Math 273a: Optimization
Convex Functions

Instructor: Wotao Yin

Department of Mathematics, UCLA

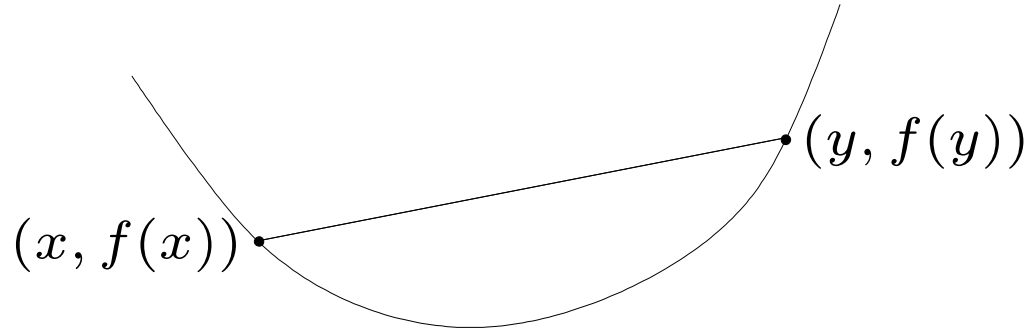
Fall 2015

online discussions on piazza.com

Definition

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \quad \forall\ \mathbf{x}, \mathbf{y}, \lambda \in [0,1]$$



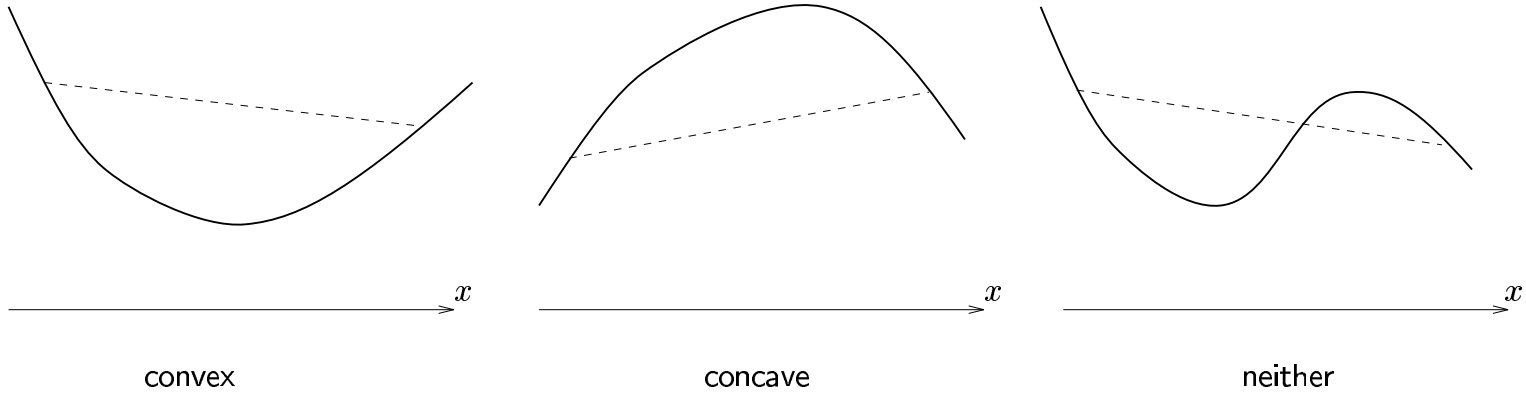
- modern definition: $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ but is proper (not identically ∞)

  then, we can ignore $\mathbf{dom}(f)$. also, f can include the indicator function of convex set S:

$$\iota_S(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in S \\ \infty, & \text{otherwise} \end{cases}$$

Concave function

- A function f is concave if $(-f)$ is convex.

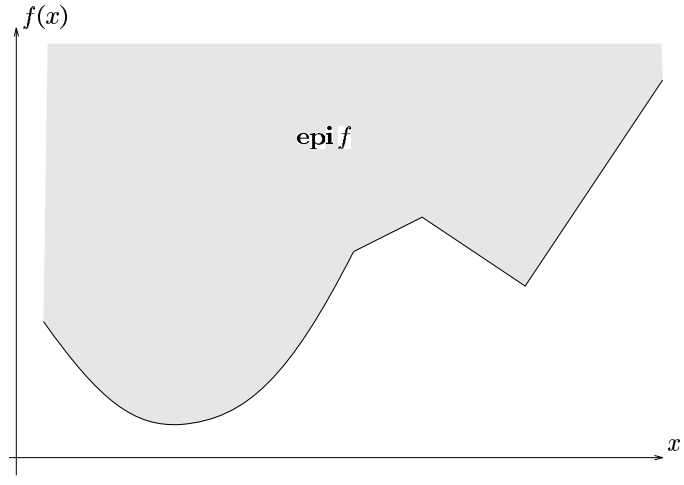


convex concave neither

Epigraph

- definition:

$$\mathbf{epi}(f) := \{(\mathbf{x}, t) : \mathbf{x} \in \mathbf{dom} f, t \geq f(\mathbf{x})\}$$



- f is convex if and only if $\mathbf{epi}(f)$ is a convex set
- "lifts" the convex function and enables set-operations such as projection, etc., enriching understanding means and numerical tools

$\mathcal{F}^1$: the set of C^1 convex functions

- A continuously differentiable function $f(\mathbf{x})$ is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T \nabla f(\mathbf{x}) \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

interpretation: $f(\mathbf{y})$ is on or above its linear support function at any point

- a twice continuously differentiable function $f(x)$ is convex if and only if

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \text{for any } \mathbf{x} \in \mathbb{R}^n$$

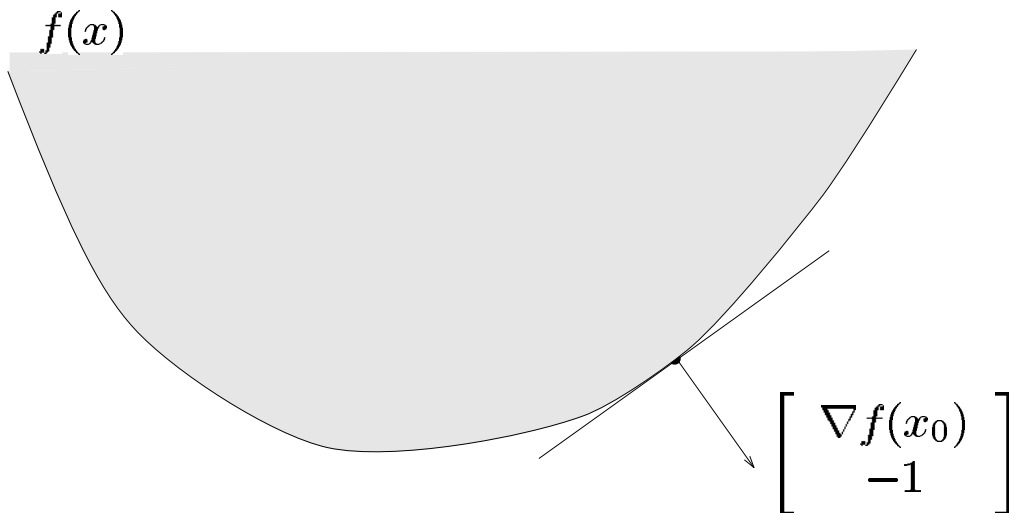
- $\mathcal{F}^k(\mathbb{R}^n)$: the set of *convex* functions on $\mathbb{R}^n$ that are k times continuously differentiable. [notation used by Nesterov'03 textbook]

- (we will show these later)

Epigraph and support hyperplane

- if $(\mathbf{x}, t) \in \mathbf{epi}(f)$, then

$$\begin{bmatrix} \nabla f(\mathbf{x}_0) \\ -1 \end{bmatrix}^T \begin{bmatrix} \mathbf{x} - \mathbf{x}_0 \\ t - f(\mathbf{x}_0) \end{bmatrix} \leq 0$$



Examples in $\mathbb{R}$

- x^α is convex over $x \geq 0$ for $\alpha \geq 1$
- $|x|$ is convex (but not differentiable)
- $\log x$ is concave, and $x \log x$ is convex, over $x > 0$
- e^x is convex
- $\max(0, x)$ and $\max(0, -x)$ are both convex

Basic properties

- f is convex if and only if $g(t) := f(\mathbf{x} + t\mathbf{d})$ is convex for all $\mathbf{x}, \mathbf{d}$
- if f is convex, then αf is convex for all $\alpha \geq 0$
- if f_1, f_2 are convex, then so is $f_1 + f_2$
- extends to infinite sums: if $g(x, y)$ is convex in x for each y, then $\int g(x, y)\mathrm{d}y$ is convex
- if f_1, f_2 are convex, then so is $h(x) = \max\{f_1(x), f_2(x)\}$
  (also extends to point-wise supremum of infinitely many convex functions)
- affine transform to x: if f is convex, so is $f(Ax + b)$

Examples in $\mathbb{R}^n$

- linear functions

$$f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$$

- quadratic functions: $Q \succeq 0$

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{a}^T\mathbf{x} + b$$

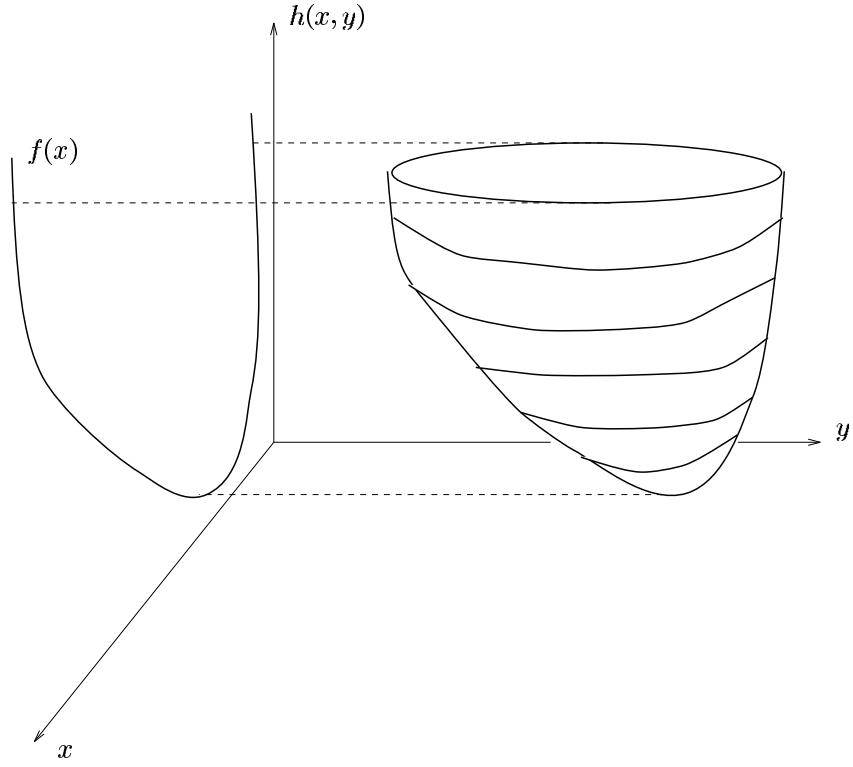
- all norms
  - $\|\mathbf{x}\|_1$
  - $\|\mathbf{x}\|_2$
  - $\|\mathbf{x}\|_\infty$
- max distance to any set (convex or not): $f(\mathbf{x}) = \sup_{\mathbf{s}\in S} \|\mathbf{s} - \mathbf{x}\|$

Partial min of jointly-convex functions

- if $h(\mathbf{x}, \mathbf{y})$ is convex (i.e., jointly convex in $(\mathbf{x}, \mathbf{y})$), then

$$f(\mathbf{x}) := \inf_{\mathbf{y}} h(\mathbf{x}, \mathbf{y})$$

  is convex. (corresponds to epigraph projection)



- min distance to a convex set, $f(\mathbf{x}) = \inf_{\mathbf{x} \in S} \|\mathbf{s} - \mathbf{x}\|$, is convex.

  proof: consider $h(\mathbf{x}, \mathbf{s}) = \|\mathbf{s} - \mathbf{x}\| + \iota_S(\mathbf{s})$, which is (jointly) convex. Then,

$$f(\mathbf{x}) = \inf_{\mathbf{s}} h(\mathbf{x}, \mathbf{s})$$

- infimal post-composition, $f(\mathbf{y}) = \inf\{g(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{y}\}$, is convex

  proof:: consider $h(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \iota_{\{0\}}(\mathbf{A}\mathbf{x} - \mathbf{y})$, which is (jointly) convex. Then,

$$f(\mathbf{y}) = \inf_{\mathbf{x}} h(\mathbf{x}, \mathbf{y})$$

Jensen's inequality

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex.

- two points: for $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$,

$$f(\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2) \leq \lambda_1 f(\mathbf{x}_1) + \lambda_2 f(\mathbf{x}_2)$$

- multiple points: for $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$,

$$f\Big(\sum_i \lambda_i \mathbf{x}_i\Big) \leq \sum_i \lambda_i f(\mathbf{x}_i)$$

- continuous version: for distribution $p(x) \geq 0$, $\int p(x)\mathrm{d}x = 1$,

$$f(\mathbf{E}x) = f\left(\int x p(x)\mathrm{d}x\right) \leq \int f(x)p(x)\mathrm{d}x = \mathbf{E}(f(x)).$$

application: $f(x) = -\log(x)$ is convex over $x > 0$. Then, for $a, b > 0$,

$$-\frac{1}{2}(\log a + \log b) \geq -\log\big((a+b)/2\big) \implies \sqrt{ab} \leq (a+b)/2$$

(of course, extends to $a, b \geq 0$.)

Properties of $f, g \in \mathcal{F}^1(\mathbb{R}^n)$

- Recall: a C^1 function $f(\mathbf{x})$ is convex on $\mathbb{R}^n$ if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T \nabla f(\mathbf{x}) \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(function $f(\mathbf{y})$ is on or above its linear support function at any point)

- **Corollary**: $\nabla f(\mathbf{x}^*) = 0$ if and only if $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x}$.
- $\mathcal{F}^1$ is closed under linear operation: if $f, g \in \mathcal{F}^1$, then $\alpha f + \beta g \in \mathcal{F}^1$, $\alpha, \beta \geq 0$
- For matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^n$, $f \in \mathcal{F}^1$, we have

$$h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b}) \in \mathcal{F}^1.$$

Equivalent Definitions for $\mathcal{F}^1(\mathbb{R}^n)$

Theorem

The followings are equivalent:

1. $f \in \mathcal{F}^1(\mathbb{R}^n)$, *i.e.,* f *is convex and in* $C^1(\mathbb{R}^n)$.
2. $f \in C^1(\mathbb{R}^n)$ *and for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *and* $0 \leq \alpha \leq 1$,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$

3. $f \in C^1(\mathbb{R}^n)$ *and monotone, i.e., for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$(\mathbf{x} - \mathbf{y})^T (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})) \geq 0. \tag{1}$$

Theorem

Function $f \in \mathcal{F}^2(\mathbb{R}^n)$ *if and only if it is in* $C^2(\mathbb{R}^n)$ *and*

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{for any } \mathbf{x} \in \mathbb{R}^n.$$

Proof of Theorem 2.

($\Rightarrow$) : Fix $\mathbf{x}$ and any direction $\mathbf{s}$. Let $\mathbf{x}_t = \mathbf{x} + t\mathbf{s}$ (for $t > 0$). From (1),

$$0 \leq \frac{1}{t^2}(\mathbf{x}_t - \mathbf{x})^T(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x})) = \frac{1}{t^2}(t\mathbf{s}^T)\left(\int_0^t \nabla^2 f(\mathbf{x} + \tau\mathbf{s})\mathbf{s}\mathrm{d}\tau\right) =$$

$$= \frac{1}{t}\int_0^t \mathbf{s}^T\nabla^2 f(\mathbf{x} + \tau\mathbf{s})\mathbf{s}d\tau$$

Letting $t \to 0$ and using continuity, we conclude that $\mathbf{s}^T\nabla^2 f(\mathbf{x})\mathbf{s} \geq 0$.

($\Leftarrow$) : Suppose $\mathbf{y}, \mathbf{x} \in \mathbb{R}^n$. Observe that

$$f(\mathbf{y}) = f(\mathbf{x}) + \int_0^1 (\mathbf{y} - \mathbf{x})^T\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\mathrm{d}t$$

$$= f(\mathbf{x}) + \int_0^1 (\mathbf{y} - \mathbf{x})^T\left(\int_0^t \nabla^2 f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})d\tau + \nabla f(\mathbf{x})\right)\mathrm{d}t$$

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

$$\geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T\nabla f(\mathbf{x}).$$

□

$\mathcal{F}_L^{k,p}(\mathbb{R}^n)$: Lipschitz continuous convex functions

- $\mathcal{F}_L^{k,p}(\mathbb{R}^n) \subset \mathcal{F}^k(\mathbb{R}^n)$: the pth derivative of f is Lipschitz with constant $L \geq 0$; that is,

$$\|f^{(p)}(\mathbf{x}) - f^{(p)}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(the rate of change of $f^{(p)}$ is bounded)

Equivalent Definitions for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Theorem

The followings are equivalent:

1. $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$.
2. f *is in* $C^1(\mathbb{R}^n)$ *and for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$0 \leq f(\mathbf{y}) - f(\mathbf{x}) - (\mathbf{y} - \mathbf{x})^T \nabla f(\mathbf{x}) \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

3. f *is in* $C^1(\mathbb{R}^n)$ *and for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T \nabla f(\mathbf{x}) + \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

4. f *is in* $C^1(\mathbb{R}^n)$ *and for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq (\mathbf{x} - \mathbf{y})^T (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})).$$

Proof of Theorem 3.

$(1) \Rightarrow (2)$: Observe that

$$
\begin{aligned}
f(\mathbf{y}) &= f(\mathbf{x}) + \int_0^1 (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))\mathrm{d}t \\
&= f(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x}) + \int_0^1 (\mathbf{y}-\mathbf{x})^T (\nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x}))dt \\
&\leq f(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x}) + \int_0^1 \|\mathbf{y}-\mathbf{x}\| \|\nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\|dt \\
&\leq f(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x}) + \int_0^1 \|\mathbf{y}-\mathbf{x}\| tL \|\mathbf{y}-\mathbf{x}\|dt \\
&= f(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2
\end{aligned}
$$

$(2) \Rightarrow (3)$: Fix $\mathbf{x}_0 \in \mathbb{R}^n$. Consider convex function $\phi(\mathbf{y}) := f(\mathbf{y}) - \mathbf{y}^T \nabla f(\mathbf{x}_0)$ which achieves it minimum at $\mathbf{y}^* = \mathbf{x}_0$ since $\nabla \phi(\mathbf{y}^*) = \nabla f(\mathbf{y}^*) - \nabla f(\mathbf{x}_0) = \mathbf{0}$. In particular,

$$\phi(\mathbf{y}^*) \leq \phi(\mathbf{y} - \frac{1}{L}\nabla \phi(\mathbf{y}))$$

Applying assumption (2) to ϕ (with $\mathbf{y} := \mathbf{y} - \frac{1}{L}\nabla\phi(\mathbf{y})$, and $\mathbf{x} := \mathbf{y}$)

Continue of Proof of Theorem 3.

$$\phi(\mathbf{y} - \frac{1}{L}\nabla\phi(\mathbf{y})) \leq \phi(\mathbf{y}) - \frac{1}{2L}\|\nabla\phi(\mathbf{y})\|^2.$$

Hence,

$$\phi(\mathbf{x}_0) \leq \phi(\mathbf{y}) - \frac{1}{2L}\|\nabla\phi(\mathbf{y})\|^2.$$

Substituting and using the fact that $\nabla\phi(\mathbf{y}) = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}_0)$ implies (3).

$(3) \Rightarrow (4)$: Adding two inequalities (3) with $\mathbf{x}$ and $\mathbf{y}$ interchanged.

$(4) \Rightarrow (1)$: Notice from Theorem 1.3 that f is convex. Next we show the Lipschitz bound. From assumption (4) and Cauchy-Schwarz we have

$$\frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq (\mathbf{x}-\mathbf{y})^T(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})) \leq \|\mathbf{x}-\mathbf{y}\|\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|.$$

Thus,

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

□

Strongly Convex Function

A continuously differentiable function $f(\mathbf{x})$ is *μ-strongly convex* on $\mathbb{R}^n$ if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T \nabla f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

- $\mathcal{S}_\mu^k(\mathbb{R}^n)$: the set of μ-strongly convex functions on $\mathbb{R}^n$ that are k times continuously differentiable.

Properties:

- If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$ and $\nabla f(\mathbf{x}^*) = 0$ then for all $\mathbf{x}$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$$

- If $f_1 \in \mathcal{S}_{\mu_1}^1(\mathbb{R}^n)$, $f_2 \in \mathcal{S}_{\mu_2}^1(\mathbb{R}^n)$ and $\alpha, \beta \geq 0$ then

$$\alpha f_1 + \beta f_2 \in \mathcal{S}_{\alpha\mu_1 + \beta\mu_2}^1(\mathbb{R}^n).$$

Equivalent Definitions for $\mathcal{S}^1_\mu(\mathbb{R}^n)$

Theorem (4)

The followings are equivalent:

1. $f \in \mathcal{S}^1_\mu(\mathbb{R}^n)$ *(i.e.* f *is strongly convex).*
2. f *is in* $C^1(\mathbb{R}^n)$ *and for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *and* $0 \leq \alpha \leq 1$,

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

3. f *is in* $C^1(\mathbb{R}^n)$ *and for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$(\mathbf{x}-\mathbf{y})^T(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})) \geq \mu\|\mathbf{x}-\mathbf{y}\|^2.$$

Theorem (5)

Function $f \in \mathcal{S}^2_\mu(\mathbb{R}^n)$ *if and only if it is in* $C^2(\mathbb{R}^n)$ *and for any* $\mathbf{x} \in \mathbb{R}^n$ *we have:*

$$\nabla^2 f(\mathbf{x}) \succeq \mu I_n.$$

Application: performance of gradient methods

"black box" optimization (Nesterov'03 textbook):

- Problem to be solved (e.g., find minimizer).
- Class of functions (e.g., $\mathcal{F}^1(\mathbb{R}^n)$, $\mathcal{F}_L^{k,p}(\mathbb{R}^n)$, $\mathcal{S}_\mu^1(\mathbb{R}^n)$).
- Oracle: black box information of the problem. For example,

$$\begin{aligned} 0\text{th order:} \quad & \mathbf{x}_i \to f(\mathbf{x}_i) \\ 1\text{st order:} \quad & \mathbf{x}_i \to (f(\mathbf{x}_i), \nabla f(\mathbf{x}_i)) \\ 2\text{nd order:} \quad & \mathbf{x}_i \to (f(\mathbf{x}_i), \nabla f(\mathbf{x}_i), \nabla^2 f(\mathbf{x}_i)). \end{aligned}$$

- Solution: exact or approximate condition (i.e. $\|\mathbf{x}^* - \mathbf{x}_{\text{sol}}\| < \epsilon$).
- Algorithm using only oracle to compute or approximate $\mathbf{x}_{\text{sol}}$.

Example:

- Problem: find root of function
- Function class: all polynomial
- Oracle: 0th order
- Method: always return $\mathbf{x}^* = 0$.

Performance: cannot be beaten for some functions in the class, arbitrarily bad performance for other functions.

How to gauge performance of a method?

Use the computational cost for *the worst function* in a chosen function class (e.g. $\mathcal{F}^1(\mathbb{R}^n), \mathcal{F}_L^{k,p}(\mathbb{R}^n), \mathcal{S}_\mu^1(\mathbb{R}^n)$).

- Analytical complexity: number of calls to the oracle which is required to solve the problem up to the desired accuracy.
- Arithmetical complexity: number of arithmetic operations (including work of oracle and method) to solve the problem up to desired accuracy.

For our purposes

- Problem: $\text{minimize}_{\mathbf{x}\in\mathbb{R}^n}\ f(\mathbf{x})$.
- Function class: $f \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$ or $\mathcal{S}_\mu^1(\mathbb{R}^n)$.
- Oracle: first order
- Approximate solution: $\|f(\bar{\mathbf{x}}) - f^*\| < \epsilon$.
- Methods: Sequence $\{\mathbf{x}_k\}$ such that

$$\mathbf{x}_k \in \mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \nabla f(\mathbf{x}_1), \ldots, \nabla f(\mathbf{x}_{k-1})\}$$

(methods include gradient descent methods.)

Lower Complexity Bounds

Theorem

*For any k, $1 \leq k \leq (n-1)/2$, and any $\mathbf{x}_0 \in \mathbb{R}^n$, **there exists a function** $f \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$ such that for all first order methods generating a sequence*

$$\mathbf{x}_k \in \mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \nabla f(\mathbf{x}_1), \ldots, \nabla f(\mathbf{x}_{k-1})\},$$

we have

$$f(\mathbf{x}_k) - f^* \geq \frac{3L\|\mathbf{x} - \mathbf{x}^*\|^2}{32(k+1)^2},$$

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \geq \frac{1}{32}\|\mathbf{x}_0 - \mathbf{x}^*\|^2.$$

Gradient Method

Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \quad \text{where } f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n).$$

Scheme:

0. Choose $\mathbf{x}_0 \in \mathbb{R}^n$.
1. k^{th} iteration $(k \geq 0)$.
   a. Compute $f(\mathbf{x}_k)$ and $\nabla f(\mathbf{x}_k)$
   b. Set $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - h_k \nabla f(\mathbf{x}_k)$

To simplify the following work, we assume $h_k = h > 0$ for every k.

Upper Bound for Gradient Method

Theorem

If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ *and* $0 < h < \frac{2}{L}$ *then*

$$f(\mathbf{x}_k) - f^* \leq \frac{(f(\mathbf{x}_0) - f^*)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + (f(\mathbf{x}_0) - f^*)h(1 - \frac{L}{2}h)k}$$

Proof:

Step 1. Boundedness of $\{\mathbf{x}_k\}$**.** Let $\varepsilon_k := \|\mathbf{x}_k - \mathbf{x}^*\|$. Then

$$\begin{aligned}
\varepsilon_{k+1}^2 &= \|\mathbf{x}_k - \mathbf{x}^* - hf'(\mathbf{x}_k)\|^2 \\
&= \varepsilon_k^2 - 2h\langle f'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle + h^2\|f'(\mathbf{x}_k)\|^2 \\
&\leq \varepsilon_k^2 - h\left(\frac{2}{L} - f\right)\|f'(\mathbf{x}_k)\|^2
\end{aligned}$$

Where we used the 4^{th} equivalent definition of $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $\nabla f(\mathbf{x}^*) = 0$ in the last line. Therefore $\varepsilon_{k+1} \leq \varepsilon_k \leq \ldots \leq \varepsilon_0$.

Note: ∇f cannot vary too quickly or $\mathbf{x}_k$ go unbounded.

Proof continued

Step 2: Objective descent.

The second equivalent definition of $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ tells us

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), -h \nabla f(\mathbf{x}_k) \rangle + \frac{L}{2} h^2 \|\nabla f(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - h(1 - \frac{L}{2} h) \|\nabla f(\mathbf{x}_k)\|^2 \end{aligned}$$

To simplify notation, define $\omega := h(1 - \frac{L}{2} h)$. Also define $\Delta_k := f(\mathbf{x}_k) - f^*$. Subtracting f^* from both sides of the above inequality yields

$$\Delta_{k+1} \leq \Delta_k - \omega \|\nabla f(\mathbf{x}_k)\|^2 \tag{2}$$

Proof continued

Step 3: The descent is "sufficient."

We shall bound $\Delta(\mathbf{x})$ by $\|\nabla f(\mathbf{x})\|$, so there is sufficient descent to estimate the descent rate of Δ_k.

From the definition of convexity,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle$$

Thus

$$\Delta_k \leq -\langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle \leq \|\nabla f(\mathbf{x}_k)\| \varepsilon_k$$

$$\Leftrightarrow \quad -\omega \|\nabla f(x_k)\|^2 \leq -\frac{\omega}{\varepsilon_0^2} \Delta_k^2. \tag{3}$$

Inequalities (2) and (3) together give:

$$\Delta_{k+1} \leq \Delta_k - \frac{\omega}{\varepsilon_0^2} \Delta_k^2.$$

Proof continued

Step 4: Rate establishment.

Recall that f^* is the minimum of f, hence $\Delta_j \geq 0$. Divide the inequality by $\Delta_{k+1}\Delta_k$, rearrange and recall $\{\Delta_k\}$ decreases in k to get:

$$\begin{aligned}\frac{1}{\Delta_{k+1}} &\geq \frac{1}{\Delta_k} + \frac{\omega}{\varepsilon_0^2} \cdot \frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{\varepsilon_0^2} \\ &\geq \ldots \geq \frac{1}{\Delta_0} + \frac{\omega}{\varepsilon_0^2}(k+1)\end{aligned}$$

Reindex at k and invert the inequality to finally obtain

$$\Delta_k \leq \frac{\Delta_0 \varepsilon_0^2}{\varepsilon_0^2 + \Delta_0 \omega k}$$

which, miraculously, is what we were to show.

Performance bounds

Further analysis can prove the following:

1. The gradient method is not optimal for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$.
2. The gradient method is not optimal for $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$.

Notes:

- All standard NLP methods (conjugate gradients, variable metric) have similar *lower* efficiency estimates.
- The bounds do not necessarily reflect real-world performance, and often not.
- The lecture is not finished yet. More on the way.