Subdifferentials and Proximal Operators

Pontus Giselsson

Outline

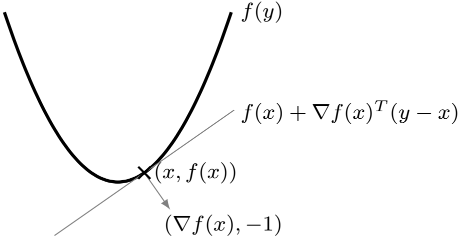
- **Subdifferential and subgradient – Definition and basic properties**
- Monotonicity
- Examples
- Strong monotonicity and cocoercivity
- Fermat's rule
- Subdifferential calculus
- Optimality conditions
- Proximal operators

Gradients of convex functions

- Recall: A *differentiable* function $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff

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

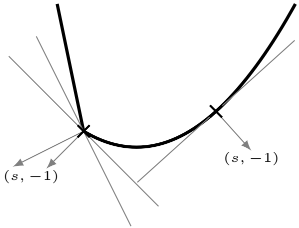
for all $x, y \in \mathbb{R}^n$



- 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
  - defines normal $(\nabla f(x), -1)$ to epigraph of f
- What if function is nondifferentiable?

Subdifferentials and subgradients

- Subgradients s define affine minorizers to the function that:

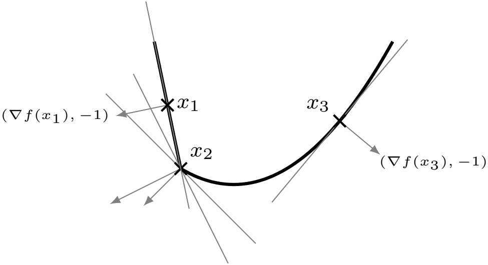


  - coincide with f at x
  - define normal vector $(s,-1)$ to epigraph of f
  - can be one of many affine minorizers at nondifferentiable points x
- Subdifferential of $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ at x is set of vectors s satisfying

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

- 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

Relation to gradient



- If f differentiable at x and $\partial f(x) \neq \emptyset$ then $\partial f(x) = \{\nabla f(x)\}$
- If f convex and $\partial f(x)$ a singleton then $\partial f(x) = \{\nabla f(x)\}$
- If f convex but not differentiable at $x \in \operatorname{int} \operatorname{dom} f$, then

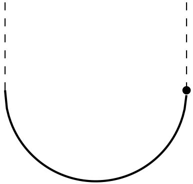
$$\partial f(x) = \operatorname{cl}(\operatorname{conv} S(x))$$

  where $S(x)$ is set of all s such that $\nabla f(x_k) \to s$ when $x_k \to x$
- In general for convex f: $\partial f(x) = \operatorname{cl}(\operatorname{conv} S(x)) + N_{\operatorname{dom} f}(x)$

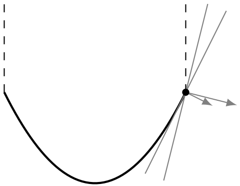
Subgradient existence – Convex setting

For *finite-valued convex* functions, a subgradient exists for every x

- In extended-valued setting, let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex:
  (i) Subgradients exist for all x in relative interior of $\mathrm{dom} f$
  (ii) Subgradients sometimes exist for x on relative boundary of $\mathrm{dom} f$
  (iii) No subgradient exists for x outside $\mathrm{dom} f$
- Examples for second case, boundary points of $\mathrm{dom} f$:



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

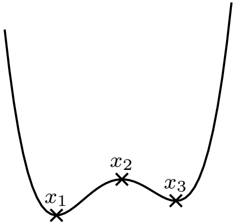


$x^2 + \iota_{[-2,2]}(x)$

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

Subgradient existence – Nonconvex setting

- 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

Outline

- Subdifferential and subgradient – Definition and basic properties
- **Monotonicity**
- Examples
- Strong monotonicity and cocoercivity
- Fermat's rule
- Subdifferential calculus
- Optimality conditions
- Proximal operators

Monotonicity of subdifferential

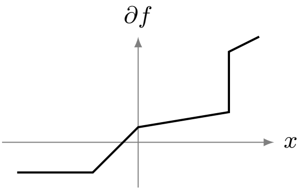
- Subdifferential operator is *monotone*:

$$(s_x - s_y)^T(x - y) \geq 0$$

  for all $s_x \in \partial f(x)$ and $s_y \in \partial f(y)$
- Proof: Add two copies of subdifferential definition

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

  with x and y swapped
- $\partial f : \mathbb{R} \to 2^{\mathbb{R}}$: Minimum slope 0 and maximum slope ∞



Monotonicity beyond subdifferentials

- Let $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be monotone, i.e.:

$$(u - v)^T (x - y) \geq 0$$

  for all $u \in Ax$ and $v \in Ay$

- There exist monotone A that are not subdifferentials

Maximal monotonicity

- Let the set $\operatorname{gph} \partial f := \{(x, u) : u \in \partial f(x)\}$ be the graph of ∂f
- ∂f is maximally monotone if no other function g exists with

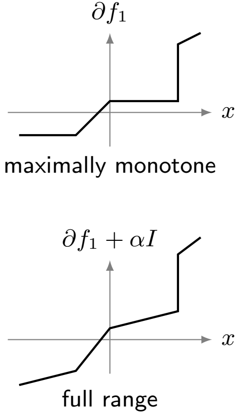
$$\operatorname{gph} \partial f \subset \operatorname{gph} \partial g,$$

  with strict inclusion
- A result (due to Rockafellar):

$$\boxed{f \text{ is closed convex if and only if } \partial f \text{ is maximally monotone}}$$

Minty's theorem

- Let $\partial f : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ and $\alpha > 0$
- ∂f is maximally monotone if and only if $\mathrm{range}(\alpha I + \partial f) = \mathbb{R}^n$



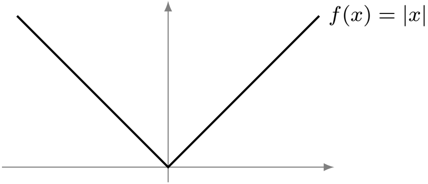
- Interpretation: No "holes" in $\mathrm{gph}\,\partial f$

Outline

- Subdifferential and subgradient – Definition and basic properties
- Monotonicity
- **Examples**
- Strong monotonicity and cocoercivity
- Fermat's rule
- Subdifferential calculus
- Optimality conditions
- Proximal operators

Example – Absolute value

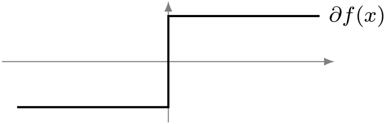
- The absolute value:



- Subdifferential
  - For $x > 0$, f differentiable and $\nabla f(x) = 1$, so $\partial f(x) = \{1\}$
  - For $x < 0$, f differentiable and $\nabla f(x) = -1$, so $\partial f(x) = \{-1\}$
  - For $x = 0$, f not differentiable, but since f convex:

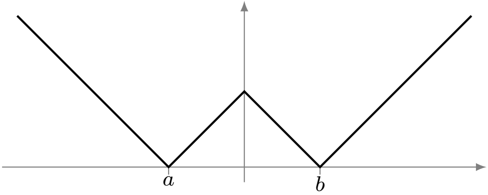
$$\partial f(0) = \mathrm{cl}(\mathrm{conv} S(0)) = \mathrm{cl}(\mathrm{conv}(\{-1, 1\}) = [-1, 1]$$

- The subdifferential operator:

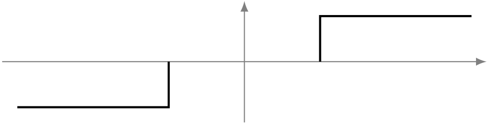


A nonconvex example

- Nonconvex function:



- Subdifferential
  - For $x > b$, f differentiable and $\nabla f(x) = 1$, so $\partial f(x) = \{1\}$
  - For $x < a$, f differentiable and $\nabla f(x) = -1$, so $\partial f(x) = \{-1\}$
  - For $x \in (a, b)$, no affine minorizer, $\partial f(x) = \emptyset$
  - For $x = a$, f not differentiable, $\partial f(x) = [-1, 0]$
  - For $x = b$, f not differentiable, $\partial f(x) = [0, 1]$
- The subdifferential operator:



Example – Separable functions

- Consider the separable function $f(x) = \sum_{i=1}^{n} f_i(x_i)$
- Subdifferential

$$\partial f(x) = \{s = (s_1, \ldots, s_n) : s_i \in \partial f_i(x_i)\}$$

- The subgradient $s \in \partial f(x)$ if and only if each $s_i \in \partial f_i(x_i)$
- Proof:
  - Assume all $s_i \in \partial f_i(x_i)$:

$$f(y) - f(x) = \sum_{i=1}^{n} f_i(y_i) - f_i(x_i) \geq \sum_{i=1}^{n} s_i(y_i - x_i) = s^T(y - x)$$

  - Assume $s_j \notin \partial f_j(x_j)$ and $x_i = y_i$ for all $i \neq j$:

$$f_j(y_j) - f_j(x_j) < s_j(y_j - x_j)$$

which gives

$$f(y) - f(x) = f_j(y_j) - f_j(x_j) < s_j(y_j - x_j) = s^T(y - x)$$

Example – 1-norm

- Consider the 1-norm $f(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$
- It is a separable function of absolute values
- From previous examples, we conclude that the subdifferential is

$$\partial f(x) = \left\{ (s_1, \ldots, s_n) : \begin{cases} s_i = -1 & \text{if } x_i < 0 \\ s_i \in [-1, 1] & \text{if } x_i = 0 \\ s_i = 1 & \text{if } x_i > 0 \end{cases} \right\}$$

Example – 2-norm

- Consider the 2-norm $f(x) = \|x\|_2 = \sqrt{\|x\|_2^2}$
- The function is differentiable everywhere except for when $x = 0$
- Divide into two cases; $x = 0$ and $x \neq 0$
- Subdifferential for $x \neq 0$: $\partial f(x) = \{\nabla f(x)\}$:
  - Let $h(u) = \sqrt{u}$ and $g(x) = \|x\|_2^2$, then $f(x) = (h \circ g)(x)$
  - The gradient for all $x \neq 0$ by chain rule (since $h : \mathbb{R}_+ \to \mathbb{R}$):

$$\nabla f(x) = \nabla h(g(x)) \nabla g(x) = \tfrac{1}{2\sqrt{\|x\|_2^2}} 2x = \frac{x}{\|x\|_2}$$

Example cont'd – 2-norm

Subdifferential of $\|x\|_2$ at $x = 0$

(i) educated guess of subdifferential from $\partial f(0) = \mathrm{cl}(\mathrm{conv} S(0))$

- recall $S(0)$ is set of all limit points of $(\nabla f(x_k))_{k\in\mathbb{N}}$ when $x_k \to 0$
- let $x_k = t^k d$ with $t \in (0,1)$ and $d \in \mathbb{R}^n \backslash \{0\}$, then $\nabla f(x_k) = \frac{d}{\|d\|_2}$
- since d arbitrary, $(\nabla f(x_k))$ can converge to any unit norm vector
- so $S(0) = \{s : \|s\|_2 = 1\}$ and $\partial f(0) = \{s : \|s\|_2 \leq 1\}$?

(ii) verify using subgradient definition $f(y) \geq f(0) + s^T(y - 0) = s^T y$

- Let $\|s\|_2 > 1$, then for, e.g., $y = 2s$

$$s^T y = 2\|s\|_2^2 > 2\|s\|_2 = f(y)$$

  so such s are not subgradients

- Let $\|s\|_2 \leq 1$, then for all y:

$$s^T y \leq \|s\|_2 \|y\|_2 \leq \|y\|_2 = f(y)$$

  so such s are subgradients

Outline

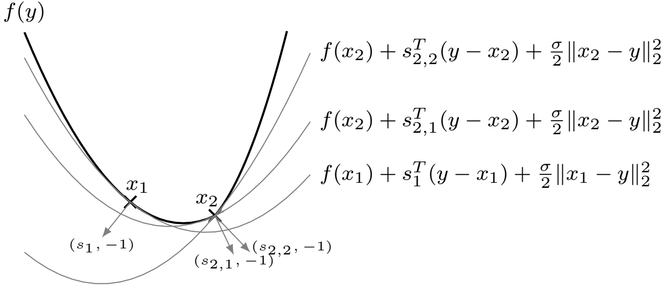
- Subdifferential and subgradient – Definition and basic properties
- Monotonicity
- Examples
- **Strong monotonicity and cocoercivity**
- Fermat's rule
- Subdifferential calculus
- Optimality conditions
- Proximal operators

Strong convexity revisited

- Recall that f is σ-strongly convex if $f - \frac{\sigma}{2}\|\cdot\|_2^2$ is convex
- If f is σ-strongly convex then

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

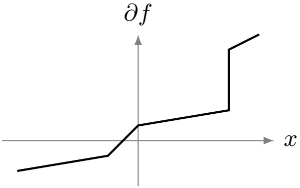
  holds for all $x \in \mathrm{dom}\partial f$, $s \in \partial f(x)$, and $y \in \mathbb{R}^n$
- The function has convex quadratic minorizers instead of affine



- Multiple lower bounds at x_2 with subgradients $s_{2,1}$ and $s_{2,2}$

Strong monotonicity

- If f σ-strongly convex function, then ∂f is *σ-strongly monotone*:
$$(s_x - s_y)^T(x - y) \geq \sigma \|x - y\|_2^2$$
for all $s_x \in \partial f(x)$ and $s_y \in \partial f(y)$
- Proof: Add two copies of strong convexity inequality
$$f(y) \geq f(x) + s_x^T(y - x) + \frac{\sigma}{2}\|x - y\|_2^2$$
with x and y swapped
- ∂f is σ-strongly monotone if and only if $\partial f - \sigma I$ is monotone
- $\partial f : \mathbb{R} \to 2^{\mathbb{R}}$: Minimum slope σ and maximum slope ∞



Strongly convex functions – An equivalence

The following are equivalent for $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$

(i) f is closed and σ-strongly convex

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

Proof:

(i)⇒(ii): we know this from before

(ii)⇒(i): (ii) $\Rightarrow \partial f - \sigma I = \partial(f - \frac{\sigma}{2}\|\cdot\|_2^2)$ maximally monotone

$\Rightarrow f - \frac{\sigma}{2}\|\cdot\|_2^2$ closed convex

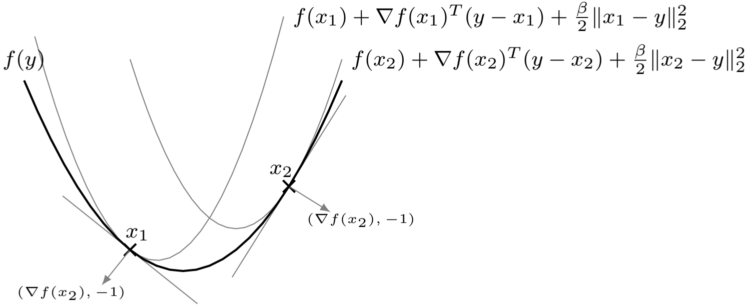
$\Rightarrow f$ closed and σ-strongly convex

Smooth convex functions

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

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

  hold for all $x, y \in \mathbb{R}^n$
- f has convex quadratic majorizers and affine minorizers



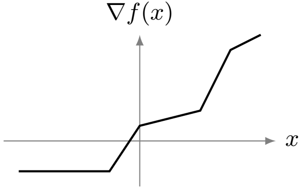
- Quadratic upper bound is called *descent lemma*

Cocoercivity of gradient

- Gradient of smooth convex function is monotone and Lipschitz

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq 0$$
$$\|\nabla f(y) - \nabla f(x)\|_2 \leq \beta \|x-y\|_2$$

- $\nabla f : \mathbb{R} \to \mathbb{R}$: Minimum slope 0 and maximum slope β



- Actually satisfies the stronger $\frac{1}{\beta}$-*cocoercivity* property:

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{\beta}\|\nabla f(y) - \nabla f(x)\|_2^2$$

due to the *Baillon-Haddad theorem*

Smooth convex functions – An equivalence

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. The following are equivalent:

(i) ∇f is $\frac{1}{\beta}$-cocoercive

(ii) ∇f is maximally monotone and β-Lipschitz continuous

(iii) f is convex and satisfies descent lemma (is β-smooth)

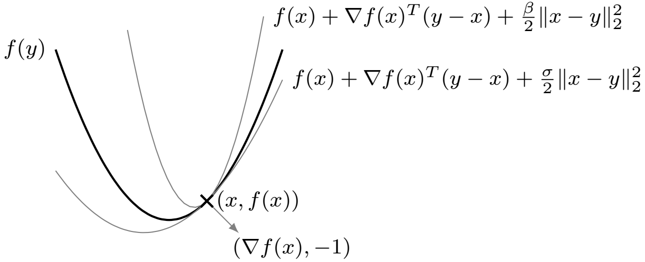
Will later connect smooth convexity and strong convexity via conjugates

Smooth strongly convex functions

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable
- f is β-smooth and σ-strongly convex with $0 < \sigma \leq \beta$ if

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

  hold for all $x, y \in \mathbb{R}^n$
- f has quadratic minorizers and quadratic majorizers



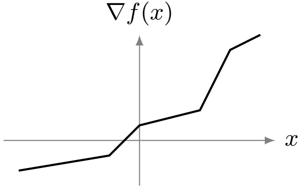
- We say that the ratio $\frac{\beta}{\sigma}$ is the *condition number* for the function

Gradient of smooth strongly convex function

- Gradient of β-smooth σ-strongly convex function f satisfies

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq \beta \|x - y\|_2$$
$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \sigma \|x - y\|_2^2$$

  so is β-Lipschitz continuous and σ-strongly monotone
- $\nabla f : \mathbb{R} \to \mathbb{R}$: Minimum slope σ and maximum slope β



- Actually satisfies this stronger property:

  $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{\beta+\sigma} \|\nabla f(y) - \nabla f(x)\|_2^2 + \frac{\sigma\beta}{\beta+\sigma} \|x - y\|_2^2$

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

Proof of stronger property

- f is σ-strongly convex if and only if $g := f - \frac{\sigma}{2}\|\cdot\|_2^2$ is convex
- Since f is β-smooth and g convex, g is $(\beta-\sigma)$-smooth
- Since g convex and $(\beta-\sigma)$-smooth, ∇g is $\frac{1}{\beta-\sigma}$-cocoercive:

$$(\nabla g(x) - \nabla g(y))^T(x-y) \geq \tfrac{1}{\beta-\sigma}\|\nabla g(x) - \nabla g(y)\|_2^2$$

which by using $\nabla g = \nabla f - \sigma I$ gives

$$(\nabla f(x) - \nabla f(y))^T(x-y) - \sigma\|x-y\|_2^2 \geq \tfrac{1}{\beta-\sigma}\|\nabla f(x) - \nabla f(y) - \sigma(x-y)\|_2^2$$

which by expanding the square and rearranging is equivalent to

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \tfrac{1}{\beta+\sigma}\|\nabla f(x) - \nabla f(y)\|_2^2 + \tfrac{\sigma\beta}{\beta+\sigma}\|x-y\|_2^2$$

Outline

- Subdifferential and subgradient – Definition and basic properties
- Monotonicity
- Examples
- Strong monotonicity and cocoercivity
- **Fermat's rule**
- Subdifferential calculus
- Optimality conditions
- Proximal operators

Fermat's rule

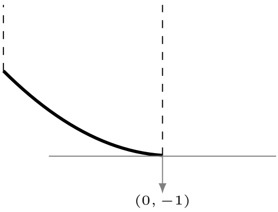
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then x minimizes f if and only if

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

- Proof: x minimizes f if and only if

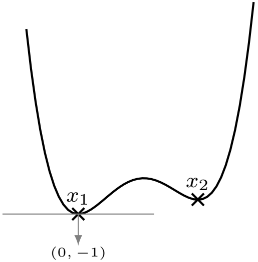
$$f(y) \geq f(x) = f(x) + 0^T(y - x) \quad \text{for all } y \in \mathbb{R}^n$$

  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

Outline

- Subdifferential and subgradient – Definition and basic properties
- Monotonicity
- Examples
- Strong monotonicity and cocoercivity
- Fermat's rule
- **Subdifferential calculus**
- Optimality conditions
- Proximal operators

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:

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

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

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

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

Outline

- Subdifferential and subgradient – Definition and basic properties
- Monotonicity
- Examples
- Strong monotonicity and cocoercivity
- Fermat's rule
- Subdifferential calculus
- **Optimality conditions**
- Proximal operators

Composite optimization problems

- We consider optimization problems on *composite form*

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

  where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, and $L \in \mathbb{R}^{m \times n}$
- Can model constrained problems via indicator function
- This model format is suitable for many algorithms

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:

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

  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:

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

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

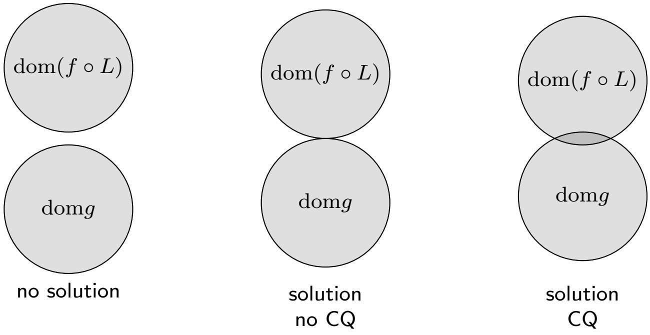
A comment on constraint qualification

- The condition

$$\operatorname{relint}\operatorname{dom}(f \circ L) \cap \operatorname{relint}\operatorname{dom} g \neq \emptyset$$

  is called *constraint qualification* and referred to as CQ

- It is a mild condition that rarely is not satisfied



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)

Outline

- Subdifferential and subgradient – Definition and basic properties
- Monotonicity
- Examples
- Strong monotonicity and cocoercivity
- Fermat's rule
- Subdifferential calculus
- Optimality conditions
- **Proximal operators**

Proximal operators

Proximal operator – Definition

- Proximal operator of $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as:

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

  where $\gamma > 0$ is a parameter
- Evaluating *prox* requires solving optimization problem
- If g closed convex, prox is single-valued mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$
  - Objective closed and strongly convex $\Rightarrow$ unique minimizing point

Prox is generalization of projection

- Recall the indicator function of a set C

$$\iota_C(x) := \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$$

- Then

$$\begin{aligned} \operatorname{prox}_{\iota_C}(z) &= \operatorname*{argmin}_x \left(\tfrac{1}{2}\|x - z\|_2^2 + \iota_C(x)\right) \\ &= \operatorname{argmin}\{\tfrac{1}{2}\|x - z\|_2^2 : x \in C\} \\ &= \operatorname{argmin}\{\|x - z\|_2 : x \in C\} \\ &= \Pi_C(z) \end{aligned}$$

- Projection onto C equals prox of indicator function of C

Prox computes a subgradient

- 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 $\partial g(x)$ where $x = \mathrm{prox}_{\gamma g}(z)$ is computed

Prox is 1-cocoercive

- For convex g, the proximal operator is 1-cocoercive:

$$(x-y)^T(\mathrm{prox}_{\gamma g}(x)-\mathrm{prox}_{\gamma f}(y)) \geq \|\mathrm{prox}_{\gamma g}(x)-\mathrm{prox}_{\gamma f}(y)\|_2^2$$

- Proof
  - Combine monotonicity of ∂g, that for all $z_u \in \partial g(u)$, $z_v \in \partial g(v)$:

$$(z_u - z_v)^T(u-v) \geq 0$$

  - with Fermat's rule on prox that evalutes subgradients of g:

$$\begin{array}{lcl} u = \mathrm{prox}_{\gamma g}(x) & \text{if and only if} & \gamma^{-1}(x-u) \in \partial g(u) \\ v = \mathrm{prox}_{\gamma g}(y) & \text{if and only if} & \gamma^{-1}(y-v) \in \partial g(v) \end{array}$$

  - which gives, by letting $z_u = \gamma^{-1}(x-u)$ and $z_v = \gamma^{-1}(y-v)$:

$$\begin{aligned} & \gamma^{-1}((x-u)-(y-v))^T(u-v) \geq 0 \\ \Leftrightarrow \quad & (x-\mathrm{prox}_{\gamma g}(x)-(y-\mathrm{prox}_{\gamma g}(y)))^T(\mathrm{prox}_{\gamma g}(x)-\mathrm{prox}_{\gamma g}(y)) \geq 0 \\ \Leftrightarrow \quad & (x-y)^T(\mathrm{prox}_{\gamma g}(x)-\mathrm{prox}_{\gamma g}(y)) \geq \|\mathrm{prox}_{\gamma g}(x)-\mathrm{prox}_{\gamma g}(y)\|_2^2 \end{aligned}$$

Prox is (firmly) nonexpansive

- We know 1-cocoercivity implies nonexpansiveness (1-Lipschitz)

$$\|\mathrm{prox}_{\gamma g}(x) - \mathrm{prox}_{\gamma g}(y)\|_2 \leq \|x - y\|_2$$

  which was shown using Cauchy-Schwarz inequality
- Actually the stronger *firm* nonexpansive inequality holds

$$\begin{aligned}\|\mathrm{prox}_{\gamma g}(x) - \mathrm{prox}_{\gamma g}(y)\|_2^2 \leq \|x - y\|_2^2 \\ - \|x - \mathrm{prox}_{\gamma g}(x) - (y - \mathrm{prox}_{\gamma g}(y))\|_2^2\end{aligned}$$

  which implies nonexpansiveness
- Proof:
  - take 1-cocoercivity and multiply both sides by 2:

$$2(x-y)^T(\mathrm{prox}_{\gamma g}(x) - \mathrm{prox}_{\gamma f}(y)) \geq 2\|\mathrm{prox}_{\gamma g}(x) - \mathrm{prox}_{\gamma f}(y)\|_2^2$$

  - use the following equality with $u = \mathrm{prox}_{\gamma g}(x)$ and $v = \mathrm{prox}_{\gamma g}(y)$:

$$(x-y)^T(u-v) = \tfrac{1}{2}\left(\|x-y\|_2^2 + \|u-v\|_2^2 - \|x-y-(u-v)\|_2^2\right)$$

Proximal operator – Separable functions

- Let $x = (x_1, \ldots, x_n)$ and $g(x) = \sum_{i=1}^{n} g_i(x_i)$ be separable, then

$$\mathrm{prox}_{\gamma g}(z) = (\mathrm{prox}_{\gamma g_1}(z_1), \ldots, \mathrm{prox}_{\gamma g_n}(z_n))$$

  decomposes into n individual proxes

- Why? Since also $\|\cdot\|_2^2$ is separable:

$$\begin{aligned}\mathrm{prox}_{\gamma g}(z) &= \underset{x\in\mathbb{R}^n}{\mathrm{argmin}}(g(x) + \tfrac{1}{2\gamma}\|x-z\|_2^2)\\ &= \underset{x\in\mathbb{R}^n}{\mathrm{argmin}}\left(\sum_{i=1}^{n} g_i(x_i) + \tfrac{1}{2\gamma}(x_i - z_i)^2\right)\end{aligned}$$

  which gives n independent optimization problems

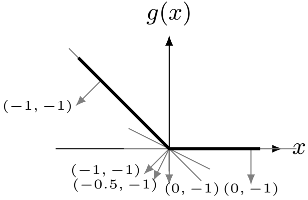
$$\underset{x_i\in\mathbb{R}}{\mathrm{argmin}}(g_i(x_i) + \tfrac{1}{2\gamma}(x_i - z_i)^2) = \mathrm{prox}_{\gamma g_i}(z_i)$$

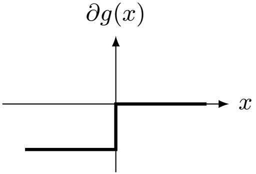
Proximal operator – Example 1

- Consider the function g with subdifferential ∂g:

$$g(x) = \begin{cases} -x & \text{if } x \le 0 \\ 0 & \text{if } x \ge 0 \end{cases} \qquad \partial g(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 0] & \text{if } x = 0 \\ 0 & \text{if } x > 0 \end{cases}$$

- Graphical representations





- Fermat's rule for $x = \mathrm{prox}_{\gamma g}(z)$:

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

Proximal operator – Example 1 cont'd

- Let $x < 0$, then Fermat's rule reads

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

  which is valid ($x < 0$) if $z < -\gamma$
- Let $x = 0$, then Fermat's rule reads

$$0 \in [-1, 0] + \gamma^{-1}(0 - z)$$

  which is valid ($x = 0$) if $z \in [-\gamma, 0]$
- Let $x > 0$, then Fermat's rule reads

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

  which is valid ($x > 0$) if $z > 0$
- The prox satisfies

$$\mathrm{prox}_{\gamma g}(z) = \begin{cases} z + \gamma & \text{if } z < -\gamma \\ 0 & \text{if } z \in [-\gamma, 0] \\ z & \text{if } z > 0 \end{cases}$$

Proximal operator – Example 2

Let $g(x) = \frac{1}{2}x^T Px + q^T x$ with P positive semidefinite

- Gradient satisfies $\nabla g(x) = Px + q$
- Fermat's rule for $x = \mathrm{prox}_{\gamma g}(z)$:

$$\begin{aligned} 0 = \nabla g(x) + \gamma^{-1}(x - z) \quad &\Leftrightarrow \quad 0 = Px + q + \gamma^{-1}(x - z) \\ &\Leftrightarrow \quad (I + \gamma P)x = z - \gamma q \\ &\Leftrightarrow \quad x = (I + \gamma P)^{-1}(z - \gamma q) \end{aligned}$$

- So $\mathrm{prox}_{\gamma g}(z) = (I + \gamma P)^{-1}(z - \gamma q)$

Computational cost

- Evaluating prox requires solving optimization problem

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

- Prox often more expensive to evaluate than gradient
  - Example: Quadratic $g(x) = \frac{1}{2}x^T P x + q^T x$:

$$\mathbf{prox}_{\gamma g}(z) = (I + \gamma P)^{-1}(z - \gamma q), \qquad \nabla g(z) = Pz + q$$

- But typically cheap to evaluate for separable functions
- Prox often used for nondifferentiable and separable functions