = 0$ by Fermat's Theorem. But

$$\phi'(\alpha) = \langle \nabla f(x^* - \alpha \nabla f(x^*)), -\nabla f(x^*) \rangle,$$
$$\phi'(0) = -\|\nabla f(x^*)\|_2^2,$$

so we must have $\nabla f(x^*) = 0$.

Part 2: Fix an arbitrary $\theta \in \mathbb{R}^d$, define $\phi_\theta(\alpha) = f(x^* + \alpha\theta)$. Use 2nd derivative test on ϕ_θ and $\phi_\theta'(0) = 0$.

2.2 Sufficient condition for optimality

Theorem 2 (Second-order sufficient condition). *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be twice continuously differentiable and assume that for some $x^* \in \mathrm{dom}(f)$,*

$$\begin{cases} \nabla f(x^*) = 0 \quad \text{and} \\ \nabla^2 f(x^*) \succ 0. \end{cases}$$

Then x^* is a $\boxed{\text{strict local minimizer}}$ of f.

Pf: Construct a nbhd around x^*, $B(x^*, \rho)$. Let $\rho > 0$, we have $\nabla^2 f(x+p) \succeq \varepsilon I$.
for some $\varepsilon > 0$. $\forall p$, $\|p\|_2 < 1$. $x^* + p$ can represents all pts in nbhd.

$\exists \gamma \in (0,1)$. $f(x^*+p) = f(x^*) + \langle \underbrace{\nabla f(x^*)}_{=0}, p \rangle + \frac{1}{2} p^T \nabla^2 f(x^* + \gamma p) p$

$= f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + \gamma p) p$

$\geq f(x^*) + \frac{1}{2} \varepsilon \|p\|_2^2 > f(x^*)$ if $\|p\|_2 \neq 0$. $\Rightarrow x^*$ is strict local min.

Remark 3. We notice that there is a gap between the conditions in last two theorems. The condition $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succeq 0$ in Theorem 1 is necessary but not sufficient: it is possible that a point x satisfies this condition but is not a local min (e.g., $f(x) = x^3$ and $x = 0$). The condition $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$ in Theorem 2 is sufficient but not necessary: it is possible that a local minimizer x^* has $\nabla^2 f(x^*) = 0$ (e.g., $f(x) = x^4$ and $x^* = 0$). In general, it is hard to check whether a point x is a local min, even for smooth unconstrained problems. For example, consider the function

$$f(x) = (x_1^2, x_2^2, \ldots, x_d^2) D (x_1^2, x_2^2, \ldots, x_d^2)^\top,$$

which is a degree-4 polynomial in x. It is NP hard to decide whether $x = 0$ is a local min (by reduction from Subset Sum; Murty-Kabadi 1987),