Introduction and Preliminaries

Ernest K. Ryu
Seoul National University

Mathematical and Numerical Optimization
Fall 2020

Last edited: 10/29/2020

Outline

Convex optimization via monotone operators

Monotone operator theory is an elegant and powerful tool.

We use this tool to provide a unified analysis of **many** classical and modern first-order convex optimization methods.

Optimization methods to cover

§2 Gradient descent, dual ascent, proximal point method, method of multipliers, proximal method of multipliers, forward-backward splitting, Douglas–Rachford splitting, Davis–Yin splitting, proximal gradient method, iterative soft thresholding, consensus optimization, forward-Douglas–Rachford, variable metric proximal point, variable metric forward-backward splitting, backward-backward method.

§3 ADMM, alternating minimization algorithm (Tseng), PDHG (Chambolle–Pock), Condat–Vũ, proximal method of multipliers with function lineraization, PAPC/PDFP^{2}O, linearized method of multipliers, PD3O, proximal ADMM, linearized ADMM, DYS 3-block ADMM, doubly linearized method of multipliers.

§5 Coordinate gradient descent block-coordinate descent, coordinate proximal-gradient descent, stochastic dual coordinate ascent, MISO/finito, coordinate updates on conic programs.

§6 ARock, asynchronous coordinate gradient descent, asynchronous ADMM.

Optimization methods to cover

§7 Stochastic forward-backward method, stochastic gradient descent, stochastic proximal gradient method, stochastic proximal simultaneous gradient method, stochastic Condat–Vũ.

§8 Function-linearized proximal ADMM, golden ratio ADMM, doubly-linearized ADMM, partial linearization, near-circulant splitting, Jacobi ADMM, 2-1-2 ADMM, Trip-ADMM, split Bregman method, four-block 2-1-2-4-3-4 ADMM.

§11 Distributed ADMM, decentralized ADMM, distributed gradient descent, method of diffusion, adapt-then-combine, PG-EXTRA, NIDS.

§12 Nesterov accelerated gradient method, FISTA, accelerated proximal point method.

1st-order vs. 2nd-order methods

2nd-order methods:

- Use second-order derivatives or their approximations.
- Focus of 70s–90s. Effective for smaller problems.
- Require fewer iterations to solve the optimization problem to high accuracy, even up to machine precision.

1st-order methods:

- Can be described and analyzed with gradients and subgradients.
- Current focus. Effective for larger problems.
- Lower computational cost per iteration. For large problems, one iteration of a 2nd-order method is infeasible, while 1st-order methods can solve to acceptable accuracy.
- 1st-order methods are extremely simple; 2- or 3-line description. Simpler methods are easy to try out and to parallelize.

1st-order vs. 2nd-order methods

Two class of methods are usually not in competition.

- When a high-accuracy solution is needed, second-order methods should be used. For small problems, use second-order methods, since no reason to forgo the high accuracy.
- In large-scale problems, one should use first-order methods and tolerate inaccuracy. Most engineering applications only require a few digits of accuracy in its solution.

Convergence and convergence rates

The total cost of a method is

$$(\text{cost per iteration}) \times (\text{number of iterations}).$$

(cost per iteration): examining the computational cost of the individual components of the method.
(number of iterations): analyzing the rate of convergence.

In optimization, methods are often compared with cost per iteration. (We just made this very argument.) However, a method with a low cost per iteration has the potential, not a guarantee, to be efficient.

Nevertheless, focusing on the cost per iteration is a useful simplification. We focus on establishing convergence without paying much attention to the rate of convergence.

Limitations of monotone operator theory

We provide streamlined convergence proofs and only discuss results that fit this approach. Such results are simple but often not the strongest.

Proofs based on monotone operator theory use monotonicity, rather than convexity, as the key property. This line of analysis does not lead to results involving function values. For example, the gradient method $x^{k+1} = x^k - \alpha \nabla f(x^k)$ converges, under suitable assumptions, with rate $\|\nabla f(x^k)\|^2 \leq \mathcal{O}(1/k)$ (proved with monotonicity) and $f(x^k) - f(x^\star) \leq \mathcal{O}(1/k)$ (proved with convexity).

Convex optimization theory goes beyond monotone operators, although monotone operators do play a central role.

Outline

Overloaded set notation

We overload many standard notation defined for points to sets:
For $\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$, $A, B \subseteq \mathbb{R}^n$, $M \in \mathbb{R}^{m \times n}$:

$$\begin{aligned} \alpha A &= \{\alpha a \,|\, a \in A\} \\ x + A &= \{x + a \,|\, a \in A\} \\ MA &= \{Ma \,|\, a \in A\} \\ A + B &= \{a + b \,|\, a \in A,\, b \in B\} \end{aligned}$$

The sum $A + B$ is called the Minkowski sum.

Lipschitz continuity

$\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^m$ is L-Lipschitz (continuous) if

$$\|\mathbb{T}(x) - \mathbb{T}(y)\| \leq L\|x - y\| \qquad \forall\, x, y \in \mathbb{R}^n.$$

$\mathbb{T}$ is Lipschitz (continuous) if L-Lipschitz for some $L \in (0, \infty)$.

- If $\mathbb{T}$ is Lipschitz, it is continuous.
- If $\mathbb{T}_1$ and $\mathbb{T}_2$ are L_1- and L_2-Lipschitz, then $\mathbb{T}_1 \circ \mathbb{T}_2$ is $L_1 L_2$-Lipschitz.
- If $\mathbb{T}_1$ and $\mathbb{T}_2$ are L_1- and L_2-Lipschitz, then $\alpha_1 \mathbb{T}_1 + \alpha_2 \mathbb{T}_2$ is $(|\alpha_1| L_1 + |\alpha_2| L_2)$-Lipschitz.

Interior

Closed ball of radius r centered at x:

$$B(x, r) = \{y \in \mathbb{R}^n \mid \|y - x\| \leq r\}$$

Interior of $C \subseteq \mathbb{R}^n$:

$$\operatorname{int} C = \{x \in C \mid B(x, r) \subseteq C \text{ for some } r > 0\}$$

Closure of $C \subseteq \mathbb{R}^n$: $\operatorname{cl} C$

Boundary of $C \subseteq \mathbb{R}^n$: $\operatorname{cl} C \backslash \operatorname{int} C$

Relative interior

Affine set: $x_0 + V$, where $x_0 \in \mathbb{R}^n$ and $V \subseteq \mathbb{R}^n$ is a subspace.

Affine hull of $C \subseteq \mathbb{R}^n$:

$$\operatorname{aff} C = \{\theta_1 x_1 + \cdots + \theta_k x_k \,|\, x_1, \ldots, x_k \in C,\ \theta_1 + \cdots + \theta_k = 1,\ k \geq 1\}$$

Affine hull is the smallest affine set containing C.

Relative interior of $C \subset \mathbb{R}^n$:

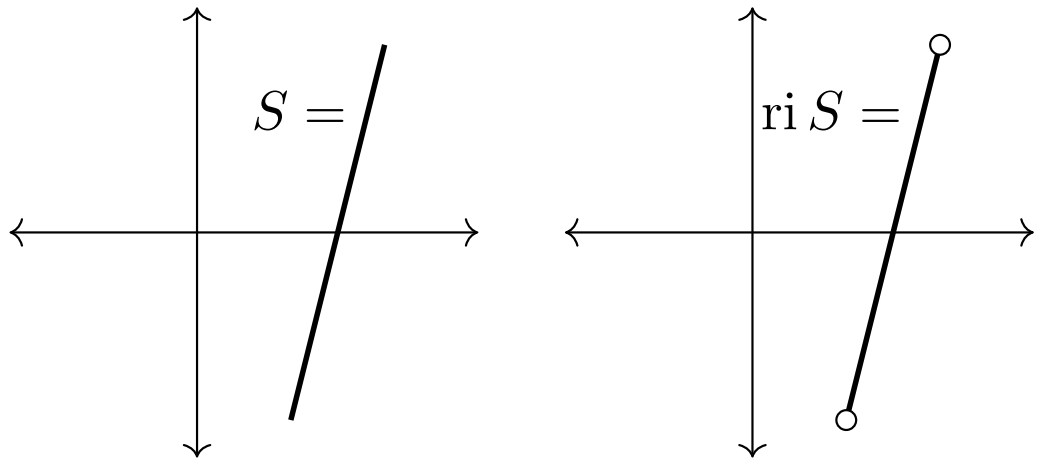
$$\operatorname{ri} C = \{x \in C \,|\, B(x, r) \cap \operatorname{aff} C \subseteq C \text{ for some } r > 0\}$$

$\operatorname{ri} C$ of a nonempty convex set C is nonempty.

Relative boundary of $C \subseteq \mathbb{R}^n$: $\operatorname{cl} C \backslash \operatorname{ri} C$

Relative interior example

$$S = \left\{ (x, y) \in \mathbb{R}^2 \,|\, x \in [0.5, 1],\, y = 4x - 3 \right\}.$$



Functions

Extended-valued functions map $\mathbb{R}^n$ to the extended real line $\mathbb{R} \cup \{\pm\infty\}$. (Some saddle functions have value $\pm\infty$.)

(Effective) domain of f:

$$\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$$

f is convex if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \quad \forall\, x, y \in \operatorname{dom} f,\ \theta \in (0,1).$$

f is strictly convex if the inequality is strict when $x \neq y$. f is (strictly) concave if $-f$ is (strictly) convex. When f is convex, $\operatorname{dom} f$ is convex.

CCP functions

f is CCP if closed, convex, and proper:

- f is proper if $f(x) = -\infty$ never and $f(x) < \infty$ somewhere.
- Proper f is closed if epigraph of f

$$\operatorname{epi} f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$$

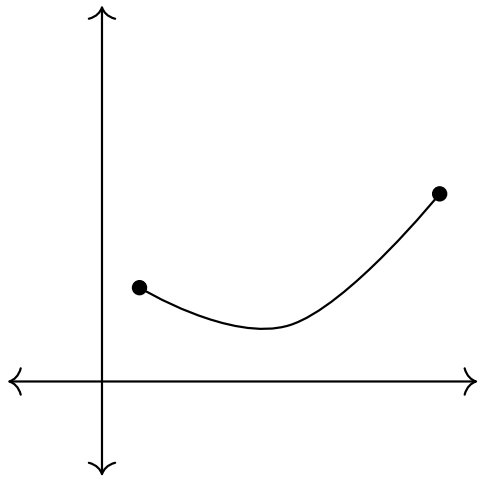
  is closed.

Properties:

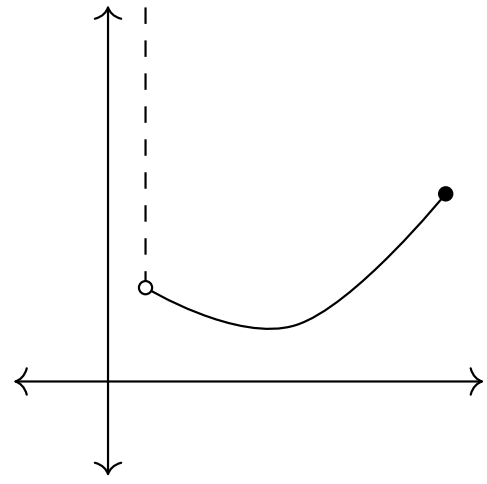
- Most convex functions of interest are closed and proper.
- [f is convex] $\Leftrightarrow$ [$\operatorname{epi} f$ is convex]
- For proper f, [f closed] $\Leftrightarrow$ [f is lower semi-continuous]
- [f CCP] $\Leftrightarrow$ [$\operatorname{epi} f$ nonempty closed convex without a vertical line]

vertical line $= \{x_0\} \times \mathbb{R}$.

CCP function example



Closed convex function

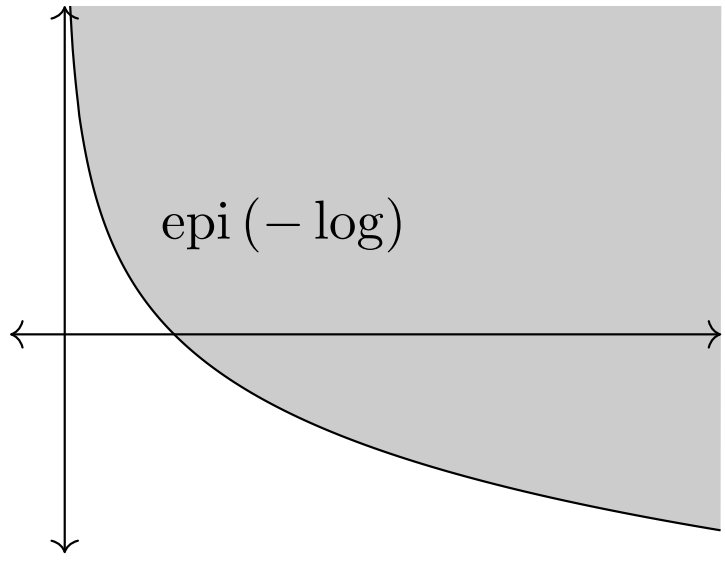


Convex but not closed

The dashed line denotes the function value of ∞.

CCP function example

Epigraph of the CCP $-\log$ is a nonempty closed convex set.



Operations preserving CCP

If f and g are CCP functions, $\alpha > 0$, and A is a matrix, then

- αf is CCP
- $f + g$ is CCP provided that $\exists\, x$ such that $f(x) + g(x) < \infty$
- $f(Ax)$ is CCP provided that $\exists\, x$ such that $f(Ax) < \infty$

Indicator function

For $S \subseteq \mathbb{R}^n$, define the *indicator function*

$$\delta_S(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{otherwise.} \end{cases}$$

If S is convex, closed, and nonempty, then δ_S is CCP.

Argmin

Set of minimizers of f:

$$\operatorname{argmin} f = \left\{ x \in \mathbb{R}^n \,\middle|\, f(x) = \inf_{z \in \mathbb{R}^n} f(z) \right\}$$

When f is CCP, $\operatorname{argmin} f$ is closed convex, possibly empty.

When f is strictly convex, $\operatorname{argmin} f$ has at most one point.

Subgradient

$g \in \mathbb{R}^n$ is a *subgradient* of convex f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \qquad \forall\, y \in \mathbb{R}^n.$$

The *subdifferential* of convex f at x is

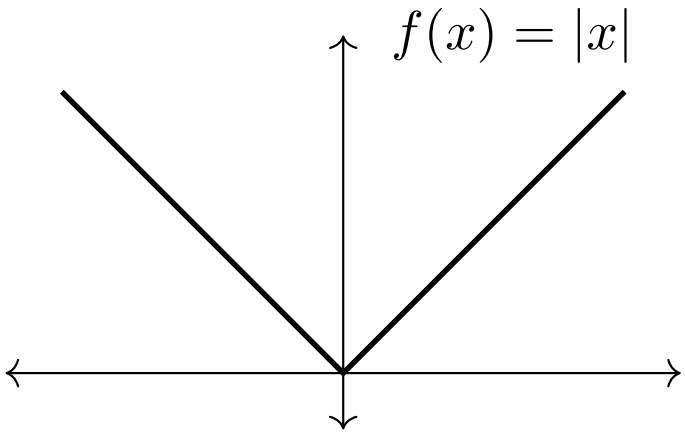
$$\partial f(x) = \{g \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle g, y - x \rangle,\ \forall\, y \in \mathbb{R}^n\},$$

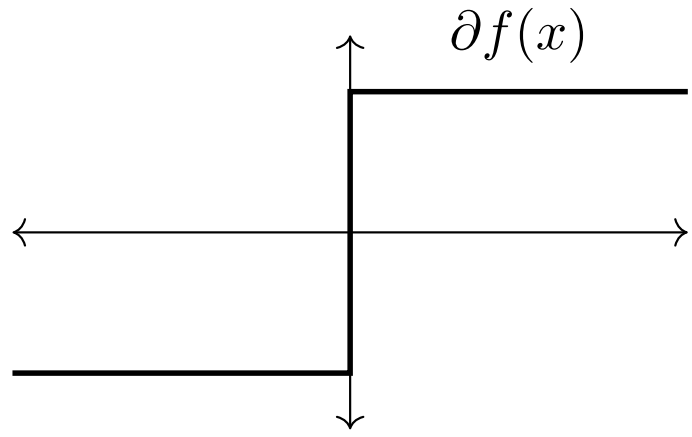
i.e., $\partial f(x) = \{\text{subgradients of } f \text{ at } x\}$.

- $\partial f(x)$ is closed convex
- [Convex f is differentiable at x] $\Leftrightarrow$ [$\partial f(x)$ is a singleton]
- $[x^\star \in \operatorname{argmin} f] \Leftrightarrow [0 \in \partial f(x^\star)]$

Subdifferential example

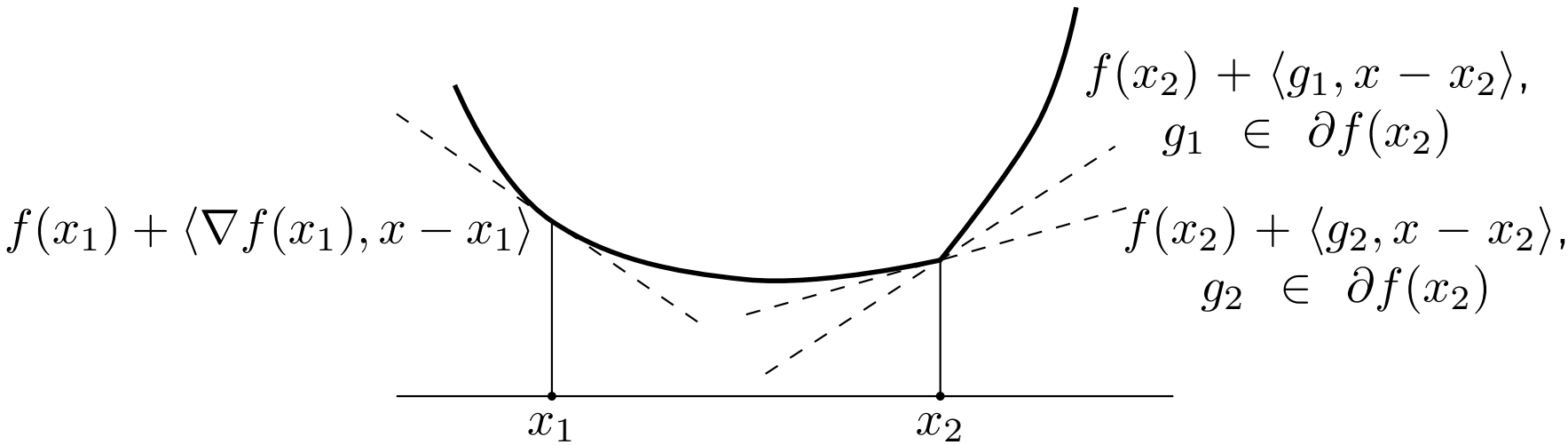
The absolute value function is differentiable everywhere except at 0.





Subdifferential example

At x_1, f is differentiable and $\partial f(x_1) = \{\nabla f(x_1)\}$.
At x_2, f is not differentiable and has many subgradients.

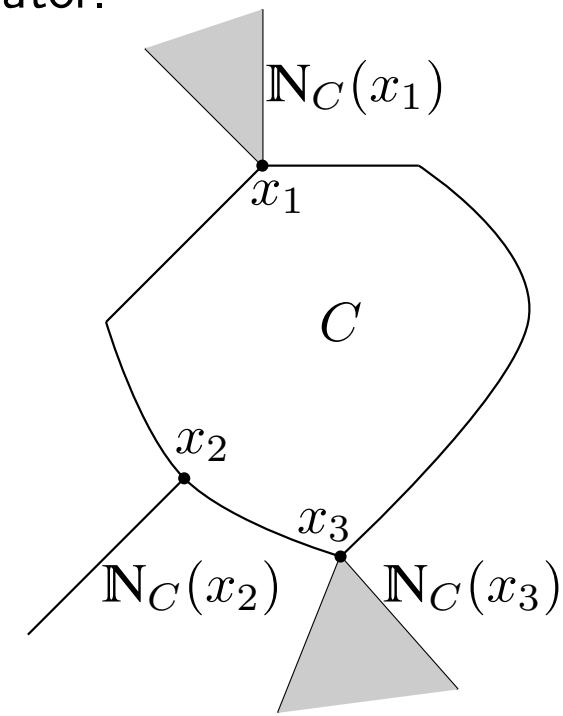


Normal cone operator

$C \subseteq \mathbb{R}^n$ closed convex. Then

$$\partial\delta_C(x) = \mathbb{N}_C(x) = \begin{cases} \emptyset & x \notin C \\ \{y \mid \langle y, z - x\rangle \leq 0 \ \forall z \in C\} & x \in C \end{cases}$$

is the normal cone operator.



We primarily use $\mathbb{N}_C$ as notational shorthand for $\partial\delta_C$.

Subdifferentiability

Convex f is subdifferentiable at x if $\partial f(x) \neq \emptyset$.

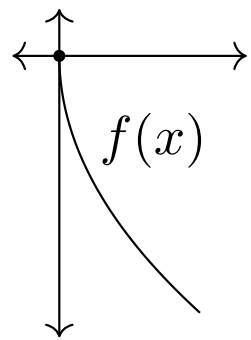
When f is CCP,

- $\partial f(x) = \emptyset$ where $x \notin \operatorname{dom} f$
- $\partial f(x) \neq \emptyset$ for any $x \in \operatorname{ri} \operatorname{dom} f$
- f may or may not be subdifferentiable on $\operatorname{dom} f \backslash \operatorname{ri} \operatorname{dom} f$

Non-subdifferentiability example

$$f(x) = \begin{cases} -\sqrt{x} & \text{for } x \geq 0 \\ \infty & \text{for } x < 0 \end{cases}$$

is not subdifferentiable at $x = 0$, although $0 \in \operatorname{dom} f$ and f is CCP.



Subgradient identities

Several standard identities for gradients also hold for subdifferentials when regularity conditions hold:

- $\partial \alpha f = \alpha \partial f$, if $\alpha > 0$
- $g(x) = f(Ax)$, $\partial g = A^\mathsf{T} \partial f A$, if $\mathcal{R}(A) \cap \operatorname{ri} \operatorname{dom} f \neq \emptyset$
- $\partial(f + g) = \partial f + \partial g$, if $\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g \neq \emptyset$

Without regularity conditions,

$$\partial g(x) \supseteq A^\mathsf{T} \partial f(Ax), \qquad \partial(f + g)(x) \supseteq \partial f(x) + \partial g(x)$$

Regularity conditions

Say we have "P $\Rightarrow$ Q".

Then, if P "usually" holds then Q "usually" holds, and we say P is a regularity condition, since P is satisfied in the usual "regular" case.

Examples:

- $[\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g \neq \emptyset] \Rightarrow [\partial(f+g) = \partial f + \partial g]$.
- [Slater's constraint qualification] $\Rightarrow$ [strong duality]

Constraint qualifications are regularity conditions ensuring strong duality.

Conjugate function

Conjugate function of f:

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - f(x)\}$$

Properties: when f is CCP

- f^* is CCP and $f^{**} = f$
- $(\nabla f)^{-1} = \nabla f^*$ when f and f^* are differentiable
- $(\partial f)^{-1} = \partial f^*$ in general (more on this next section)

Strong convexity

CCP f is μ-strongly convex if:

- $f(x) - (\mu/2)\|x\|^2$ is convex.
- $\langle \partial f(x) - \partial f(y), x - y \rangle \geq \mu \|x - y\|^2$ for all x, y.
- $\nabla^2 f(x) \succeq \mu I$ for all x if f is twice-differentiable.

These conditions are equivalent.

If f is μ-strongly convex and g is convex, then $f + g$ is μ-strongly convex. To clarify, strong convexity does not imply differentiability.

L-smooth function

CCP f is L-smooth if:

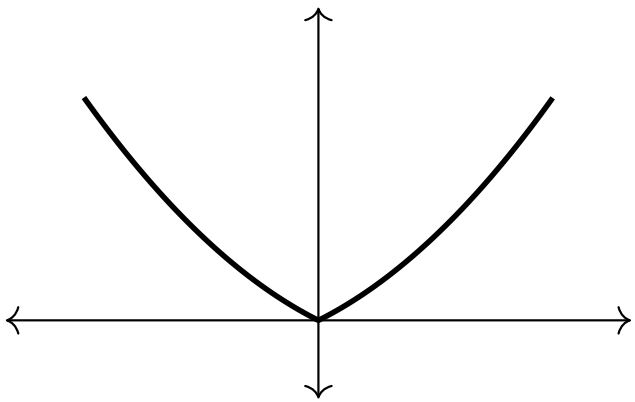
- $f(x) - (L/2)\|x\|^2$ is concave.
- f is differentiable and $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq (1/L)\|\nabla f(x) - \nabla f(y)\|^2$ for all x, y.
- f is differentiable and ∇f is L-Lipschitz.
- $\nabla^2 f(x) \preceq LI$ for all x if f is twice-differentiable.

These conditions are equivalent.

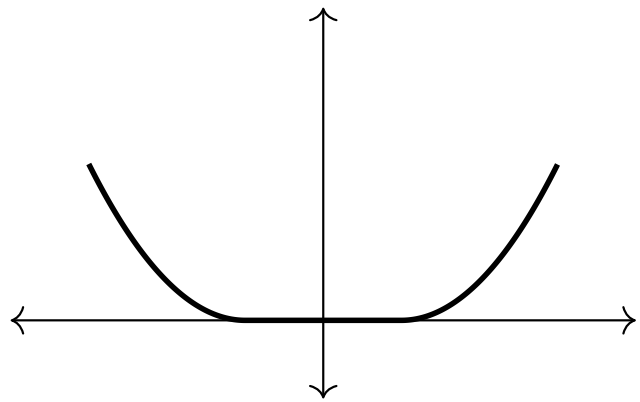
"L-smoothness", which implies once-continuous differentiability, is somewhat non-standard; "smoothness" often means infinite differentiability in other fields of mathematics.

Strong convexity and smoothness

Informally speaking, μ-strongly convex functions have upward curvature of at least μ and L-smooth convex functions have upward curvature of no more than L. We can think of nondifferentiable points to be points with infinite curvature.



Strongly convex but not smooth



Smooth but not strongly convex.

Strong convexity and smoothness

If f is μ-strongly convex and L-smooth, then $\mu \leq L$ since

$$\begin{aligned}\mu\|x-y\|^2 &\leq \langle \nabla f(x) - \nabla f(y), x-y\rangle \\ &\leq \|\nabla f(x) - \nabla f(y)\|\|x-y\| \\ &\leq L\|x-y\|^2.\end{aligned}$$

Strong convexity and smoothness are dual properties:
if f CCP, [f is μ-strongly convex] $\Leftrightarrow$ [f^* is $(1/\mu)$-smooth]

Convex-concave saddle function and saddle point

Let $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\pm\infty\}$. We say $\mathbf{L}(x, u)$ is convex-concave if $\mathbf{L}$ is convex in x when u is fixed and concave in u when x is fixed.

$(x^\star, u^\star)$ is a saddle point of $\mathbf{L}$ if

$$\mathbf{L}(x^\star, u) \leq \mathbf{L}(x^\star, u^\star) \leq \mathbf{L}(x, u^\star) \qquad \forall\, x \in \mathbb{R}^n,\, u \in \mathbb{R}^m.$$

Duality from saddle functions

Primal problem generated by $\mathbf{L}$:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \sup_{u \in \mathbb{R}^m} \mathbf{L}(x, u)$$

Dual problem generated by $\mathbf{L}$:

$$\underset{u \in \mathbb{R}^m}{\text{maximize}} \quad \inf_{x \in \mathbb{R}^n} \mathbf{L}(x, u)$$

Trick is to find $\mathbf{L}$ that generates the primal problem of interest.

Duality example: linearly constrained minimization

$$\mathbf{L}(x, u) = f(x) + \langle u, Ax - b \rangle$$

generates the primal problem

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

and dual problem

$$\underset{u \in \mathbb{R}^m}{\text{maximize}} \quad -f^*(-A^\intercal u) - b^\intercal u.$$

If $\{x \mid Ax = b\} \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$ holds, then $d^\star = p^\star$.

Duality example: Fenchel–Rockafellar dual

$$\mathbf{L}(x, u) = f(x) + \langle u, Ax \rangle - g^*(u)$$

generates the primal problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) + g(Ax)$$

and dual problem

$$\underset{u \in \mathbb{R}^m}{\text{maximize}} \quad -f^*(-A^\intercal u) - g^*(u).$$

If $A\operatorname{dom} f \cap \operatorname{int}\operatorname{dom} g \neq \emptyset$ holds, then $d^\star = p^\star$.

Weak and strong duality

Weak duality: $d^\star \leq p^\star$. Always holds.

Proof. For any x, u we have

$$\inf_x \mathbf{L}(x,u) \leq \mathbf{L}(x,u)$$

$$\sup_u \inf_x \mathbf{L}(x,u) \leq \sup_u \mathbf{L}(x,u)$$

$$d^\star = \sup_u \inf_x \mathbf{L}(x,u) \leq \inf_x \sup_u \mathbf{L}(x,u) = p^\star.$$

□

Strong duality: $d^\star = p^\star$. Holds often but not always in convex optimization. Regularity conditions that ensure strong duality are called constraint qualifications.

Total duality

Total duality: a primal solution exists, a dual solution exists, and strong duality holds.

[Total duality] $\Leftrightarrow$ [$\mathbf{L}$ has a saddle point]

When total duality holds, solving the primal and dual optimization problems is equivalent to finding a saddle point of $\mathbf{L}$.

We will later see that total duality is the regularity condition that ensures primal-dual methods converge.

Total duality

Proof. Assume $\mathbf{L}$ has a saddle point $(x^\star, u^\star)$. Then

$$\begin{aligned}\mathbf{L}(x^\star, u^\star) &= \inf_x \mathbf{L}(x, u^\star) \\ &\leq \sup_u \inf_x \mathbf{L}(x, u) = d^\star \\ &\leq \inf_x \sup_u \mathbf{L}(x, u) = p^\star \\ &\leq \sup_u \mathbf{L}(x^\star, u) = \mathbf{L}(x^\star, u^\star),\end{aligned}$$

and equality holds throughout.

$\inf_x \sup_u \mathbf{L}(x, u) = \sup_u \mathbf{L}(x^\star, u)$, so $x^\star$ is a primal solution.
$\inf_x \mathbf{L}(x, u^\star) = \sup_u \inf_x \mathbf{L}(x, u)$, so $u^\star$ is a dual solution.
$d^\star = \sup_u \inf_x \mathbf{L}(x, u) = \inf_x \sup_u \mathbf{L}(x, u) = p^\star$, so strong duality holds.

Total duality

Assume total duality and $x^\star$, $u^\star$ are primal, dual solutions. Then

$$\begin{aligned}\inf_x \mathbf{L}(x, u^\star) &= \sup_u \inf_x \mathbf{L}(x, u) = d^\star \\ &= \inf_x \sup_u \mathbf{L}(x, u) = p^\star \\ &= \sup_u \mathbf{L}(x^\star, u).\end{aligned}$$

Since

$$\mathbf{L}(x^\star, u^\star) \leq \sup_u \mathbf{L}(x^\star, u) = \inf_x \mathbf{L}(x, u^\star) \leq \mathbf{L}(x^\star, u^\star)$$

equality holds throughout and we conclude

$$\sup_u \mathbf{L}(x^\star, u) = \mathbf{L}(x^\star, u^\star) = \inf_x \mathbf{L}(x, u^\star),$$

i.e., $(x^\star, u^\star)$ is a saddle point. □

Proximal operator

Proximal operator with respect to αf:

$$\mathrm{Prox}_{\alpha f}(y) = \underset{x \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \alpha f(x) + \frac{1}{2}\|x - y\|^2 \right\}$$

for CCP f and $\alpha > 0$. When $\alpha = 1$, write Prox_f.

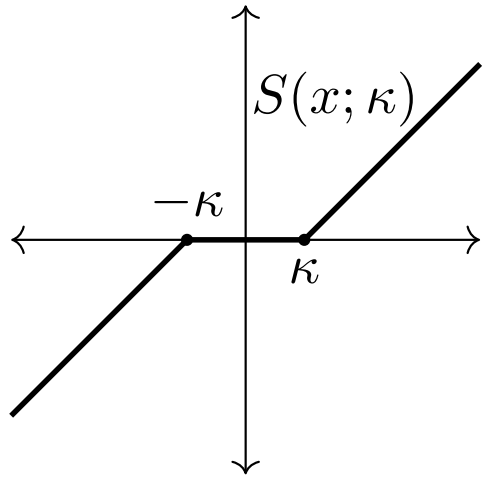
If f is CCP, then $\mathrm{Prox}_{\alpha f}$ is well defined, i.e., argmin uniquely exists.

Proximal operator example: Soft-thresholding

Soft-thresholding operator $S(x;\kappa) = \mathrm{Prox}_{\kappa\|\cdot\|_1}(x)$ has closed-form

$$(S(x;\kappa))_i = \begin{cases} x_i - \kappa & \text{for } \kappa < x_i \\ 0 & \text{for } -\kappa \leq x_i \leq \kappa \\ x_i + \kappa & \text{for } x_i < -\kappa \end{cases}$$

for $i = 1, \ldots, n$.



Proximal operator example: Projection

Projection onto nonempty closed convex $C \subseteq \mathbb{R}^n$:

$$\Pi_C(y) = \operatorname*{argmin}_{x \in C} \|x - y\|$$

Since $\operatorname{Prox}_{\alpha\delta_C} = \operatorname{Prox}_{\delta_C} = \Pi_C$ for any $\alpha > 0$, proximal operators generalize projections.

Proximable functions

Eevaluating $\mathrm{Prox}_{\alpha f}$ is an optimization problem itself. However, many interesting convex f has a closed-form solution for $\mathrm{Prox}_{\alpha f}$ so it is a useful subroutine.

f is proximable (informal definition) if $\mathrm{Prox}_{\alpha f}$ is computationally efficient to evaluate. Catalog of proximable functions in several papers.

We decompose an optimization problem into smaller, simpler differentiable or proximable functions and operate on them separately.