Recap

Pontus Giselsson

Outline

- Convex analysis
- Composite optimization and duality
- Solving composite optimization problems – Algorithms

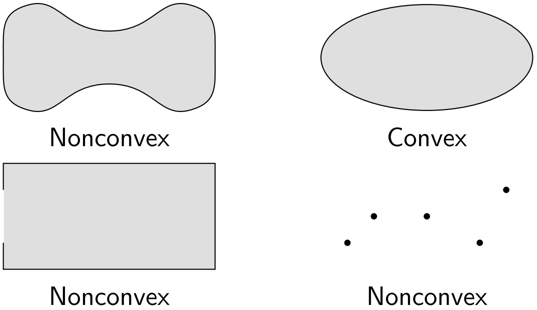
Convex Analysis

Convex sets

- A set C is convex if for every $x, y \in C$ and $\theta \in [0,1]$:

$$\theta x + (1-\theta)y \in C$$

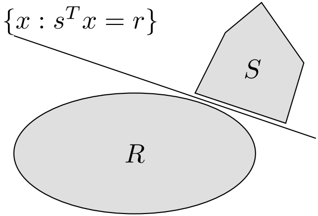
- "Every line segment that connect any two points in C is in C"



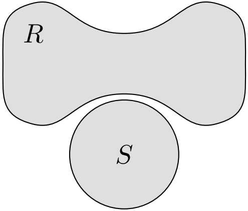
- Will assume that all sets are nonempty and closed

Separating hyperplane theorem

- Suppose that $R, S \subseteq \mathbb{R}^n$ are two non-intersecting convex sets
- Then there exists hyperplane with S and R in opposite halves



Example



Counter-example
R nonconvex

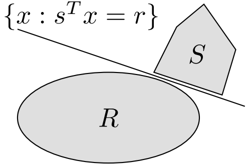
- Mathematical formulation: There exists $s \neq 0$ and r such that

$$
\begin{aligned}
s^T x &\leq r \qquad\qquad \text{for all } x \in R \\
s^T x &\geq r \qquad\qquad \text{for all } x \in S
\end{aligned}
$$

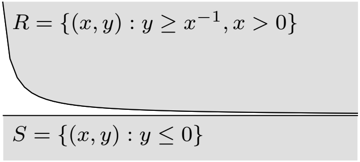
- The hyperplane $\{x : s^T x = r\}$ is called *separating hyperplane*

A strictly separating hyperplane theorem

- Suppose that $R, S \subseteq \mathbb{R}^n$ are non-intersecting closed and convex sets and that one of them is compact (closed and bounded)
- Then there exists hyperplane with strict separation



Example



Counter example
R, S not compact

- Mathematical formulation: There exists $s \neq 0$ and r such that

$$s^T x < r \qquad \text{for all } x \in R$$

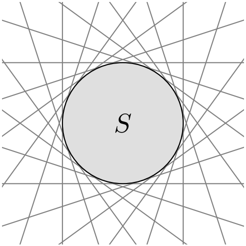
$$s^T x > r \qquad \text{for all } x \in S$$

Consequence – S is intersection of halfspaces

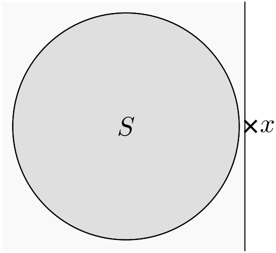
$\boxed{\text{a closed convex set } S \text{ is the intersection of all halfspaces that contain it}}$

proof:

- let H be the intersection of all halfspaces containing S
- $\Rightarrow$: obviously $x \in S \Rightarrow x \in H$
- $\Leftarrow$: assume $x \notin S$, since S closed and convex and x compact (a point), there exists a strictly separating hyperplane, i.e., $x \notin H$:



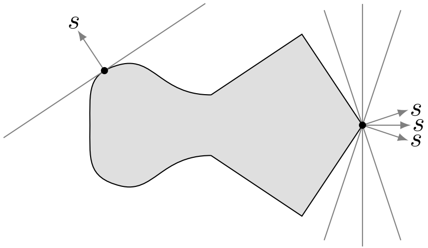
$\Rightarrow$



$\Leftarrow$

Supporting hyperplanes

- Supporting hyperplanes touch set and have full set on one side:



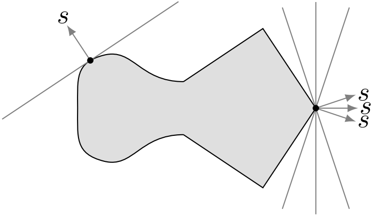
- We call the halfspace that contains the set *supporting halfspace*
- s is called *normal vector* to S at x
- Definition: Hyperplane $\{y : s^T y = r\}$ supports S at $x \in$ bd S if

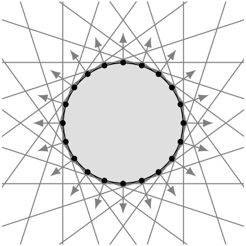
$$s^T y \leq r \text{ for all } y \in S \qquad \text{and} \qquad s^T x = r$$

Supporting hyperplane theorem

Let S be a nonempty convex set and let $x \in \mathrm{bd}(S)$. Then there exists a supporting hyperplane to S at x.

- Does not exist for all point on boundary for nonconvex sets
- Many supporting hyperplanes exist for points of nonsmoothness





Connection to duality and subgradients

Supporting hyperplanes are at the core of convex analysis:

- Subgradients define supporting hyperplanes to $\mathrm{epi} f$
- Conjugate functions define supporting hyperplanes to $\mathrm{epi} f$
- Duality is based on subgradients, hence supporting hyperplanes:
  - Consider $\mathrm{minimize}_x(f(x) + g(x))$ and primal solution $x^\star$
  - Dual problem $\mathrm{minimize}_\mu(f^*(\mu) + g^*(-\mu))$ solution $\mu^\star$ satisfies

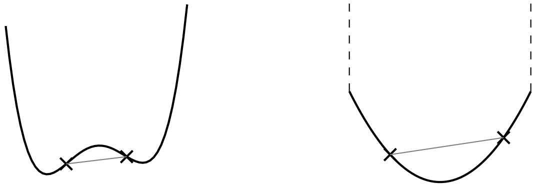
$$\mu^\star \in \partial f(x^\star) \qquad\qquad -\mu^\star \in \partial g(x^\star)$$

  i..e, dual problem finds subgradients at optimal point[1]

[1] When solving $\min_x(f(Lx) + g(x))$ dual problem finds μ such that $L^T\mu \in \partial(f \circ L)(x)$ and $-L^T\mu \in \partial g(x)$.

Convex functions

- Graph below line connecting any two pairs $(x, f(x))$ and $(y, f(y))$



nonconvex function convex function

- Function $f\ :\ \mathbb{R}^n \to \overline{\mathbb{R}}$ is *convex* if for all $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$:

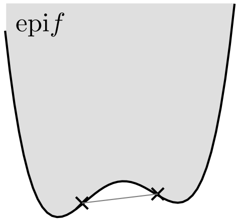
$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

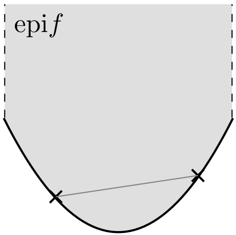
(in extended valued arithmetics)

- A function f is *concave* if $-f$ is convex

Epigraphs and convexity

- Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$
- Then f is convex if and only $\mathrm{epi} f$ is a convex set in $\mathbb{R}^n \times \mathbb{R}$





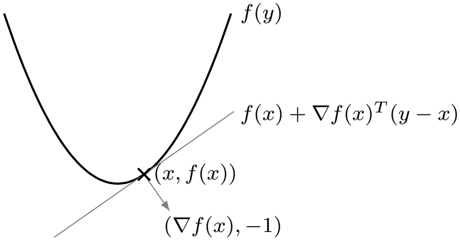
- f is called closed (lower semi-continuous) if $\mathrm{epi} f$ is closed set

First-order condition for convexity

- A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

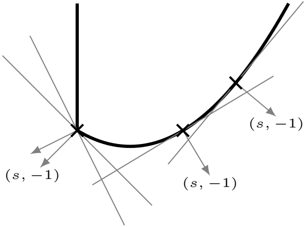
for all $x, y \in \mathbb{R}^n$



- Function f has for all $x \in \mathbb{R}^n$ an affine minorizer that:
  - has slope s defined by ∇f
  - coincides with function f at x
  - is supporting hyperplane to epigraph of f
  - defines normal $(\nabla f(x), -1)$ to epigraph of f

Subdifferentials and subgradients

- Subgradients s define affine minorizers to the function that:



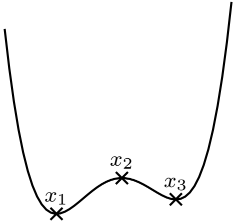
  - coincide with f at x
  - define normal vector $(s, -1)$ to epigraph of f
  - can be one of many affine minorizers at nondifferentiable points x
- Subdifferential of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at x is set of vectors s satisfying

$$f(y) \geq f(x) + s^T(y - x) \quad \text{for all } y \in \mathbb{R}^n, \tag{1}$$

- Notation:
  - subdifferential: $\partial f : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ (power-set notation $2^{\mathbb{R}^n}$)
  - subdifferential at x: $\partial f(x) = \{s : (1) \text{ holds}\}$
  - elements $s \in \partial f(x)$ are called *subgradients* of f at x

Subgradient existence – Nonconvex example

- Function can be differentiable at x but $\partial f(x) = \emptyset$

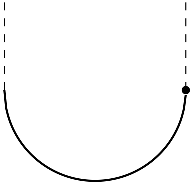


  - x_1: $\partial f(x_1) = \{0\}$, $\nabla f(x_1) = 0$
  - x_2: $\partial f(x_2) = \emptyset$, $\nabla f(x_2) = 0$
  - x_3: $\partial f(x_3) = \emptyset$, $\nabla f(x_3) = 0$

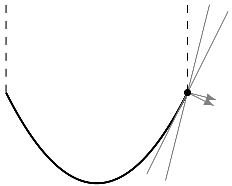
- Gradient is a local concept, subdifferential is a global property

Existence for extended-valued convex functions

- Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex, then:
  1. Subgradients exist for all x in relative interior of $\mathrm{dom} f$
  2. Subgradients sometimes exist for x on boundary of $\mathrm{dom} f$
  3. No subgradient exists for x outside $\mathrm{dom} f$
- Examples for second case, boundary points of $\mathrm{dom} f$:



$-\sqrt{1-x^2} + \iota_{[-1,1]}(x)$



$x^2 + \iota_{[-2,2]}(x)$

- No subgradient (affine minorizer) exists for left function at $x = 1$

Fermat's rule

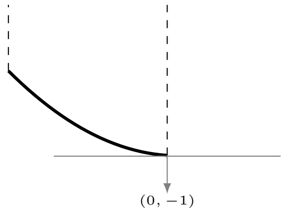
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then x minimizes f if and only if

$$0 \in \partial f(x)$$

- Proof: x minimizes f if and only if

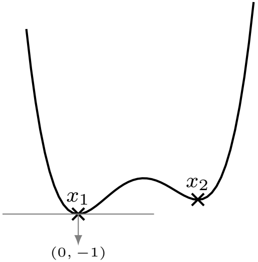
$$f(y) \geq f(x) + 0^T(y - x) \quad \text{for all } y \in \mathbb{R}^n$$

  which by definition of subdifferential is equivalent to $0 \in \partial f(x)$
- Example: several subgradients at solution, including 0



Fermat's rule – Nonconvex example

- Fermat's rule holds also for nonconvex functions
- Example:



  - $\partial f(x_1) = 0$ and $\nabla f(x_1) = 0$ (global minimum)
  - $\partial f(x_2) = \emptyset$ and $\nabla f(x_2) = 0$ (local minimum)
- For nonconvex f, we can typically only hope to find local minima

Subdifferential calculus rules

- Subdifferential of sum $\partial(f_1 + f_2)$
- Subdifferential of composition with matrix $\partial(g \circ L)$

Subdifferential of sum

If f_1, f_2 closed convex and $\mathrm{relint}\,\mathrm{dom} f_1 \cap \mathrm{relint}\,\mathrm{dom} f_2 \neq \emptyset$:

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2$$

- One direction always holds: if $x \in \mathrm{dom}\partial f_1 \cap \mathrm{dom}\partial f_2$:

  $$\partial(f_1 + f_2)(x) \supseteq \partial f_1(x) + \partial f_2(x)$$

  Proof: let $s_i \in \partial f_i(x)$, add subdifferential definitions:

  $$f_1(y) + f_2(y) \geq f_1(x) + f_2(x) + (s_1 + s_2)^T(y - x)$$

  i.e. $s_1 + s_2 \in \partial(f_1 + f_2)(x)$

- If f_1 and f_2 differentiable, we have (without convexity of f)

  $$\nabla(f_1 + f_2) = \nabla f_1 + \nabla f_2$$

Subdifferential of composition

$$\boxed{\begin{array}{c}\text{If } f \text{ closed convex and } \operatorname{relint}\operatorname{dom}(f \circ L) \neq \emptyset\text{:}\\ \partial(f \circ L)(x) = L^T \partial f(Lx)\end{array}}$$

- One direction always holds: If $Lx \in \mathrm{dom} f$, then

$$\partial(f \circ L)(x) \supseteq L^T \partial f(Lx)$$

Proof: let $s \in \partial f(Lx)$, then by definition of subgradient of f:

$$(f \circ L)(y) \geq (f \circ L)(x) + s^T(Ly - Lx) = (f \circ L)(x) + (L^T s)^T(y - x)$$

i.e., $L^T s \in \partial(f \circ L)(x)$

- If f differentiable, we have chain rule (without convexity of f)

$$\nabla(f \circ L)(x) = L^T \nabla f(Lx)$$

A sufficient optimality condition

Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, and $L \in \mathbb{R}^{m \times n}$ then:

$$\operatorname{minimize} f(Lx) + g(x) \tag{1}$$

is solved by every $x \in \mathbb{R}^n$ that satisfies

$$0 \in L^T \partial f(Lx) + \partial g(x) \tag{2}$$

- Subdifferential calculus inclusions say:

$$0 \in L^T \partial f(Lx) + \partial g(x) \subseteq \partial((f \circ L)(x) + g(x))$$

  which by Fermat's rule is equivalent to x solution to (1)
- Note: (1) can have solution but no x exists that satisfies (2)

A necessary and sufficient optimality condition

Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $L \in \mathbb{R}^{m\times n}$ with f, g closed convex and assume $\operatorname{relint} \operatorname{dom}(f \circ L) \cap \operatorname{relint} \operatorname{dom} g \neq \emptyset$ then:

$$\operatorname{minimize} f(Lx) + g(x) \tag{1}$$

is solved by $x \in \mathbb{R}^n$ if and only if x satisfies

$$0 \in L^T \partial f(Lx) + \partial g(x) \tag{2}$$

- Subdifferential calculus equality rules say:

$$0 \in L^T \partial f(Lx) + \partial g(x) = \partial((f \circ L)(x) + g(x))$$

  which by Fermat's rule is equivalent to x solution to (1)
- Algorithms search for x that satisfy $0 \in L^T \partial f(Lx) + \partial g(x)$

Evaluating subgradients of convex functions

- Obviously need to evaluate subdifferentials to solve

$$0 \in L^T \partial f(Lx) + \partial g(x)$$

- Explicit evaluation:
  - If function is differentiable: ∇f (unique)
  - If function is nondifferentiable: compute element in ∂f
- Implicit evaluation:
  - Proximal operator (specific element of subdifferential)

Proximal operator

- Proximal operator of (convex) g defined as:

$$\mathrm{prox}_{\gamma g}(z) = \underset{x}{\mathrm{argmin}}(g(x) + \tfrac{1}{2\gamma}\|x - z\|_2^2)$$

  where $\gamma > 0$ is a parameter
- Evaluating prox requires solving optimization problem
- Objective is strongly convex $\Rightarrow$ solution exists and is unique

Prox evaluates the subdifferential

- Fermat's rule on prox definition: $x = \mathrm{prox}_{\gamma g}(z)$ if and only if

$$0 \in \partial g(x) + \gamma^{-1}(x - z) \quad \Leftrightarrow \quad \gamma^{-1}(z - x) \in \partial g(x)$$

  Hence, $\gamma^{-1}(z - x)$ is element in $\partial g(x)$
- A subgradient in $\partial g(x)$ where $x = \mathrm{prox}_{\gamma g}(z)$ is computed
- Often used in algorithms when g nonsmooth (no gradient exists)

Conjugate functions

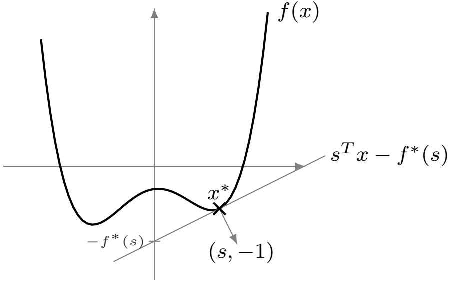
- The conjugate function of $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(s) := \sup_x \left(s^T x - f(x) \right)$$

- Implicit definition via optimization problem

Conjugate interpretation

- Conjugate $f^*(s)$ defines affine minorizer to f with slope s:



where $f^*(s)$ decides the constant offset to have support at x^*

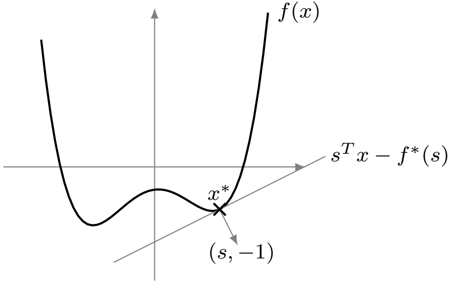
- "Affine minorizor generator: Pick slope s, get offset for support"
- Why? Consider $f^*(s) = \sup_x \left(s^Tx - f(x)\right)$ with maximizer x^*:

$$f^*(s) = s^Tx^* - f(x^*) \quad \Leftrightarrow \quad f^*(s) \geq s^Tx - f(x) \text{ for all } x$$
$$\Leftrightarrow \quad f(x) \geq s^Tx - f^*(s) \text{ for all } x$$

- Support at x^* since $f(x^*) = s^Tx^* - f^*(s)$

Fenchel Young's equality

- Going back to conjugate interpretation:



- Fenchel's inequality: $f(x) \geq s^T x - f^*(s)$ for all x, s
- Fenchel-Young's equality and equivalence:

$$\boxed{f(x^*) = s^T x^* - f^*(s) \text{ holds if and only if } s \in \partial f(x^*)}$$

A subdifferential formula

$$\boxed{\text{Assume } f \text{ closed convex, then } \partial f(x) = \mathrm{Argmax}_s(s^T x - f^*(s))}$$

- Since $f^{**} = f$, we have $f(x) = \sup_s(x^T s - f^*(s))$ and

$$\begin{aligned} s^* \in \underset{s}{\mathrm{Argmax}}(x^T s - f^*(s)) \quad &\iff \quad f(x) = x^T s^* - f^*(s^*) \\ &\iff \quad s^* \in \partial f(x) \end{aligned}$$

- The last equivalence is Fenchel-Young

Subdifferential of conjugate – Inversion formula

$$\boxed{\text{Suppose } f \text{ closed convex, then } s \in \partial f(x) \iff x \in \partial f^*(s)}$$

- Consequence of Fenchel-Young
- Another way to write the result is that for closed convex f:

$$\partial f^* = (\partial f)^{-1}$$

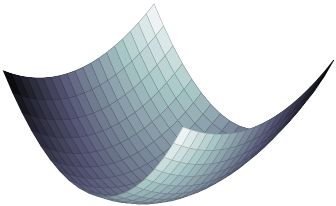
(Definition of inverse of set-valued A: $x \in A^{-1}u \iff u \in Ax$)

Strong convexity

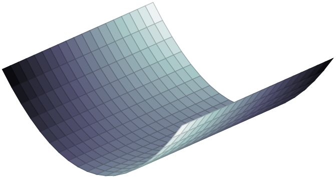
- Let $\sigma > 0$
- A function f is *σ-strongly convex* if $f - \frac{\sigma}{2}\|\cdot\|_2^2$ is convex
- Alternative equivalent definition of σ-strong convexity:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{\sigma}{2}\theta(1-\theta)\|x-y\|^2$$

  holds for every $x, y \in \mathbb{R}^n$ and $\theta \in [0,1]$
- Strongly convex functions are strictly convex and convex
- Example: f 2-strongly convex since $f - \|\cdot\|_2^2$ convex:



$f(x)$



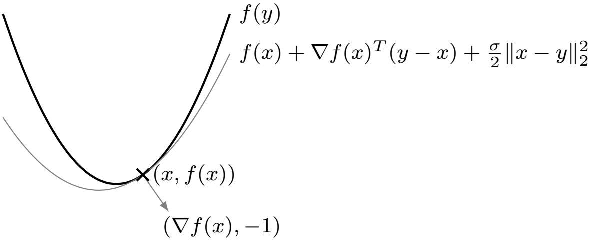
$f(x) - \|x\|_2^2$

First-order condition for strong convexity

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable
- f is σ-strongly convex with $\sigma > 0$ if and only if

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\|x-y\|_2^2$$

for all $x, y \in \mathbb{R}^n$



- Function f has for all $x \in \mathbb{R}^n$ a quadratic minorizer that:
  - has curvature defined by σ
  - coincides with function f at x
  - defines normal $(\nabla f(x), -1)$ to epigraph of f

Smoothness

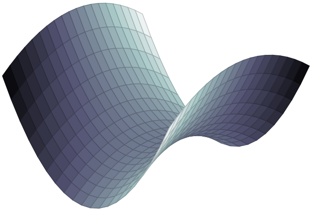
- A function is called β-*smooth* if its gradient is β-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$$

  for all $x, y \in \mathbb{R}^n$ (it is not necessarily convex)
- Alternative equivalent definition of β-smoothness

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y) - \frac{\beta}{2}\theta(1-\theta)\|x-y\|^2$$
$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) + \frac{\beta}{2}\theta(1-\theta)\|x-y\|^2$$

  hold for every $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$
- Smoothness does not imply convexity
- Example:

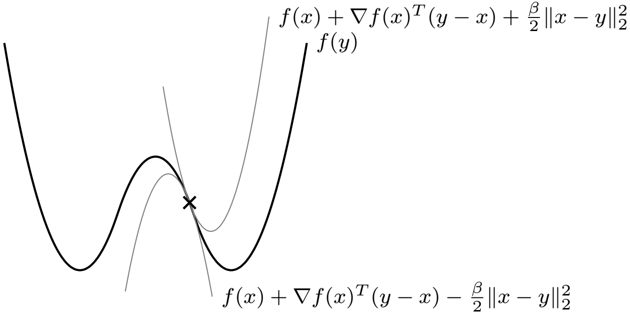


First-order condition for smoothness

- f is β-smooth with $\beta \geq 0$ if and only if

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|x-y\|_2^2$$
$$f(y) \geq f(x) + \nabla f(x)^T(y-x) - \frac{\beta}{2}\|x-y\|_2^2$$

for all $x, y \in \mathbb{R}^n$



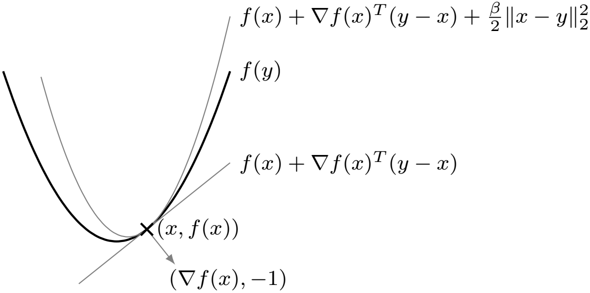
- Quadratic upper/lower bounds with curvatures defined by β
- Quadratic bounds coincide with function f at x

First-order condition for smooth convex

- f is β-smooth with $\beta \geq 0$ and convex if and only if

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|x-y\|_2^2$$
$$f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

for all $x, y \in \mathbb{R}^n$



- Quadratic upper bound and affine lower bound
- Bounds coincide with function f at x
- Quadratic upper bound is called *descent lemma*

Duality correspondance

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Then the following are equivalent:

(i) f is closed and σ-strongly convex

(ii) ∂f is maximally monotone and σ-strongly monotone

(iii) ∇f^* is σ-cocoercive

(iv) ∇f^* is maximally monotone and $\frac{1}{\sigma}$-Lipschitz continuous

(v) f^* is closed convex and satisfies descent lemma (is $\frac{1}{\sigma}$-smooth)

where $\nabla f^* : \mathbb{R}^n \to \mathbb{R}^n$ and $f^* : \mathbb{R}^n \to \mathbb{R}$

Comments:

- Relation (i) $\Leftrightarrow$ (v) most important for us
- Since $f = f^{**}$ the result holds with f and f^* interchanged
- Full proof available on course webpage

Composite Optimization

Composite optimization

We consider composite optimization problems of the form

$$\underset{x}{\text{minimize}}\, f(Lx) + g(x)$$

Optimality conditions and dual problem

- Assume f, g closed convex and that CQ holds
- Problem $\mathrm{minimize}_x(f(Lx) + g(x))$ is solved by x iff

$$0 \in L^T \underbrace{\partial f(Lx)}_{\mu} + \partial g(x)$$

  where dual variable μ has been defined
- Primal dual necessary and sufficient optimality conditions:

$$\begin{cases} \mu \in \partial f(Lx) \\ -L^T\mu \in \partial g(x) \end{cases} \qquad \begin{cases} Lx \in \partial f^*(\mu) \\ -L^*\mu \in \partial g(x) \end{cases}$$

$$\begin{cases} \mu \in \partial f(Lx) \\ x \in \partial g^*(-L^T\mu) \end{cases} \qquad \begin{cases} Lx \in \partial f^*(\mu) \\ x \in \partial g^*(-L^T\mu) \end{cases}$$

- Dual optimality condition

$$0 \in \partial f^*(\mu) + \partial(g^* \circ -L^T)(\mu) \tag{1}$$

  solves dual problem $\mathrm{minimize}_\mu\, f^*(\mu) + g^*(-L^T\mu)$
- If CQ-D holds, all dual problem solutions satisfy (1)
- Dual searches for μ such that $L^T\mu \in \partial f(x)$ and $-L^T\mu \in \partial g(x)$

Solving the primal via the dual

- Why solve dual? Sometimes easier to solve than primal
- Only interesting if primal solution can be recovered
- Assume f, g closed convex and CQ
- Assume optimal dual μ known: $0 \in \partial f^*(\mu) + \partial(g^* \circ -L^T)(\mu)$
- Optimal primal x must satisfy any and all primal-dual conditions:

$$\begin{cases} \mu \in \partial f(Lx) \\ -L^T\mu \in \partial g(x) \end{cases} \qquad \begin{cases} Lx \in \partial f^*(\mu) \\ -L^T\mu \in \partial g(x) \end{cases}$$

$$\begin{cases} \mu \in \partial f(Lx) \\ x \in \partial g^*(-L^T\mu) \end{cases} \qquad \begin{cases} Lx \in \partial f^*(\mu) \\ x \in \partial g^*(-L^T\mu) \end{cases}$$

- If one of these uniquely characterizes x, then must be solution:
  - ∂g^* is differentiable at $-L^T\mu$ for dual solution μ
  - ∂f^* is differentiable at dual solution μ and L invertible
  - $\ldots$

Algorithms

Proximal gradient method

- Consider $\underset{x}{\mathrm{minimize}}\, f(x) + g(x)$ where
  - f is β-smooth $f : \mathbb{R}^n \to \mathbb{R}$ (not necessarily convex)
  - g is closed convex
- Due to β-smoothness of f, we have

$$f(y) + g(y) \leq f(x) + \nabla f(x)^T(y - x) + \tfrac{\beta}{2}\|y - x\|_2^2 + g(y)$$

  for all $x, y \in \mathbb{R}^n$, i.e., r.h.s. is majorizing function for fixed x
- Majorization minimization with majorizer if $\gamma_k \in [\epsilon, \beta^{-1}]$, $\epsilon > 0$:

$$\begin{aligned} x_{k+1} &= \underset{y}{\mathrm{argmin}}\left(f(x_k) + \nabla f(x_k)^T(y - x) + \tfrac{1}{2\gamma_k}\|y - x_k\|_2^2 + g(y)\right) \\ &= \underset{y}{\mathrm{argmin}}\left(g(y) + \tfrac{1}{2\gamma_k}\|y - (x_k - \gamma_k \nabla f(x_k))\|_2^2\right) \\ &= \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) \end{aligned}$$

  gives proximal gradient method

Proximal gradient – Fixed-points

- Denote $T_{\rm PG}^{\gamma} := \mathrm{prox}_{\gamma g}(I - \gamma\nabla f)$, gives algorithm $x_{k+1} = T_{\rm PG}^{\gamma} x_k$
- Proximal gradient fixed-point set definition

$$\mathrm{fix} T_{\rm PG}^{\gamma} = \{x : x = T_{\rm PG}^{\gamma} x\} = \{x : x = \mathrm{prox}_{\gamma g}(x - \gamma\nabla f(x))\}$$

i.e., set of points for which $x_{k+1} = x_k$

Let $\gamma > 0$. Then $\bar{x} \in \mathrm{fix} T_{\rm PG}^{\gamma}$ if and only if $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$.

- Consequence: fixed-point set same for all $\gamma > 0$
- We call inclusion $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$ *fixed-point characterization*
    - For convex problems: global solutions
    - For nonconvex problems: critical points

Applying proximal gradient to primal problems

Problem $\underset{x}{\text{minimize}}\, f(x) + g(x)$:

- Assumptions:
  - f β-smooth
  - g closed convex and prox friendly[1]
  - $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$
- Algorithm: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$

Problem $\underset{x}{\text{minimize}}\, f(Lx) + g(x)$:

- Assumptions:
  - f β-smooth (implies $f \circ L$ $\beta\|L\|_2^2$-smooth)
  - g closed convex and prox friendly[1]
  - $\gamma_k \in [\epsilon, \frac{2}{\beta\|L\|_2^2} - \epsilon]$
- Gradient $\nabla(f \circ L)(x) = L^T \nabla f(Lx)$
- Algorithm: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k L^T \nabla f(Lx_k))$

---

[1]Prox friendly: proximal operator cheap to evaluate, e.g., g separable

Applying proximal gradient to dual problem

Dual problem $\underset{\nu}{\text{minimize}}\, f^*(\nu) + g^*(-L^T\nu)$:

- Assumptions:
  - f closed convex and prox friendly
  - g σ-strongly convex (which implies $g^* \circ -L^T$ $\frac{\|L\|_2^2}{\sigma}$-smooth)
  - $\gamma_k \in [\epsilon, \frac{2\sigma}{\|L\|_2^2} - \epsilon]$
- Gradient: $\nabla(g^* \circ -L^T)(\nu) = -L\nabla g^*(-L^T\nu)$
- Prox (Moreau): $\text{prox}_{\gamma_k f^*}(\nu) = \nu - \gamma_k \text{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1}\nu)$
- Algorithm:

$$\begin{aligned}
\nu_{k+1} &= \text{prox}_{\gamma_k f^*}(\nu_k - \gamma_k \nabla(g^* \circ -L^T)(\nu_k)) \\
&= (I - \gamma_k \text{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} \circ I))(\nu_k + \gamma_k L \nabla g^*(-L^T\nu_k))
\end{aligned}$$

- Problem must be convex to have dual!
- Enough to know prox of f

What problems cannot be solved (efficiently)?

Problem $\underset{x}{\operatorname{minimize}}\, f(x) + g(x)$

- Assumptions: f and g convex and nonsmooth
- No term differentiable, another method must be used:
  - Subgradient method
  - Douglas-Rachford splitting
  - Primal-dual methods

Problem $\underset{x}{\operatorname{minimize}}\, f(x) + g(Lx)$

- Assumptions:
  - f smooth
  - g nonsmooth convex
  - L arbitrary structured matrix
- Can apply proximal gradient method, but

$$\operatorname{prox}_{\gamma_k(g \circ L)}(z) = \underset{x}{\operatorname{argmin}}\, g(Lx) + \tfrac{1}{2\gamma}\|x - z\|_2^2)$$

often not "prox friendly", i.e., it is expensive to evaluate

Training problems

- Training problem format

$$\underset{\theta}{\text{minimize}} \underbrace{\sum_{i=1}^{N} L(m(x_i;\theta), y_i)}_{f(X\theta)} + \underbrace{\sum_{j=1}^{n} g_j(\theta_j)}_{g(\theta)}$$

  where f is data misfit term and g is regularizer
- Regularizers ($\theta = (w, b)$)
  - Tikhonov $g(\theta) = \|w\|_2^2$ is prox-friendly
  - Sparsity inducing 1-norm $g(\theta) = \|w\|_1$ is prox-friendly
- Data misfit terms (with $m(x;\theta) = \phi(x)^T\theta$ for convex problems)
  - Least squares $L(u, y) = \|u - y\|_2^2$ smooth, hence f smooth
  - Logistic $L(u, y) = \log(1 + e^u) - yu$ smooth, hence f smooth
  - SVM $L(u, y) = \max(0, 1 - yu)$ not smooth, hence f not smooth
- Proximal gradient method
  - Least squares: can efficiently solve primal
  - Logistic regression: can solve primal
  - SVM: add strongly convex regularization and solve dual
    - Strongly convex regulariztion to have one conjugate smooth
    - If bias term not regularized, only strongly convex in w
    - SVM with $\|\cdot\|_1$-regularization not solvable with prox-grad

Dual training problem

- Convex training problem

$$\underset{\theta}{\text{minimize}} \underbrace{\sum_{i=1}^{N} L(\phi(x_i)^T\theta, y_i)}_{f(X\theta)} + \underbrace{\sum_{j=1}^{n} g_j(\theta_j)}_{g(\theta)}$$

has dual

$$\underset{\theta}{\text{minimize}} \underbrace{\sum_{i=1}^{N} L^*(\mu_i)}_{f^*(\mu)} + \underbrace{\sum_{j=1}^{n} g_j^*((-X^T\mu)_j)}_{g^*(-X^T\mu)}$$

where the conjugate of L is w.r.t. first argument

- Dual has same structure as primal, finite-sum plus separable

Training problem structure

- Primal training problem

$$\underset{\theta}{\text{minimize}} \underbrace{\sum_{i=1}^{N} L(m(x_i;\theta), y_i)}_{f(X\theta)} + \underbrace{\sum_{j=1}^{n} g_j(\theta_j)}_{g(\theta)}$$

- Dual training problem

$$\underset{\theta}{\text{minimize}} \underbrace{\sum_{i=1}^{N} L^*(\mu_i)}_{f^*(\mu)} + \underbrace{\sum_{j=1}^{n} g_j^*((-X^T\mu)_j)}_{g^*(-X^T\mu)}$$

- Common structure, finite sum plus separable:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{N} f_i((X\theta)_i) + \sum_{j=1}^{n} \psi_j(\theta_j)$$

  - Primal: $f_i = L(m(x_i;\cdot), y_i)$ (one summand per training example)
  - Dual: $f_i = g_j^*((-X^T\cdot)_j)$, $\psi_j = L^*$

Exploiting structure

- Common structure, finite sum plus separable:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{N} f_i((X\theta)_i) + \sum_{j=1}^{n} \psi_j(\theta_j)$$

- Stochastic gradient descent exploits finite-sum structure:
  - Computes stochastic gradient of *smooth* part f
  - Pick summand f_i at random and perform gradient step
  - Primal formulations: Pick training example and compute gradient
  - Deep learning: evaluted via backpropagation
- Coordinate gradient descent exploits separable structure:
  - Coordinate-wise updates if *nonsmooth* ϕ_j separable
  - Requires efficient coordinate-wise evaluations of ∇f