Subdifferentials and Proximal Operators

Pontus Giselsson

Learning goals

- Be able to derive subdifferential and proximal operator formulas
- Understand that subdifferentials define affine minorizors
- Existence of subgradient for convex functions
- Understand maximal monotonicity and Minty's theorem
- Know strong monotonicity and relation to strong convexity
- Know different characterizations of smoothness
- Understand and be able to use Fermat's rule
- Know subdifferential calculus rules
- Understand that prox evaluates subdifferential

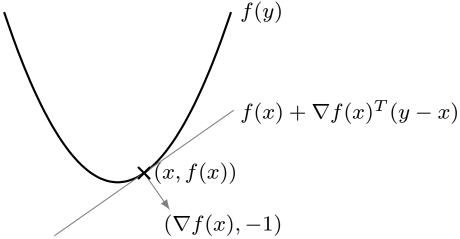
Subdifferentials

Gradients of convex functions

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

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

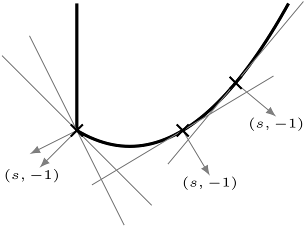
  for all $x, y \in \mathbb{R}^n$



- 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:



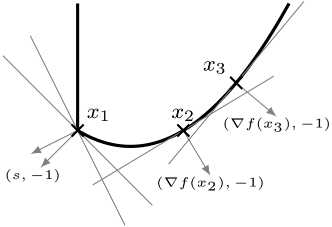
  - coincide with f at x
  - define normal vector $(s, -1)$ to epigraph of f
  - can be one of many affine minorizers at nondifferentiable points x
- Subdifferential of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at x is set of vectors s satisfying

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

- 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

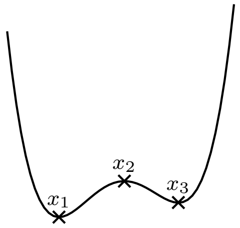
- If f differentiable at x and $\partial f(x) \neq \emptyset$ then $\partial f(x) = \{\nabla f(x)\}$:



- i.e., subdifferential (if nonempty) at x consists of only gradient

Subgradient existence – Nonconvex example

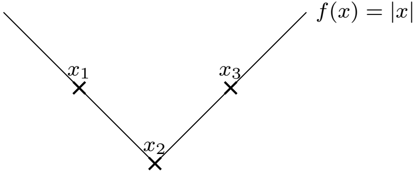
- Function can be differentiable at x but $\partial f(x) = \emptyset$



  - 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

Subgradient existence – Convex example

- Consider the convex function:



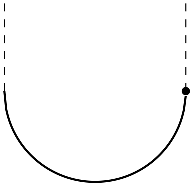
- What are the subdifferentials at points x_1, x_2, x_3?
  - Subdifferential at x_1 is -1 (affine minorizer with slope -1)
  - Subdifferential at x_2 is [-1,1] (affine minorizers with slope [-1,1])
  - Subdifferential at x_3 is 1 (affine minorizer with slope 1)

Fact:

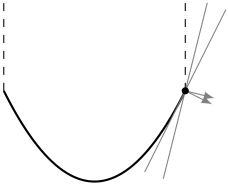
- For *finite-valued* convex functions, a subgradient exists for every x

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 $x = 1$

Monotonicity

- Subdifferential operator is *monotone*:

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

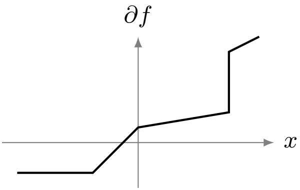
for all $s_x \in \partial f(x)$ and $s_y \in \partial f(y)$

- 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 $v \in Ay$

- If $n = 1$, then $A = \partial f$ for some function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$
- If $n \geq 2$ 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

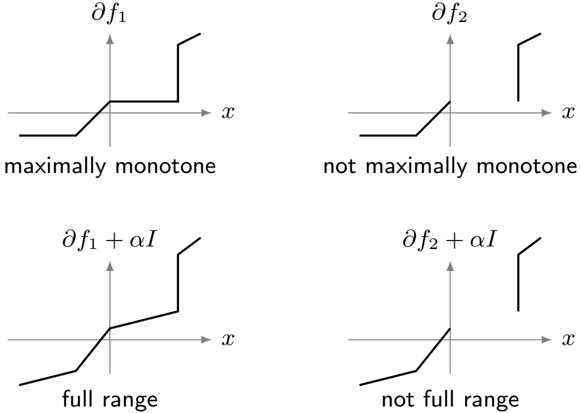
$$\operatorname{gph} \partial f \subset \operatorname{gph} \partial g,$$

  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 $\mathrm{range}(\alpha I + \partial f) = \mathbb{R}^n$



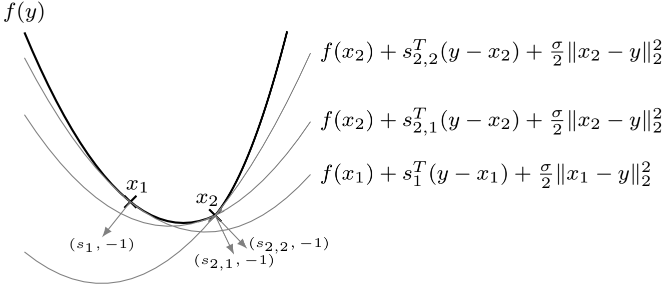
- Interpretation: No "holes" in $\mathrm{gph}\,\partial f$

Strong convexity

- 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

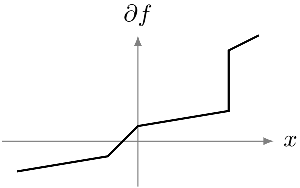
- If f σ-strongly convex function, then ∂f is *σ-strongly monotone*:

$$(s_x - s_y)^T(x - y) \geq \sigma \|x - y\|_2^2$$

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

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

  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

(i) f is closed and σ-strongly convex

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

Proof:

(i)$\Rightarrow$(ii): we know this from before

(ii)$\Rightarrow$(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

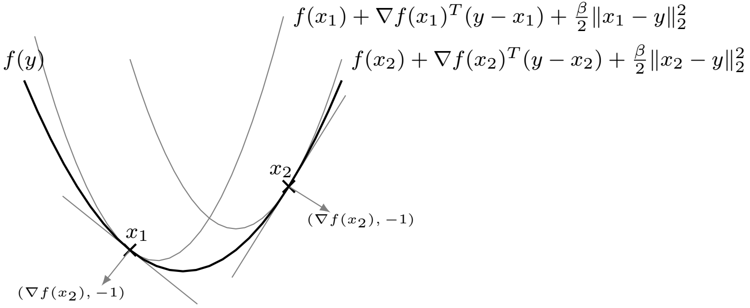
Smoothness and convexity

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

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

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

- f has convex quadratic majorizers and affine minorizers



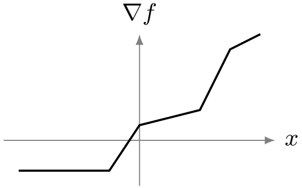
- Quadratic upper bound is called *descent lemma*

Gradient of smooth convex function

- 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:

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

Smooth convex functions – Equivalences

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 closed convex and satisfies descent lemma (is β-smooth)

- Implication (ii)$\Rightarrow$(i) is called the Baillon-Haddad theorem
- Will connect smoothness and strong convexity via conjugates in next lecture

Fermat's rule

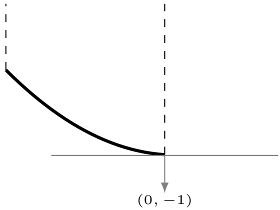
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then x minimizes f if and only if

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

- Proof: x minimizes f if and only if

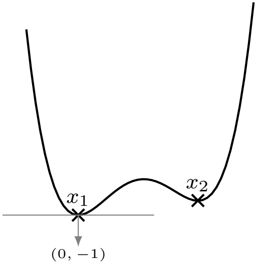
$$f(y) \geq f(x) + 0^T(y - x) \quad \text{for all } y \in \mathbb{R}^n$$

  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 $\operatorname{relint} \operatorname{dom} f_1 \cap \operatorname{relint} \operatorname{dom} f_2 \neq \emptyset$:

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

- 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

If f closed convex and $\operatorname{relint}\operatorname{dom}(f \circ L) \neq \emptyset$:

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

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

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

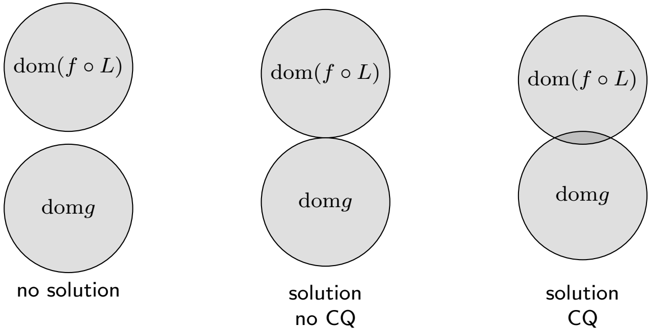
  which by Fermat's rule is equivalent to x solution to (1)
- Algorithms search for x that satisfy $0 \in L^T \partial f(Lx) + \partial g(x)$

A comment to 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)

Proximal operators

Proximal operator

- Proximal operator of (convex) g defined as:

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

  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: $x = \text{prox}_{\gamma g}(z)$ 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 $x = \text{prox}_{\gamma g}(z)$ is computed
- Often used in algorithms when g nonsmooth (no gradient exists)

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

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

- Projection onto C equals prox of indicator function of C

Proximal operator – Example 1

Let $g(x) = \frac{1}{2}x^T Hx + h^T x$ with H positive semidefinite

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

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

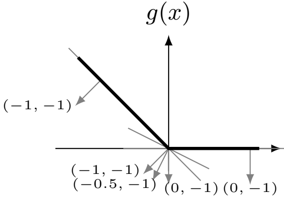
- So $\mathrm{prox}_{\gamma g} z = (I + \gamma H)^{-1}(z - \gamma h)$

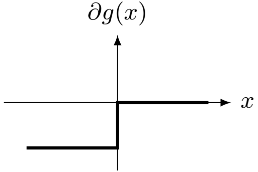
Proximal operator – Example 2

- Consider the function g with subdifferential ∂g:

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

- Graphical representations





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

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

Proximal operator – Example 2 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

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

  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

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

Computational cost

- Evaluating prox requires solving optimization problem

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

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

$$\mathrm{prox}_{\gamma g}(z) = (I + \gamma H)^{-1}(z - \gamma h), \qquad \nabla g(z) = Hz - h$$