Special classes of function in optimization in machine learning

Andersen Ang

ECS, Uni. Southampton, UK
andersen.ang@soton.ac.uk

Homepage angms.science

Version: April 28, 2023
First draft: June 6, 2017

Content

Some old-school terminology

Notation used by Nesterov, Mordukhovich, or any classical real analysis textbooks:

- $f \in C^0$: $f(\boldsymbol{x})$ is continuous
- $f \in C^1$: $f(\boldsymbol{x})$ and $\nabla f(\boldsymbol{x})$ are continuous
- $f \in C^2$: $f(\boldsymbol{x})$, $\nabla f(\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x})$ are continuous
- $f \in C^{1,1}$: $f(\boldsymbol{x})$ and $\nabla f(\boldsymbol{x})$ are continuous, $\nabla f(\boldsymbol{x})$ is L-Lipschitz with $L < +\infty$
- $f \in C_L^{k,p}$: f is k times continuously differentiable and pth derivative is L-Lipschitz
- $f \in \mathcal{F}_L^k$: f is $\mathcal{C}_L^k$ and convex
- $f \in \mathcal{S}_{M,L}^k$: f is $\mathcal{F}_L^k$ and M-strongly convex

Table of Contents

Convex function

A function $f(\boldsymbol{x}) : \mathrm{dom}\, f \to \mathbb{R}$ is **convex** if :

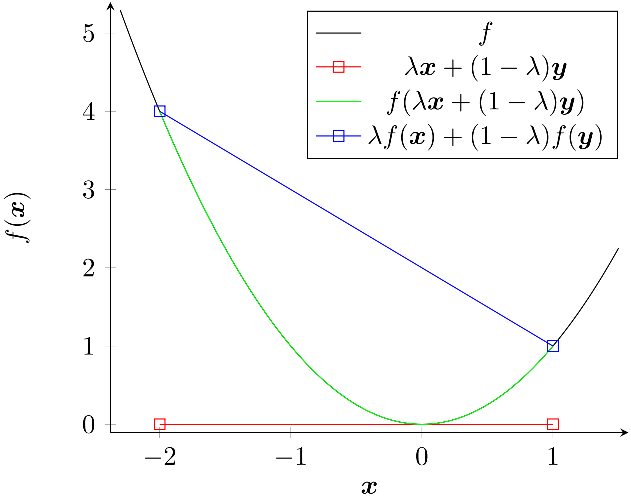
- $\mathrm{dom}\, f$ is a convex set[1]
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}\, f$, we have any one of the following
  1. Jensen's inequality: $f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$.
  2. Epigraph of f is a convex set.
  3. 1st-order Taylor series at $\boldsymbol{x}$ is a global under-estimator: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle$.
  4. Gradient is monotonic: $\big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \geq 0$.

  (For 3,4, if f is not differentiable, we replace gradient by subgradient.)
- The 4 definitions are equivalent ("if and only if"). See optimization books for proof. See here for proof of $1 \iff 3$.
- If f is twice differentiable, it is convex iff $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$.
- f is **strictly convex** if $\leq, \geq$ became $<, >$ (i.e. strict inequality).

[1] $\mathrm{dom}\, f$ can be open set. However, in optimization usually $\mathrm{dom}\, f$ is closed because optimization over an open set has no solution.

Convexity: the geometry of Jensen's inequality

$f : \operatorname{dom} f \to \mathbb{R}$ is **convex** IF (1) $\operatorname{dom} f$ is a convex set and
(2) $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$

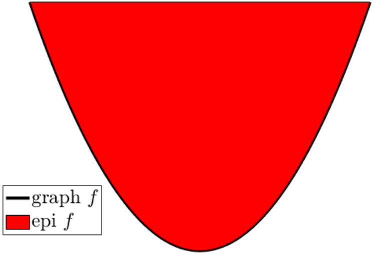


Convexity: the convex geometry of epigraph

$f : \mathrm{dom}\, f \to \mathbb{R}$ is **convex** IF epigraph of f is a convex set

Visualization of $\mathrm{graph} f$ and $\mathrm{epi}\, f$

- $\mathrm{epi}\, f$ = **all** the points of $\mathbb{R}^{n+1}$ lying on or above $\mathrm{graph} f$.
- Example: $f(x) = x^2$
  - $n = 1$ (1-dimensional)
  - $\mathrm{graph} f := \Big\{(x, y) \in \mathbb{R} \times \mathbb{R} \,:\, y = f(x)\Big\}$ is a 1d curve in a 2d space.
  - $\mathrm{epi}\, f := \Big\{(x, \alpha) \in \mathbb{R} \times \mathbb{R} \,:\, \alpha \geq f(x)\Big\}$ is a 2d set in a 2d space.



5 / 28

Details.

Convexity: the geometry of 1st-order Taylor series

- $f : \operatorname{dom} f \to \mathbb{R}$ is **convex** if :
  1. $\operatorname{dom} f$ is a convex set
  2. $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, we have

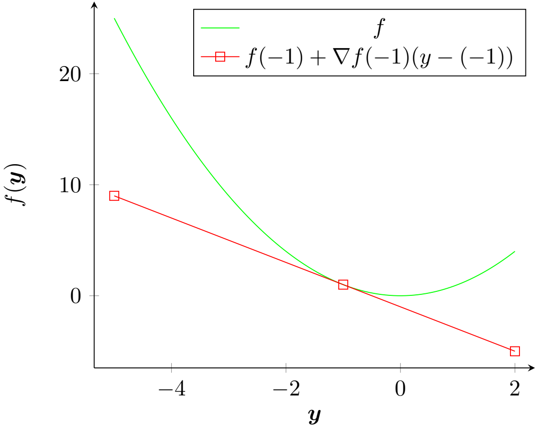
$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \qquad (*)$$

- $(*)$ assumes f is differentiable at $\boldsymbol{x}$. If f is not differentiable at $\boldsymbol{x}$, we generalize gradient with **subgradient**:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{q}, \boldsymbol{y} - \boldsymbol{x} \rangle. \qquad (\#)$$

  I.e., we replace $\nabla f(\boldsymbol{x})$ by any vector $\boldsymbol{q}$ that $(\#)$ holds.

- In fact, subgradient is defined using $(\#)$



- The gap between f and the 1st-order Taylor series is known as the Bregman Divergence.

Why convex and differentiable f is lower-bounded by their own 1st-order Taylor series?

- Consider a pedagogical case: f is (twice) differentiable of single variable, then

$$\begin{aligned} f(y) &= f(x) + f'(x)(y-x) + o(y-x) && \text{Taylor series} \\ &= f(x) + f'(x)(y-x) + \tfrac{f''(\xi)}{2}(y-x)^2 && \text{see 1} \\ &\geq f(x) + f'(x)(y-x) && \text{see 2} \end{aligned}$$

  1. Lagrange remainder theorem: using mean-value theorem, the remainder term $o(y-x) = \frac{f''(\xi)}{2}(y-x)^2$ for some ξ in the interval $[x, y]$.
  2. As f is convex, which means $f'' \geq 0$ so the last term is nonnegative.

- The arguments above generalize to multi-variable f.

- Note: this is not a prove but an illustration, because
  - apart from assuming f is differentiable, we assumed f is twice differentiable,
  - we didn't show that f is convex $\iff$ its Hessian is positive semi-definite.

α-strongly convex function

A function $f : \operatorname{dom} f \to \mathbb{R}$ is α-**strongly convex** if:

- $\operatorname{dom} f$ is a convex set.
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, we have any one of the following
  1. Jensen's inequality with an additional quadratic term with $\alpha > 0$

  $$f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\alpha}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

  2. grad$f$ is monotonic with an additional quadratic term with $\alpha > 0$

  $$\big\langle \boldsymbol{x}-\boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\big\rangle \;\geq\; \alpha\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \geq 0.$$

  3. 1st-order Taylor series at $\boldsymbol{x}$ is global under-estimator with an additional quadratic term with $\alpha > 0$

  $$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\big\rangle + \frac{\alpha}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,$$

  or we say f is lower bounded by a quadratic function.
  4. With $\alpha > 0$, the function $f(\boldsymbol{x}) - \frac{\alpha}{2}\|\boldsymbol{x}\|_2^2$ is convex.
- These definitions are equivalent.
- If f is twice differentiable, it is α-strongly convex iff $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}$.

Illustrating equivalence between definitions of strong convexity

For $\alpha > 0$ and f twice differentiable, $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I} \implies \big\langle \boldsymbol{x}-\boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \geq \alpha \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$.

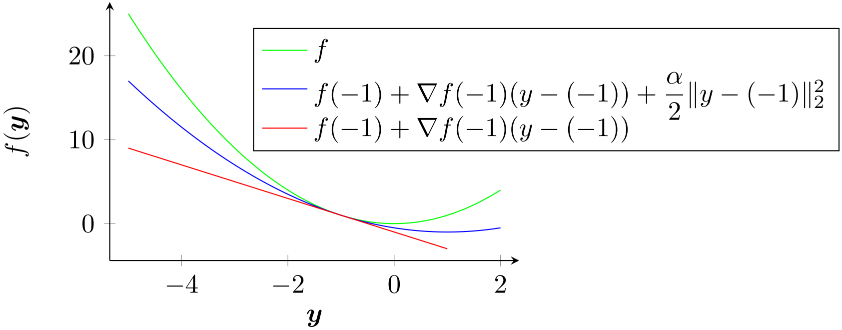
- **Proof**. Recall from calculus $G(b) - G(a) = \int_a^b g(\theta) d\theta$. Next, a smart step, let $\theta = \boldsymbol{y} + \tau(\boldsymbol{x}-\boldsymbol{y})$, then $d\theta = (\boldsymbol{x}-\boldsymbol{y})d\tau$. Consider integral range from 0 to 1 for τ we let G be ∇f and g be $\nabla^2 f$, this gives
$$\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) = \int_0^1 \nabla^2 f\big(\boldsymbol{y} + \tau(\boldsymbol{x}-\boldsymbol{y})\big)(\boldsymbol{x}-\boldsymbol{y}) d\tau.$$
(left hand side is a vector, right hand side is matrix-vector product, also a vector)

- Take dot product with $\boldsymbol{x}-\boldsymbol{y}$ on the whole equation on both sides
$$\begin{aligned} \big\langle \boldsymbol{x}-\boldsymbol{y},\ \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle &= \Big\langle \boldsymbol{x}-\boldsymbol{y},\ \int_0^1 \nabla^2 f\big(\boldsymbol{y} + \tau(\boldsymbol{x}-\boldsymbol{y})\big)(\boldsymbol{x}-\boldsymbol{y}) d\tau \Big\rangle \\ &\geq \Big\langle \boldsymbol{x}-\boldsymbol{y},\ \int_0^1 \alpha(\boldsymbol{x}-\boldsymbol{y}) d\tau \Big\rangle \\ &= \alpha \|\boldsymbol{x}-\boldsymbol{y}\|_2^2, \end{aligned}$$
where the inequality is due to $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}$ for all $\boldsymbol{x}$: we have $\nabla^2 f\big(\boldsymbol{y} + \tau(\boldsymbol{x}-\boldsymbol{y})\big) \succeq \alpha \boldsymbol{I}$. ■

α-strongly convex: the geometry of the lower bounded

$f(x) : \mathrm{dom}\, f \to \mathbb{R}$ is α-**strongly convex** if

(1) $\mathrm{dom}\, f$ is a convex and (2) $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}\, f$: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \dfrac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$



Meaning: f is lower bounded by a quadratic curve with some curvature, which is also lower bounded by the 1st order Taylor series (zero curvature)

$\Longrightarrow$ f is not "too flat" / at least "as curved as" the lower bound

In other words: f is at least α-amount of "bumpy".

Remarks on convexity

- Strongly convex $\implies$ strictly convex $\implies$ convex.

  The opposite is false.
  - e.g., x^4 is strictly convex but not strongly convex.
    Why: x^4 is not globally lower-bounded by x^2.
- Convexity function needs not to be differentiable.
  - That's why we have Jansen's definition

    $$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}),$$

    which does not involve ∇f.
- Strongly convex functions are coercive.

Table of Contents

Lipschitz continuity

A function $f(\boldsymbol{x}) : \operatorname{dom} f \to \mathbb{R}$ is *Lipschitz* if for any $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, there exists a constant $L \geq 0$ (the Lipschitz constant) such that

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

- Re-arrange gives

$$\frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|} \leq L \quad \overset{\boldsymbol{y}\to\boldsymbol{x}}{\approx} \quad \text{size of } \nabla f(\boldsymbol{x}) \leq L$$

$\implies$ f is Lipschitz means the "slope" (rate of change) of f is bounded above globally by L.

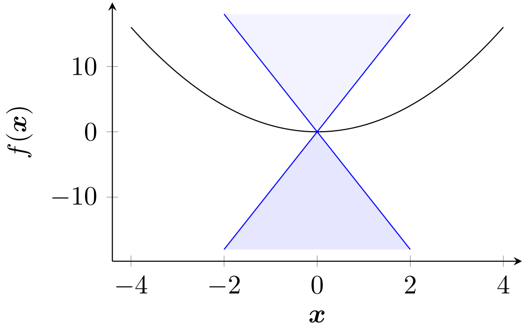
- Removing the absolute value sign:

$$\begin{cases} f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + L\|\boldsymbol{x} - \boldsymbol{y}\| \\ f(\boldsymbol{x}) \geq f(\boldsymbol{y}) - L\|\boldsymbol{x} - \boldsymbol{y}\| \end{cases}$$

means that f for all $\boldsymbol{x}$ is bounded above and below by a linear function constructed at $\boldsymbol{y}$.

The geometry of Lipschitz continuity

f is Lipschitz $\iff$ f does not have sharp change everywhere: $\forall \boldsymbol{x}$ the curve f is entirely outside a cone which is modeled by the linear functions in the last page.



Important note: such property is **global**, such cone **exists for all points on** f. i.e. the cone can "slide" along the curve and the argument still holds.

Lipschitz continuity and differentiability

- **Q**: If f is Lipschitz continuous, is f differentiable?
  A: No.

- **Rademacher's theorem**: Lipschitz function is *almost everywhere* differentiable.
  Almost everywhere $\neq$ everywhere.

- Example. $|x|$
  - $|x|$ is 1-Lipschitz but not differentiable at $x = 0$.
  - However, the single point $x = 0$ has a measure zero[2] on $\mathbb{R}$, this is what "almost everywhere" means in Rademacher's theorem.

---

[2] The probability of getting this number in a random guess on the real line is zero, because there are infinitely many real numbers.

Composition of Lipschitz functions

- Suppose f_1 is L_1-Lipschitz and f_2 is L_2-Lipschitz. Then $f_1 \circ f_2$ is $L_1 L_2$-Lipschitz.
- $f_1 \circ f_2$ means the composition of f_1 and f_2, i.e., $f_1(f_2)$
- The proof: direct proof

$$
\begin{aligned}
\|(f_1 \circ f_2)(\boldsymbol{x}) - (f_1 \circ f_2)(\boldsymbol{y})\| &\leq \|f_1\big(f_2(\boldsymbol{x})\big) - f_1\big(f_2(\boldsymbol{y})\big)\| \\
&\leq L_1 \|f_2(\boldsymbol{x}) - f_2(\boldsymbol{y})\| && f_1 \text{ is } L_1\text{-Lipschitz} \\
&\leq L_1 L_2 \|\boldsymbol{x} - \boldsymbol{y}\| && f_2 \text{ is } L_2\text{-Lipschitz}
\end{aligned}
$$

  (The proof holds for any norm, not only for ℓ_2 norm)

- This result is commutative: $f_1 \circ f_2$ and $f_2 \circ f_1$ are both $L_1 L_2$-Lipschitz
- A small detail: in Euclidean space $f_1 \circ f_2$ assumes the output dimension of f_2 match the input dimension of f_1
- Corollary: $f_1 \circ f_2 \circ \cdots \circ f_n$ is $L_1 L_2 \cdots L_n$-Lipschitz

Sum of Lipschitz functions

- Suppose f_1 is L_1-Lipschitz and f_2 is L_2-Lipschitz. Then $\alpha_1 f_1 + \alpha_2 f_2$ is $|\alpha_1|L_1 + |\alpha_2|L_2$-Lipschitz.

- **Proof** First we group the terms

$$\Big\|\alpha_1 f_1(\boldsymbol{x}) + \alpha_2 f_2(\boldsymbol{x}) - \alpha_1 f_1(\boldsymbol{y}) + \alpha_2 f_2(\boldsymbol{y})\Big\| \leq \Big\|\alpha_1\Big(f_1(\boldsymbol{x}) - f_1(\boldsymbol{y})\Big) + \alpha_2\Big(f_1(\boldsymbol{y}) - f_2(\boldsymbol{y})\Big)\Big\|$$

Use triangle inequality[3]

$$\begin{aligned}
\Big\|\alpha_1 f_1(\boldsymbol{x}) + \alpha_2 f_2(\boldsymbol{x}) - \alpha_1 f_1(\boldsymbol{y}) + \alpha_2 f_2(\boldsymbol{y})\Big\| &\leq \Big\|\alpha_1\Big(f_1(\boldsymbol{x}) - f_1(\boldsymbol{y})\Big)\Big\| + \Big\|\alpha_2\Big(f_1(\boldsymbol{y}) - f_2(\boldsymbol{y})\Big)\Big\| \\
&\leq |\alpha_1|\, \|f_1(\boldsymbol{x}) - f_1(\boldsymbol{y})\| + |\alpha_2|\, \|f_1(\boldsymbol{y}) - f_2(\boldsymbol{y})\| \\
&\leq |\alpha_1|\, L_1\|\boldsymbol{x} - \boldsymbol{y}\| + |\alpha_2|\, L_2\|\boldsymbol{x} - \boldsymbol{y}\| \\
&= \Big(|\alpha_1|L_1 + |\alpha_2|L_2\Big)\|\boldsymbol{x} - \boldsymbol{y}\|
\end{aligned}$$

[3] First for the squared term $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + 2|\langle \boldsymbol{a}, \boldsymbol{b}\rangle| \leq \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + 2\|\boldsymbol{a}\|\|\boldsymbol{b}\| = (\|\boldsymbol{a}\| + \|\boldsymbol{b}\|)^2$.
Remove the square we have $\|\boldsymbol{a}+\boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$

Max of Lipschitz functions

- Suppose f_1 is L_1-Lipschitz and f_2 is L_2-Lipschitz. Then $\max\{f_1, f_2\}$ is $\max\{L_1, L_2\}$-Lipschitz.
- Tools we need

$a \leq |a|$ $\qquad$ $a \leq \max\{a, b\}$ $\qquad$ $\begin{cases} a \leq M \\ b \leq M \end{cases} \iff \max\{a, b\} \leq M$ $\qquad$ $a \leq M$ and $-a \leq M$ imply $|a| \leq M$

- **Proof** f_1 is Lipschitz so $|f_1(\boldsymbol{x}) - f_1(\boldsymbol{y})| \leq L_1\|\boldsymbol{x} - \boldsymbol{y}\|$. By $f_1(\boldsymbol{x}) - f_1(\boldsymbol{y}) \leq L_1\|\boldsymbol{x} - \boldsymbol{y}\|$, which gives

$$f_1(\boldsymbol{x}) \leq f_1(\boldsymbol{y}) + L_1\|\boldsymbol{x} - \boldsymbol{y}\| \quad \iff \quad f_1(\boldsymbol{x}) \leq \max\{f_1(\boldsymbol{y}), f_2(\boldsymbol{y})\} + \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\| \qquad (1)$$

Similarly,

$$f_2(\boldsymbol{x}) \;\leq\; \max\{f_1(\boldsymbol{y}), f_2(\boldsymbol{y})\} + \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\| \qquad (2)$$

By , (1) and (2) gives

$$\max\{f_1(\boldsymbol{x}), \boldsymbol{f}_2(\boldsymbol{x})\} \;\leq\; \max\{f_1(\boldsymbol{y}), f_2(\boldsymbol{y})\} + \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\| \qquad (3)$$

(3) holds by swapping $(\boldsymbol{x}, \boldsymbol{y})$ as $(\boldsymbol{y}, \boldsymbol{x})$, we have

$$\max\{f_1(\boldsymbol{y}), \boldsymbol{f}_2(\boldsymbol{y})\} \;\leq\; \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\} + \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\| \qquad (4)$$

$$(3) \iff \underbrace{\max\{f_1(\boldsymbol{x}), \boldsymbol{f}_2(\boldsymbol{x})\} - \max\{f_1(\boldsymbol{y}), f_2(\boldsymbol{y})\}}_{a} \;\leq\; \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\|$$

$$(4) \iff \underbrace{\max\{f_1(\boldsymbol{y}), f_2(\boldsymbol{y})\} - \max\{f_1(\boldsymbol{x}), \boldsymbol{f}_2(\boldsymbol{x})\}}_{-a} \;\leq\; \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\|$$

By ,

$$\Big|\max\{f_1(\boldsymbol{x}), \boldsymbol{f}_2(\boldsymbol{x})\} - \max\{f_1(\boldsymbol{y}), f_2(\boldsymbol{y})\}\Big| \;\leq\; \max\{L_1, L_2\}\|\boldsymbol{x} - \boldsymbol{y}\|. \qquad \blacksquare$$

L-smooth function / Lipschitz continuous gradient

A function $f : \operatorname{dom} f \to \mathbb{R}$ is *L-smooth* if for any two points $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, there exists a constant $L < +\infty$ such that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

- This assume f is differentiable.
- "f is L-smooth" is also called L-Lipschitz gradient, or $\mathcal{C}^{1,1}$.
- "f is L-smooth" is equivalent to

$$\left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \right| \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Removing the absolute value sign gives

$$\begin{cases} f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \\ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle - \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \end{cases}$$

meaning that f is bounded above and below by a quadratic function.

Equivalent definitions of L-smooth function

A function $f(x)$ is L-smooth if

- grad$f$ is L-Lipschitz with $L \geq 0$. I.e. $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$ we have $L \geq 0$
$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$
- f is bounded by a quadratic function with $L > 0$:
$$\left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \right| \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$
- the gradient of f is monotonic with additional term with $L > 0$:
$$\left\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \right\rangle \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2.$$
- the norm of the slope of ∇f (which is $\nabla^2 f$) is bounded above.
- If f is twice differentiable, $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, or all the eigenvalue of $\nabla^2 f(\boldsymbol{x})$ is below L.

These definitions are equivalent. See here for more about the 2nd definition.

Proof of equivalence

We show for $L > 0$, $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$ implies $\Big|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \Big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \Big\rangle\Big| \leq \dfrac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$.

Recall calculus $G(b) - G(a) = \int_a^b g(\theta)d\theta$. Next, a smart step, let $g(\tau) = \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle$ be a function in τ and $d\theta = d\tau$. Consider the definite integral of $g(\tau)$ from 0 to 1, let $G(b) = f(\boldsymbol{y})$ and $G(a) = f(\boldsymbol{x})$, hence

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \Big\langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \Big\rangle d\tau \\ &= \int_0^1 \Big\langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \Big\rangle d\tau. \end{aligned}$$

As $\nabla f(\boldsymbol{x})$ is independent of τ, can take out from the integral

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \Big\langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \Big\rangle d\tau.$$

The idea is to create the term $\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$ so that we can move it to the left and get $\Big|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \Big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \Big\rangle\Big|$

Proof of equivalence - continue

$$\begin{aligned}
|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle| &= \Big| \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \, d\tau \Big| \\
&\leq \int_0^1 \Big| \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \Big| d\tau \\
&\overset{\text{Cauchy - Schwarz}}{\leq} \int_0^1 \|\nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| d\tau.
\end{aligned}$$

Look at $\|\nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\|$, this is exactly where we can apply the Lipschitz gradient inequality

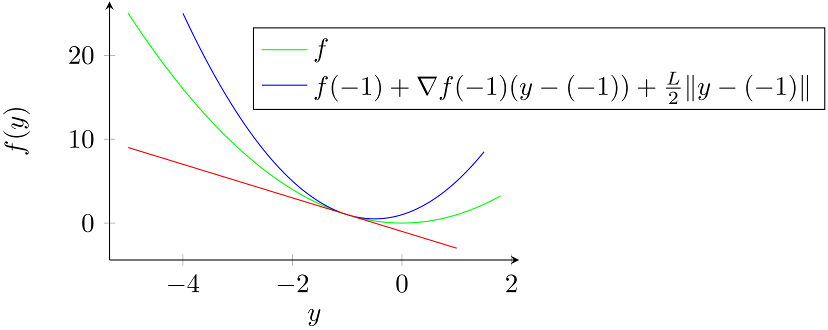
$$\|\nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \leq L\|\tau(\boldsymbol{y} - \boldsymbol{x})\| \leq L|\tau|\|\boldsymbol{y} - \boldsymbol{x}\| = L\tau\|\boldsymbol{y} - \boldsymbol{x}\|$$

where $\|\tau(\boldsymbol{y} - \boldsymbol{x})\| = |\tau|\|\boldsymbol{y} - \boldsymbol{x}\|$ as norm is non-negative. Note that the integral range is from 0 to 1 so the absolute sign in τ can be removed. Lastly

$$\Big| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \Big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \Big\rangle \Big| \leq \int_0^1 L\tau d\tau \cdot \|\boldsymbol{y} - \boldsymbol{x}\|^2 = \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \quad \square$$

L-smoothness: the geometry of the upper bound

f is L-**smooth** if $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2$



Meaning: f is globally bounded above by a quadratic function.
i.e. f cannot be "too sharp" (f is flatter than the upper bound), or f cannot grow "too fast".

Relatively-smooth function

- f is L-smooth

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle + L\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

- f is L-smooth relative to the distance kernel h

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle + L D_h(\boldsymbol{x}, \boldsymbol{y}),$$

  where D_h is the Bregman divergence on the distance kernel h.

- Why relative smoothness
  - for proving convergence of gradient descent on non-Euclidean geometry
  - for function that is not uniformly smooth,
    e.g. the slope of $x^2 - \log(x)$ approaches to ∞ as $x \to 0$, the value L change dramatically as x moves.
  - application in minimizing $\frac{1}{4}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_4^4$.
  - mirror descent

Lipschitz continuous Hessian

A function $f(\boldsymbol{x}) : \operatorname{dom} f \to \mathbb{R}$ has L-Lipschitz Hessian, if $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f, \exists L < \infty$ such that

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|.$$

- This assumes f is twice differentiable.
- This means the norm of $\nabla^3 f(\boldsymbol{x})$ is bounded above by L.
- f has L-Lipschitz Hessian is equivalent to

$$\left| f(\boldsymbol{x}) - f(\boldsymbol{y}) - \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle - \big\langle \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle \right| \leq \frac{L}{6} \|\boldsymbol{y} - \boldsymbol{x}\|_2^3$$

see here for the proof.
Removing the absolute value sign, and make $\boldsymbol{y}$ the subject:

$$\begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{x}) - \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle - \big\langle \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle - \frac{L}{6} \|\boldsymbol{y} - \boldsymbol{x}\|_2^3 \\ f(\boldsymbol{y}) \leq f(\boldsymbol{x}) - \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle - \big\langle \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle + \frac{L}{6} \|\boldsymbol{y} - \boldsymbol{x}\|_2^3 \end{cases}$$

which means $f(\boldsymbol{y})$ is bounded above and below by two cubic functions parameterized at the point $\boldsymbol{x}$ for all $\boldsymbol{y}$.

Table of Contents

Strongly convex smooth function

- A function $f : \mathrm{dom} \to \mathbb{R}$ is α-strongly convex and β-smooth if
  - f is β-smooth, which means f is differentiable and ∇f is monotone

$$\big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \;\geq\; \frac{1}{\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2.$$

  - f is α-strongly convex, which means gradient is strongly monotone

$$\big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \;\geq\; \alpha\|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

- As f satisfies both monotone inequalities, so we have

$$\big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \;\geq\; \frac{\alpha\beta}{\alpha+\beta}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2.$$

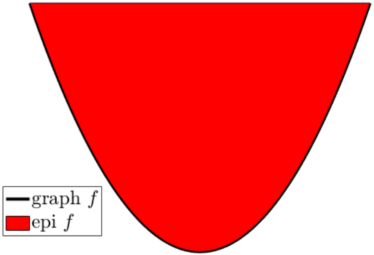
Details here.

Table of Contents

Epigraph

Visualization of $\operatorname{graph} f$ and $\operatorname{epi} f$

- $\operatorname{epi} f$ = **all** the points of $\mathbb{R}^{n+1}$ lying on or above $\operatorname{graph} f$.
- Example: $f(x) = x^2$
  - $n = 1$ (1-dimensional)
  - $\operatorname{graph} f := \left\{(x, y) \in \mathbb{R} \times \mathbb{R} \,:\, y = f(x)\right\}$ is a 1d curve in a 2d space.
  - $\operatorname{epi} f := \left\{(x, \alpha) \in \mathbb{R} \times \mathbb{R} \,:\, \alpha \geq f(x)\right\}$ is a 2d set in a 2d space.

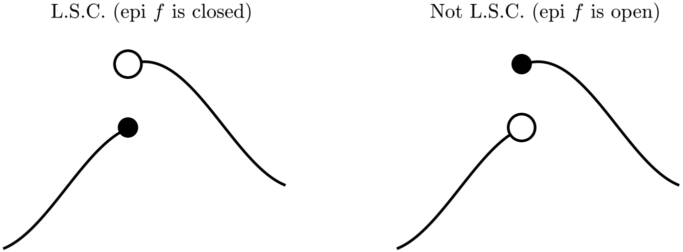


5 / 28

Details.

Lower semicontinuous function (l.s.c.)

- $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is the extended real line.
- A function is continuous means it has no "jump".
- A function is l.s.c. means it allow jump but still continuous if viewed from below.
- A function f is l.s.c. if its epigraph is closed.



- Why care about l.s.c. function: indicator function of a closed convex set are all l.s.c..

Closed and proper function

- A function f is proper if it never takes the value $-\infty$ and $\operatorname{dom} f \neq \varnothing$

  OR euivalently,

  $\operatorname{epi} f \neq \varnothing$ without svertical line[4].

- A proper function f is closed if $\operatorname{dom} f$ is closed and f is lower semicontinuous at each $x \in \operatorname{dom} f$

  OR equivalently,

  epi$f$ is closed.

[4] which can move downward and touch $-\infty$

argmin (argument of minimum = set of minimizer)

- argmin is a set defined as

$$\operatorname{argmin} f := \left\{ \boldsymbol{x} \in \operatorname{dom} f \mid f(\boldsymbol{x}) = \inf_{\boldsymbol{z} \in \operatorname{dom} f} f(\boldsymbol{z}) \right\}.$$

- If f is closed convex proper, then $\operatorname{argmin} f$ is closed convex and possibly empty[5]

[5] If $\operatorname{argmin} f$ is an empty set that means there is no minimizer for f

Polyak-Łojasiewicz and Kurdyka-Łojasiewicz

- f is Polyak-Łojasiewicz (PŁ) if $\exists \mu > 0$ such that $\|\nabla f(\boldsymbol{x})\|_2^2 \geq \mu\big(f(\boldsymbol{x}) - f^*\big)$ for all $\boldsymbol{x} \in \operatorname{dom} f$.
  - PŁ is weaker than strong convexity.
  - If f is μ-strongly convex, then f is μ-PŁ.
  - PŁ can be used as a tool to prove convergence of gradient descent, see here for more.

- Kurdyka-Łojasiewicz
  - Generalized PŁ : it can handles nonsmooth function
  - KŁ is a tool for proving convergence of gradient method on nonsmooth optimization.
  - Very technical. The original full definition is long, so we give a simplified one here. f is KŁ at a point $\bar{\boldsymbol{x}}$ if there exists $c > 0$ and $\mu \in [0, 1)$ such that $\|\partial f(\boldsymbol{x})\|_2 \geq \frac{1}{c(1-\mu)}\big(f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})\big)^{\mu}$ holds for all $\boldsymbol{x}$ within a neighbourhood of $\bar{\boldsymbol{x}}$. For $\partial f(\boldsymbol{x})$, we use the norm of the subgradient with smallest ℓ_2 norm to define $\|\partial f(\boldsymbol{x})\|_2$.
  - If f is a semi-algebraic function, the f is KŁ

- Semi-algebraic function
  - A function is semi-algebraic if epi$f$ is a semialgebraic set.
  - A set is semialgebaric if it is defined by polynomial equations and polynomial inequalities

f is proper if $\mathrm{epi}\ f$ is non-empty and has no vertical line | proper
proper f is closed if $\mathrm{epi}\ f$ is closed | closed of proper f
f is l.s.c. if epi f is closed. | Lower semicontinuous
$\mathrm{argmin}\ f$ is closed convex if f is closed convex proper | $\mathrm{argmin}\ f$ closed convex

f is convex if dom f is convex and

1. $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$ — Jansen
2. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge 0$ — Gradient is monotone
3. $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle$ — 1st-order Taylor series is global support
4. $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$, if f is twice differentiable — Hessian argument
5. epi f is convex — epigraph is convex set

f is α-strongly convex if dom f is convex and

1. $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\alpha}{2}\lambda(1-\lambda)\|x - y\|_2^2$ — Jansen
2. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge \alpha\|x - y\|_2^2$ — Strongly monotone
3. $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|x - y\|_2^2$ — Global quadratic lower bound
4. $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex — Convexity
5. $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}$, if f is twice differentiable — Hessian argument

f is L-Lipschitz gradient (L-smooth) if f is differentiable and

1. $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$ — Definition of Lipschitz
2. $\left| f(y) - f(x) - \langle \nabla f(x), y - x\rangle \right| \le \frac{L}{2}\|y - x\|_2^2$ — Quadratic inequality
3. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$ — monotone
4. $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, if f is twice differentiable — Hessian argument

f is L-Lipschitz Hessian if f is twice differentiable and

1. $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x - y\|$ — Definition of Lipschitz
2. $\left| f(x) - f(y) - \langle \nabla f(x), y - x\rangle - \langle \nabla^2 f(x)(y - x), y - x\rangle \right| \le \dfrac{L}{6}\|y - x\|_2^3$ — Cubic inequality

f is α-strongly convex and β-smooth

1. $\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\rangle \ \ge\ \dfrac{\alpha\beta}{\alpha+\beta}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$

End of document