Some special classes of function in optimization

Lipschitz, convex, α-strongly convex, β-smooth, etc.

Andersen Ang

Department of Combinatorics and Optimization,
University of Waterloo, Waterloo, Canada

ms$\mathbf{x}$ang@uwaterloo.ca, where $\mathbf{x} = \lfloor \pi \rfloor$
Homepage: angms.science

First draft: June 6, 2017
Last update: March 17, 2021

Table of Contents

Convex function

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

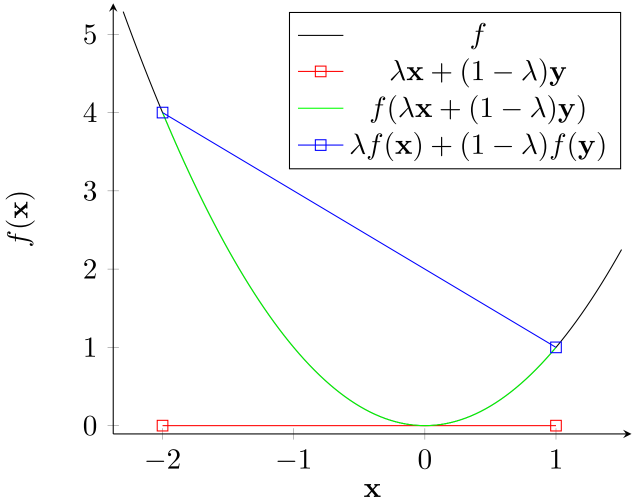
- $\mathrm{dom} f$ is a convex set
- $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, f satisfies
  - Jensen's inequality
$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$
  - Gradient of f is monotonic
$$\Big\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \Big\rangle \geq 0.$$
  - 1st-order Taylor approximation at point $\mathbf{x}$ is a global under-estimator
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \Big\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \Big\rangle.$$
  - Epigraph of f is a convex set.
- f is **strictly convex** if $\leq, \geq$ became $<, >$ (i.e. strict inequality).
- The 4 definitions are equivalent: you can move from one definition to another as "if and only if". See optimization books for the proof of equivalence between these 4 definitions.

Convexity: the geometry of Jensen's inequality

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

(1) $\mathrm{dom} f$ is a convex set and

(2) $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom} f, f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$

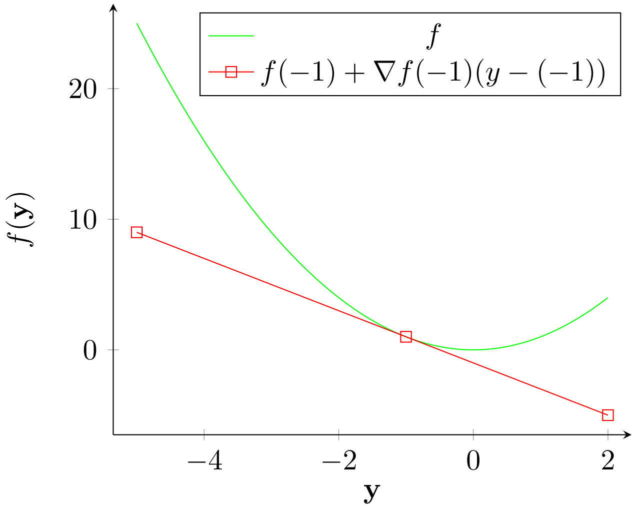


Convexity: the geometry of 1st-order Taylor approximation

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

(1) $\mathrm{dom} f$ is a convex set and

(2) $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom} f, f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$



Why if f convex then $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$?

- Why convex f are lower-bounded by their own 1st-order Taylor approximation? Simple explanation using Taylor series.
- Consider the pedagogical case that f is a function 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 setting.
- Note: this is not a prove but an illustration, because we didn't show that f is convex $\iff$ its Hessian is positive semi-definite.

Table of Contents

α-strongly convex function

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

- $\mathrm{dom} f$ is a convex set.
- $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, f satisfies
  - Jensen's inequality with an additional quadratic term with $\alpha > 0$

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

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

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

  - 1st-order Taylor approximation at point $\mathbf{x}$ is global under-estimator with an additional quadratic term with $\alpha > 0$

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

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

Equivalence between definitions of strong convexity

We show $\nabla^2 f(\mathbf{x}) \succeq \alpha \mathbf{I} \implies \big\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \big\rangle \geq \alpha \|\mathbf{x}-\mathbf{y}\|_2^2$, $\alpha > 0$.

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

$$\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) = \int_0^1 \nabla^2 f\big(\mathbf{y} + \tau(\mathbf{x}-\mathbf{y})\big)(\mathbf{x}-\mathbf{y}) d\tau.$$

(left hand side is a vector, right hand side is matrix-vector product, also a vector)

Take dot product with $\mathbf{x}-\mathbf{y}$ on the whole equation on both sides

$$\big\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \big\rangle = \Big\langle \mathbf{x}-\mathbf{y}, \int_0^1 \nabla^2 f\big(\mathbf{y} + \tau(\mathbf{x}-\mathbf{y})\big)(\mathbf{x}-\mathbf{y}) d\tau \Big\rangle.$$

By $\nabla^2 f(\mathbf{x}) \succeq \alpha \mathbf{I}$ for all $\mathbf{x}$, we have $\nabla^2 f\big(\mathbf{y} + \tau(\mathbf{x}-\mathbf{y})\big) \succeq \alpha \mathbf{I}$ and

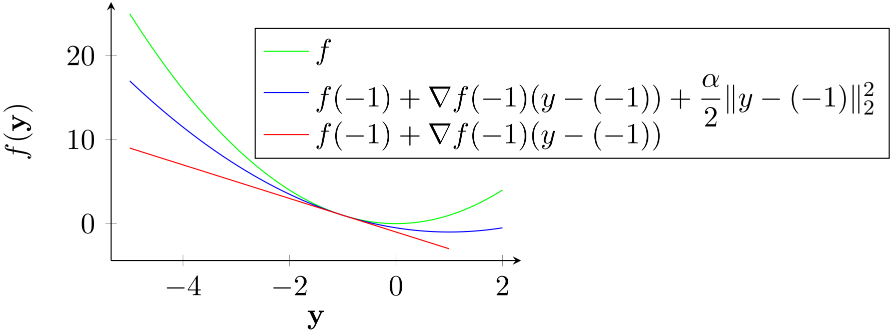
$$\big\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \big\rangle \geq \Big\langle \mathbf{x}-\mathbf{y}, \int_0^1 \alpha(\mathbf{x}-\mathbf{y}) d\tau \Big\rangle = \alpha \|\mathbf{x}-\mathbf{y}\|_2^2. \quad \square$$

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 set and

(2) for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$



Interpretation: f is lower bounded by a quadratic curve with some curvature, which is also lower bounded by the 1st order Taylor approximation (zero curvature) $\Longrightarrow$ f is not "too flat" (at least not "as flat as" the lower bound). In other words: f is at least α-amount of "bumpy".

Table of Contents

Lipschitz continuity

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

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

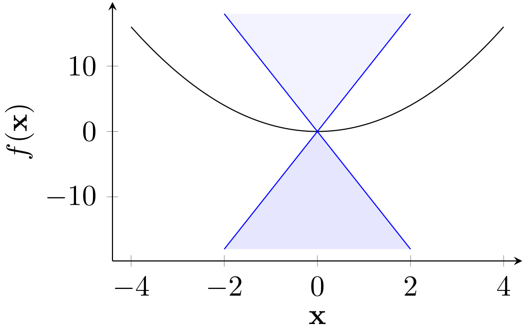
- Re-arrange gives $\dfrac{|f(x) - f(y)|}{\|x - y\|} \leq L$, which is approximately the magnitude of the gradient when $\mathbf{x}, \mathbf{y}$ are close $\implies f$ is Lipschitz means the "slope" (rate of change) of f is bounded above by a global constant L.
- Removing the absolute value sign:

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

  meaning that f for all $\mathbf{x}$ is bounded above and below by a linear function.

The geometry of Lipschitz continuity

A function is Lipschitz means function does not have sharp changes everywhere: $\forall \mathbf{x}$, the function value 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.

Table of Contents

L-smooth function / Lipschitz continuous gradient

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

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

- This assume f is differentiable.
- "f is L-smooth" is also called L-Lipschitz gradient.
- "f is L-smooth" is equivalent to

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

Removing the absolute value sign gives

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

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

These definitions are equivalent. e.g., take the norm of the 3rd condition gives the 1st condition.

Proof of equivalence

We show for $L > 0$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ implies

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

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

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

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

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

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

Proof of equivalence - continue

$$\begin{array}{rcl} |f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| & = & | \int_0^1 \langle \nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \, d\tau | \\ & \leq & \int_0^1 \big| \langle \nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \big| d\tau \\ & \overset{\text{c.s.}}{\leq} & \int_0^1 \|\nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\| \cdot \|\mathbf{y} - \mathbf{x}\| d\tau. \end{array}$$

c.s. means Cauchy – Schwarz inequality.

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

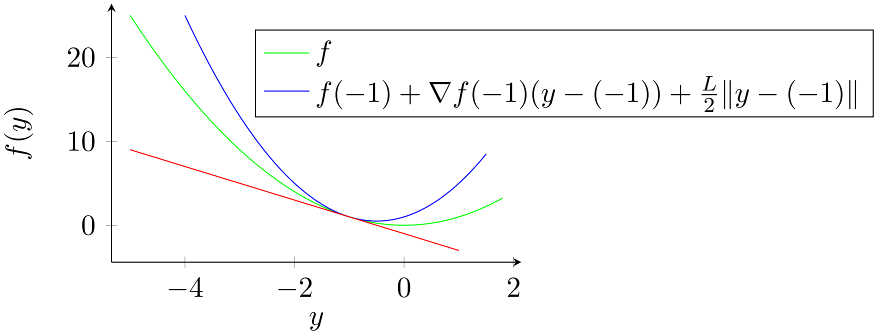
$$\|\nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\| \leq L\|\tau(\mathbf{y} - \mathbf{x})\| \leq L|\tau|\|\mathbf{y} - \mathbf{x}\| = L\tau\|\mathbf{y} - \mathbf{x}\|$$

where $\|\tau(\mathbf{y} - \mathbf{x})\| = |\tau|\|\mathbf{y} - \mathbf{x}\|$ 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

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

L-smoothness: the geometry of the upper bound

A function f is L-**smooth** if for any two points $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$,
$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$



Interpretation : 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".

Table of Contents

Relative-smooth function

- f is L-smooth

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

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

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

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

- Why relative smoothness: it has application for proving convergence of gradient descent on non-Euclidean geometry.

Table of Contents

Lipschitz continuous Hessian

A function $f(\mathbf{x}) : \mathrm{dom} f \to \mathbb{R}$ has L-Lipschitz Hessian, if for any two points $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, there exists a constant L (the Lipschitz constant) such that

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

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

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

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

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

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

Table of Contents

Last page - summary

f is convex if dom$f$ is convex and

1. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$
2. $\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq 0$
3. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

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

1. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\alpha}{2}\lambda(1-\lambda)\|x-y\|_2^2$
2. $\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \alpha\|x-y\|_2^2$
3. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x-y\|_2^2$
4. $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex
5. $\nabla^2 f(\mathbf{x}) \succeq \alpha \mathbf{I}$, if f is twice differentiable

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

1. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$
2. $\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \leq \frac{L}{2}\|y-x\|_2^2$
3. $\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$
4. $\nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$, if f is twice differentiable

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

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

End of document