Special classes of function in optimization in machine learning

(alternative title – some basic convex analysis for optimization)

Andersen Ang

ECS, Uni. Southampton, UK

andersen.ang@soton.ac.uk

Homepage angms.science

Version: April 11, 2024

First draft: June 6, 2017

Content

Some "old" 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

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

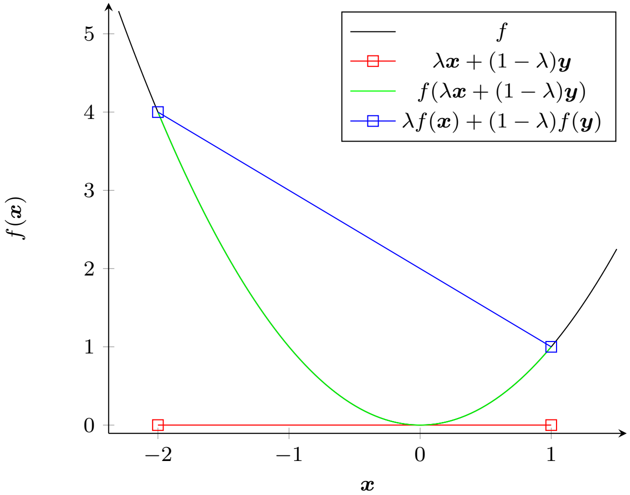
- $\mathrm{dom} f$ is a convex set[1]
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, we have any one of the following
  1. Jensen's inequality: $f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$. chord description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

  (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 the proofs. here is a proof of $1 \iff 3$.
- If f is twice differentiable, it is convex iff $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$. Hessian description
- 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 open set has no solution. For example, maximizing x over the open set $x < 3$ has no solution.

Convexity: the geometry of Jensen's inequality (chord description)

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

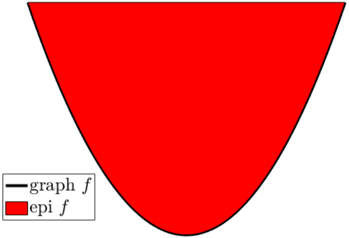


Convexity: epigraph is a convex set

$f : \mathrm{dom} f \to \mathbb{R}$ is **a convex function** $\iff$ epigraph of f is **a convex set** proof in p.10

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)
  - $\mathrm{graph} f := \left\{(x, y) \in \mathbb{R} \times \mathbb{R} \,:\, y = f(x)\right\}$ is a 1d curve in a 2d space.
  - $\mathrm{epi}\, f := \left\{(x, \alpha) \in \mathbb{R} \times \mathbb{R} \,:\, \alpha \geq f(x)\right\}$ is a 2d set in a 2d space.



5 / 28

Details.

Convexity: epi f is a convex set

- $f : \mathrm{dom} f \to \mathbb{R}$ is a convex function $\iff$ $\mathrm{epi}\, f$ is a convex set.
- What's the big deal: we connected **the function language** to **the set language**
- Suppose $\mathrm{epi}\, f$ is a closed set for a function f
- If f is a convex function, then $\mathrm{epi}\, f$ is a convex set
- **Fact**: "any closed convex sets can be written as an intersection of half space" (not go to the details here)
- In other words, if $\mathrm{epi}\, f$ is convex, then

$$\mathrm{epi}\, f = \bigcap_{H \in \mathcal{H}} H = \bigcap_{i=1} \Big\{ \boldsymbol{x} : \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \geq b_i \Big\}.$$

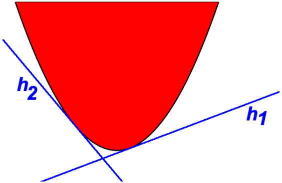
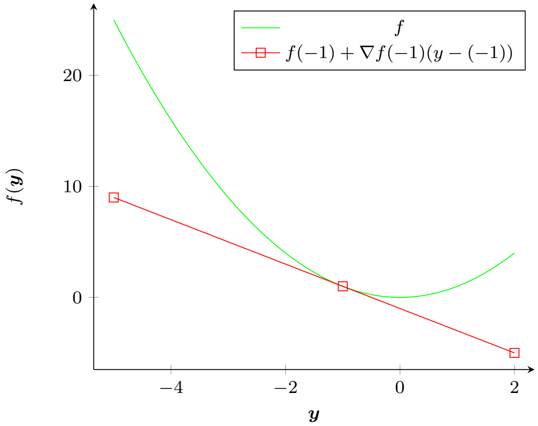


Figure: An illustrative example: two hyperplane $\boldsymbol{h}_1, \boldsymbol{h}_2$

Convexity: the geometry of 1st-order Taylor series

- The halfspace description of $\operatorname{epi} f$ can be translated to an inequality on function
- $f : \operatorname{dom} f \to \mathbb{R}$ is **convex** if :
  1. $\operatorname{dom} f$ is a convex set
  2. $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, we have
  $$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \qquad (*)$$
  i.e. a tangent supports f at a fixed point x
- $(*)$ assumes f is differentiable at $\boldsymbol{x}$. If f is not differentiable at $\boldsymbol{x}$, we generalize gradient to **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.

Convexity: the geometry of supporting hyperplane

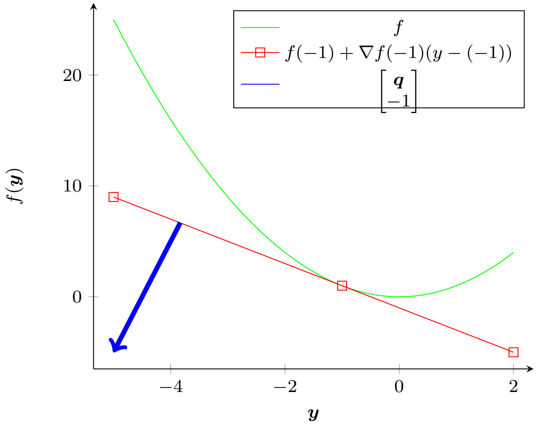
- $f : \mathrm{dom} f \to \mathbb{R}$ is **convex** if :

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

$$\iff \left\langle \begin{bmatrix} \boldsymbol{q} \\ -1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{y} - \boldsymbol{x} \\ f(\boldsymbol{y}) - f(\boldsymbol{x}) \end{bmatrix} \right\rangle \leq 0 \text{ for all } (\boldsymbol{y}, t) \in \mathrm{epi}\, f$$

where $\begin{bmatrix} \boldsymbol{q} \\ -1 \end{bmatrix}$ is the normal of the supporting hyperplane.

- Example. Te figure to the right show a $f : \mathbb{R} \to \mathbb{R}$.
  - Here f is a single variable function, so $\boldsymbol{q}$ is a scalar.
  - The slop of f at $x = -1$ is shown by the red line
  - The slop of f at $x = -1$ is a negative value, say -0.5
  - Therefore the normal $\begin{bmatrix} \boldsymbol{q} \\ -1 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -1 \end{bmatrix}$ points towards the lower left corner, and this arrow is the normal to the supporting hyperplane
  - The term "support" here means the hyperplane just touch $\mathrm{epi}\, f$



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

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

  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.

- This is not a proof 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.

Convexity: gradient is monotone

- A differentiable $f : \mathrm{dom} f \to \mathbb{R}$ is a convex function $\iff \big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \geq 0$.

- A possibly non-differentiable $f : \mathrm{dom} f \to \mathbb{R}$ is a convex function $\iff \big\langle \boldsymbol{x} - \boldsymbol{y}, \partial f(\boldsymbol{x}) - \partial f(\boldsymbol{y}) \big\rangle \geq 0$.

- **Proof** f is convex, so

$$
\begin{aligned}
f(\boldsymbol{x}) &\geq f(\boldsymbol{y}) + \big\langle \partial f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \big\rangle && (1) \\
f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \big\langle \partial f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle && (2) \\
0 &\geq \big\langle \partial f(\boldsymbol{y}) - \partial f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{y} \big\rangle && (1+2) \\
0 &\leq \big\langle \partial f(\boldsymbol{x}) - \partial f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \big\rangle && \text{flip the sign of } (1+2)
\end{aligned}
$$

- What is monotone: a scalar-valued function $g : \mathbb{R} \to \mathbb{R}$ is monotone if $a \geq b$ implies $g(a) \geq g(b)$.
  - $a \geq b$ and $g(a) \geq g(b)$ mean $a - b \geq 0$ implies $g(a) - g(b) \geq 0$, so we have two non-negative things.
  - These two non-negative things can be captured by a single inequality $(a - b)(g(a) - g(b)) \geq 0$.
  - For vector-valued function ∇f, we replace multiplication by inner product, thus $\big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \geq 0$

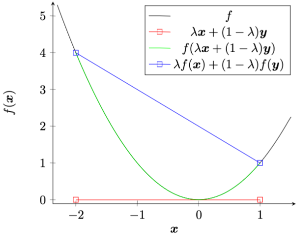
- **Kachurovskii's theorem**: a convex function has monotonic operators as their derivatives.

- Some histroy
  - Kachurovskii, R. I. (1960). "On monotone operators and convex functionals".
  - Minty, G. J. (1964). "On the monotonicity of the gradient of a convex function".

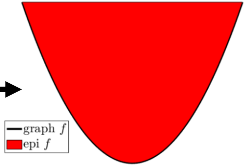
Convexity: a big picture

Function language	**Set language**

``Jansen inequality / inner chord description''



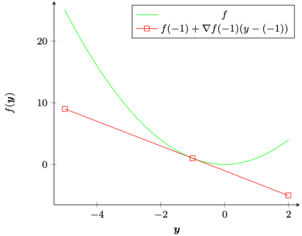
``epigraph is a convex set''



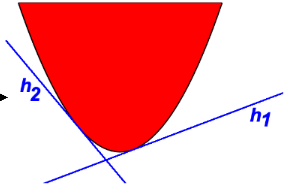
inner chord as an inner line in the epi *f*

convex set is

intersection of halfspaces



a tangential hyperplane is support



``Taylor series is a global support / under estimator''

``epigraph is intersection of halfspaces''

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

- $\mathrm{dom} f$ is a convex set.
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, 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, $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I} \implies \big\langle \boldsymbol{x}-\boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \big\rangle \geq \alpha \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$.

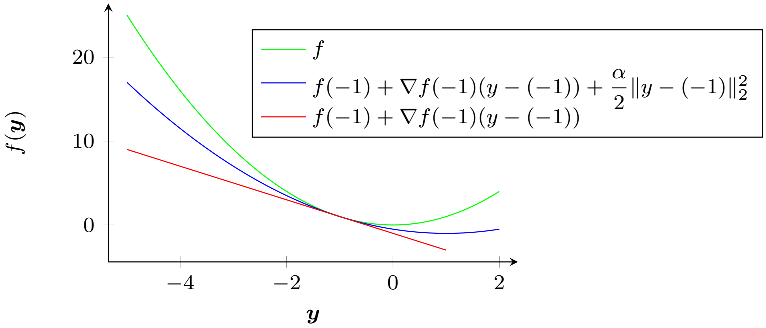
- **Proof**. Recall from calculus $G(b) - G(a) = \int_a^b g(\theta) d\theta$. Next, a smart step, let $\theta = \boldsymbol{y} + \tau(\boldsymbol{x}-\boldsymbol{y})$, then $d\theta = (\boldsymbol{x}-\boldsymbol{y})d\tau$. Consider integral range from 0 to 1 for τ we let G be ∇f and g be $\nabla^2 f$, 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

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

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



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)

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

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

ρ-weakly convex

- **Recall about strong-convexity.** For $\alpha > 0$, a function f is α-strongly convex $\iff$ $f - \frac{\alpha}{2}\|\boldsymbol{x}\|_2^2$ is convex
- **Weak = the opposite of strong.** For $\rho > 0$, a function is ρ-weakly convex $\iff$ $f + \frac{\rho}{2}\|\boldsymbol{x}\|_2^2$ is convex
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, we have any one of the following
  1. f is ρ-weakly convex
  2. 1st-order Taylor series at $\boldsymbol{x}$ is global under-estimator with an additional quadratic term with $\rho > 0$
  $$f(\boldsymbol{y}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle,$$
  or we say f plus a quadratic is lower bounded by a linear function.
  3. Jensen's inequality with an additional quadratic term with $\rho > 0$
  $$f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) + \frac{\rho}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Remarks on convexity ... 1/2

- Strongly convex $\implies$ strictly convex $\implies$ convex $\implies$ weakly 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. (recall if f is strongly convex than there exists a μ such that $f - \frac{\mu}{2}x^2$ is convex, for $f = x^4$, there is no such μ)

- Convexity function needs not to be differentiable.

  - That's why we have epigraph and 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.

- Other convexity

  - log-convex
  - invex
  - pseudoconvex
  - quasiconvex

Remarks on convexity ... 2/2

- Convexity is only about "all local minima are global minima".
- Q: If a function f is convex, is f differentiable?
  A: Differentiability of f has nothing to do with convexity.
- Q: If a function f is convex, does $\min f$ has a solution?
  A: The existence of solution of $\min f$ has nothing to do with convexity.
- Q: If a function f is convex, is the solution $\min f$ unique?
  A: The uniqueness of the solution of $\min f$ has nothing to do with convexity, but it has something to do with strict convexity
- Strict convexity: f has no more than 1 minimum
  - can be none (no minimum)
  - can be 1 (one minimum)
  - no more than 1 (minimum is unique)

Table of Contents

Global Lipschitz continuity

A function $f(\boldsymbol{x}) : \mathrm{dom} f \to \mathbb{R}$ is *globally Lipschitz* if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, 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:

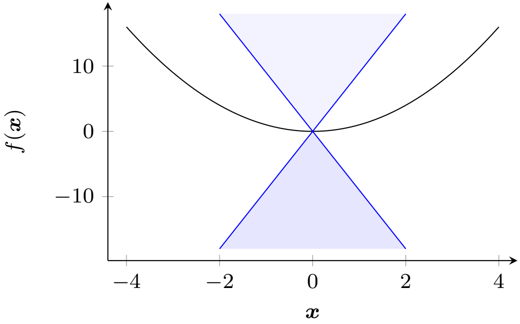
$$\begin{cases} f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + L\|\boldsymbol{x} - \boldsymbol{y}\| \\ f(\boldsymbol{x}) \geq f(\boldsymbol{y}) - L\|\boldsymbol{x} - \boldsymbol{y}\| \end{cases}$$

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

The geometry of global Lipschitz continuity

f is globally Lipschitz $\iff$ f has no sharp change everywhere

$\iff$ $\forall \boldsymbol{x}$ the curve f is entirely outside a cone generated by the two linear functions in the previous 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.

The importance of “global” Lipschitz

- L is defined in the least-upper-bound sense

$$L := \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|} < +\infty$$

- Since L is “global”, so it **holds for any** $\boldsymbol{x}, \boldsymbol{y}$
  - Including derivative case $\boldsymbol{x} \to \boldsymbol{y}$
  - In this case $\left| \frac{df(\boldsymbol{x})}{d\boldsymbol{x}} \right| \leq L$
  - So L is like "the largest slope you can have"
- **holds for any** $\boldsymbol{x}, \boldsymbol{y} \implies L$ is a *pessimistic global constant*
  - Not adaptive to local structure

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.

- Global Lipschitz vs local Lipschitz
  - f is locally Lipschitz at x there exists a neighborhood of x such that f is Lipschitz continuous in thus neighborhood
  - For example, $\sqrt{x}$ in $[0, 1]$ is not globally Lipschitz

---

[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 (globoally) 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

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

  Use triangle inequality[3]

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

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

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

$a \leq |a|$ $\quad$ $a \leq \max\{a, b\}$ $\quad$ $\begin{cases} a \leq M \\ b \leq M \end{cases} \iff \max\{a, b\} \leq M$ $\quad$ $a \leq M$ 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 (blue box) $f_1(\boldsymbol{x}) - f_1(\boldsymbol{y}) \leq L_1\|\boldsymbol{x} - \boldsymbol{y}\|$, which gives

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

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 (red box), (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 (yellow box),

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

L-smooth function / Lipschitz continuous gradient

A function $f : \mathrm{dom} f \to \mathbb{R}$ is L-*smooth* if for any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, 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 $\mathcal{C}_L^{1,1}$.
- "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.
- The word "smooth" ($\mathcal{C}^1$) in machine learning is different from the one used in analysis / manifold, in which smooth means $\mathcal{C}^\infty$ (infinitely differentiable), although all C^1 functions are C^∞ (2nd/higher-order derivative s all equal to zero)

Equivalent definitions of L-smoothness: A function $f(x)$ is L-smooth if

- grad$f$ is L-Lipschitz with $L \geq 0$. I.e. $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$ 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 $\left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \right| \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$.

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 $\left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \right|$

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

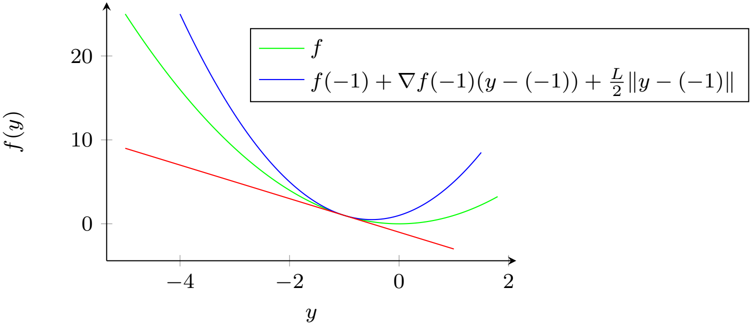
$$
\|\nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \leq L\|\tau(\boldsymbol{y} - \boldsymbol{x})\| \leq L|\tau|\|\boldsymbol{y} - \boldsymbol{x}\| = L\tau\|\boldsymbol{y} - \boldsymbol{x}\|
$$

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

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

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

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

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

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

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

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

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

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

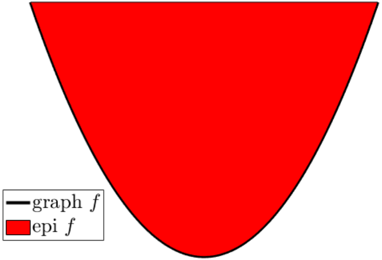
  Details here.

Table of Contents

Epigraph: many properties of f can be translated to the language of epigraph

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)
  - $\mathrm{graph} f := \left\{(x, y) \in \mathbb{R} \times \mathbb{R} \,:\, y = f(x)\right\}$ is a 1d curve in a 2d space.
  - $\mathrm{epi}\, f := \left\{(x, \alpha) \in \mathbb{R} \times \mathbb{R} \,:\, \alpha \geq f(x)\right\}$ is a 2d set in a 2d space.



5 / 28

Details.

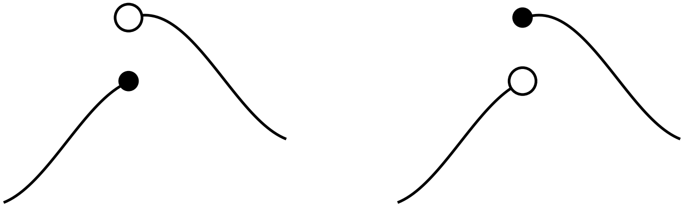
Lower semicontinuity (l.s.c.)

- $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is the extended real line.

- A function is continuous means it has no "jump".

$$
\begin{aligned}
f \text{ is l.s.c. at } \bar{\boldsymbol{x}} \quad &\iff \quad \liminf_{\boldsymbol{x} \to \bar{\boldsymbol{x}}} f(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) \\
&\iff \quad f \text{ allows jump but still continuous if viewed from below} \\
&\iff \quad f \text{ has a closed epigraph}
\end{aligned}
$$

L.S.C. (epi f is closed) Not L.S.C. (epi f is open)



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

Closed, proper function & lower level-bounded

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

  i.e., $f(\boldsymbol{x}) > -\infty \ \forall \boldsymbol{x}$ and $f(\boldsymbol{x}) < +\infty$ for at least one $\boldsymbol{x}$

  OR equivalently, $\operatorname{epi} f \neq \varnothing$ without a vertical 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, $\operatorname{epi} f$ is closed.

- A function f is **lower level-bounded** if if all its level sets are bounded

[4] a vertical line in the graph of f can move downward and touch $-\infty$

argmin (argument of minimum = set of minimizer)

- Given a function f, its argmin is the set of minimizer defined as

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

  Such set can be

  - empty — no minimizer for f
  - singleton — has minimizer for f, unique
  - set-valued (multiple elements) — has minimizers for f, not unique

- IF f is closed convex proper
  THEN $\operatorname{argmin} f$ is closed convex and possibly empty[5]

- IF f is proper, lsc, level bounded
  THEN $\operatorname{argmin} f$ is nonempty and compact. See Theorem 1.9 (attainment of a minimum)[6]

[5] $\operatorname{argmin} f = \varnothing$ that means there is no minimizer for f

[6] Rockafella and Wets, Variational Analysis

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 $\operatorname{epi} f$ is a semialgebraic set.
  - A set is semialgebraic if it is defined by polynomial equations and polynomial inequalities

Cheat sheet

f is proper if $\operatorname{epi} f$ is non-empty and has no vertical line	proper
proper f is closed if $\operatorname{epi} f$ is closed	closedness of proper f
f is l.s.c. if $\operatorname{epi} f$ is closed.	Lower semicontinuous
$\operatorname{argmin} f$ is closed convex if f is closed convex proper	$\operatorname{argmin} f$ closed convex
$\operatorname{argmin} f$ nonempty compact if f is proper, lsc, level bounded	$\operatorname{argmin} f$ nonempty compact
f is convex if $\operatorname{dom} f$ is convex and	
1. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$	Jansen
2. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq 0$	Gradient is monotone
3. $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$	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 $\operatorname{dom} f$ is convex and	
1. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\alpha}{2}\lambda(1-\lambda)\lVert x - y\rVert_2^2$	Jansen
2. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \alpha \lVert x - y\rVert_2^2$	Strongly monotone
3. $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\lVert x - y\rVert_2^2$	Global quadratic lower bound
4. $f(x) - \frac{\alpha}{2}\lVert x\rVert_2^2$ is convex	Convexity
5. $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}$, if f is twice differentiable	Hessian argument
f is ρ-weak convex if $\operatorname{dom} f$ is convex and	
1. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) + \frac{\rho}{2}\lambda(1-\lambda)\lVert x - y\rVert_2^2$	Jansen
3. $f(y) + \frac{\rho}{2}\lVert x - y\rVert_2^2 \geq f(x) + \langle \nabla f(x), y - x\rangle$	1st-order Taylor series is global support
4. $f(x) + \frac{\rho}{2}\lVert x\rVert_2^2$ is convex	Convexity
f is L-Lipschitz gradient (L-smooth) if f is differentiable and	
1. $\lVert \nabla f(x) - \nabla f(y)\rVert \leq L\lVert x - y\rVert$	Definition of Lipschitz
2. $\left\lvert f(y) - f(x) - \langle \nabla f(x), y - x\rangle \right\rvert \leq \frac{L}{2}\lVert y - x\rVert_2^2$	Quadratic inequality
3. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \frac{1}{L}\lVert \nabla f(x) - \nabla f(y)\rVert_2^2$	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. $\lVert \nabla^2 f(x) - \nabla^2 f(y)\rVert \leq L\lVert x - y\rVert$	Definition of Lipschitz
2. $\left\lvert f(x) - f(y) - \langle \nabla f(x), y - x\rangle - \langle \nabla^2 f(x)(y-x), y - x\rangle \right\rvert \leq \dfrac{L}{6}\lVert y - x\rVert_2^3$	Cubic inequality

f is α-strongly convex and β-smooth $\quad \langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\rangle \geq \dfrac{\alpha\beta}{\alpha+\beta}\lVert \boldsymbol{x} - \boldsymbol{y}\rVert_2^2 + \frac{1}{\alpha+\beta}\lVert \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\rVert_2^2$

Read all these to get a Permanent Head Damage

- David G. Luenberger, Optimization by Vector Space Methods (1969)
- Rockafellar, R. Tyrrell, Convex analysis (1970)
- Nemirovskij, Arkadij Semenovic & David Borisovich Yudin, Problem complexity and method efficiency in optimization (1983).
- Jean-Baptiste Hiriart-Urruty & Claude Lemarechal, Convex Analysis and Minimization Algorithms I & II (1993)
- Rockafellar, R. Tyrrell & Roger J-B. Wets, Variational analysis (1997,2004,2009)
- Yurii Nesterov, Introductory Lectures on Convex Optimization (1998)
- Ivar Ekeland and Roger Témam Convex Analysis and Variation Problems (1999)
- Jorge Nocedal & Stephen Wright Numerical Optimization (1999,2006)
- Jonathan M. Borwein & Adrian S. Lewis, Convex analysis and nonlinear optimization (2000)
- A Ben-Tal, Lectures on modern convex optimization (2001)
- Dimitri Bertsekas & Angelia Nedic & Asuman Ozdaglar, Convex Analysis and Optimization (2003)
- Stephen Boyd & Lieven Vandenberghe, Convex optimization (2004)
- P.-A. Absil, R. Mahony & Rodolphe Sepulchre, Optimization Algorithms on Matrix Manifolds (2008)
- Boris Mordukhovich & Mau Nam Nguyen An Easy Path to Convex Analysis and Applications (2013), Convex Analysis and Beyond (2022)
- Heinz H. Bauschke and Patrick L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces (2011)
- Amir Beck, First-Order Methods in Optimization (2017)
- Boris Mordukhovich Variational Analysis and Applications (2018)
- Nicolas Boumal, An Introduction to Optimization on Smooth Manifolds (2023)
- Ernest K. Ryu & Wotao Yin, Large-Scale Convex Optimization Algorithms and Analyses via Monotone Operators (2023)

End of document