Introduction to Convex Optimization

Lieven Vandenberghe

Electrical Engineering Department, UCLA

IPAM Graduate Summer School: Computer Vision
August 5, 2013
Convex optimization problem

minimize $f_0(x)$
subject to $f_i(x) \leq 0, \quad i = 1, \ldots, m$
$Ax = b$

objective and inequality constraint functions $f_i$ are convex:

$$f_i(\theta x + (1 - \theta)y) \leq \theta f_i(x) + (1 - \theta)f_i(y) \quad \text{for } 0 \leq \theta \leq 1$$

- can be solved globally, with similar low complexity as linear programs
- surprisingly many problems can be solved via convex optimization
- provides tractable heuristics and relaxations for non-convex problems
History

• 1940s: linear programming

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad a_i^T x \leq b_i, \quad i = 1, \ldots, m
\end{align*}
\]

• 1950s: quadratic programming

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} x^T P x + q^T x \\
\text{subject to} & \quad a_i^T x \leq b_i, \quad i = 1, \ldots, m
\end{align*}
\]

• 1960s: geometric programming

• since 1990: semidefinite programming, second-order cone programming, quadratically constrained quadratic programming, robust optimization, sum-of-squares programming, . . .
New applications since 1990

- linear matrix inequality techniques in control
- semidefinite programming relaxations in combinatorial optimization
- support vector machine training via quadratic programming
- circuit design via geometric programming
- $\ell_1$-norm optimization for sparse signal reconstruction
- applications in structural optimization, statistics, machine learning, signal processing, communications, image processing, computer vision, quantum information theory, finance, power distribution, . . .
Advances in convex optimization algorithms

**Interior-point methods**

- 1984 (Karmarkar): first practical polynomial-time algorithm for LP
- 1984-1990: efficient implementations for large-scale LPs
- 1990s: high-quality software packages for conic optimization
- 2000s: convex modeling software based on interior-point solvers

**First-order algorithms**

- fast gradient methods, based on Nesterov’s methods from 1980s
- extensions to nondifferentiable or constrained problems
- multiplier/splitting methods for large-scale and distributed optimization
Overview

1. Introduction to convex optimization theory
   - convex sets and functions
   - conic optimization
   - duality

2. Introduction to first-order algorithms
   - (proximal) gradient algorithm
   - splitting and alternating minimization methods
1. Convex optimization theory

- convex sets and functions
- conic optimization
- duality
**Convex set**

contains the line segment between any two points in the set

\[ x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Rightarrow \quad \theta x_1 + (1 - \theta)x_2 \in C \]
Basic examples

**Affine set:** solution set of linear equations $Ax = b$

**Halfspace:** solution of one linear inequality $a^T x \leq b$ ($a \neq 0$)

**Polyhedron:** solution of finitely many linear inequalities $Ax \leq b$

**Ellipsoid:** solution of positive definite quadratic inequality

$$(x - x_c)^T A (x - x_c) \leq 1 \quad (A \text{ positive definite})$$

**Norm ball:** solution of $\|x\| \leq R$ (for any norm)

**Positive semidefinite cone:** $S^n_+ = \{ X \in S^n \mid X \succeq 0 \}$

the intersection of any number of convex sets is convex
Convex function

Domain $\text{dom } f$ is a convex set and Jensen’s inequality holds:

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

for all $x, y \in \text{dom } f$, $0 \leq \theta \leq 1$

$f$ is concave if $-f$ is convex
Examples

• linear and affine functions are convex and concave

• $\exp x, -\log x, x \log x$ are convex

• $x^\alpha$ is convex for $x > 0$ and $\alpha \geq 1$ or $\alpha \leq 0$; $|x|^\alpha$ is convex for $\alpha \geq 1$

• norms are convex

• quadratic-over-linear function $x^T x/t$ is convex in $x, t$ for $t > 0$

• geometric mean $(x_1 x_2 \cdots x_n)^{1/n}$ is concave for $x \geq 0$

• $\log \det X$ is concave on set of positive definite matrices

• $\log(e^{x_1} + \cdots e^{x_n})$ is convex
Differentiable convex functions

differentiable $f$ is convex if and only if $\text{dom } f$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \text{dom } f$$

twice differentiable $f$ is convex if and only if $\text{dom } f$ is convex and

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \text{dom } f$$
Subgradient

$g$ is a **subgradient** of a convex function $f$ at $x$ if

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom } f$$

the set of all subgradients of $f$ at $x$ is called the **subdifferential** $\partial f(x)$

- $\partial f(x) = \{\nabla f(x)\}$ if $f$ is differentiable at $x$
- convex $f$ is subdifferentiable ($\partial f(x) \neq \emptyset$) on $x \in \text{int dom } f$
Examples

**Absolute value** \( f(x) = |x| \)

\[
f(x) = |x|
\]

**Euclidean norm** \( f(x) = \|x\|_2 \)

\[
\partial f(x) = \begin{cases} 
\frac{1}{\|x\|_2} x & \text{if } x \neq 0, \\
\{g \mid \|g\|_2 \leq 1\} & \text{if } x = 0
\end{cases}
\]
Establishing convexity

1. verify definition

2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$

3. show that $f$ is obtained from simple convex functions by operations that preserve convexity
   - nonnegative weighted sum
   - composition with affine function
   - pointwise maximum and supremum
   - minimization
   - composition
   - perspective
Positive weighted sum & composition with affine function

Nonnegative multiple: $\alpha f$ is convex if $f$ is convex, $\alpha \geq 0$

Sum: $f_1 + f_2$ convex if $f_1, f_2$ convex (extends to infinite sums, integrals)

Composition with affine function: $f(Ax + b)$ is convex if $f$ is convex

Examples

• logarithmic barrier for linear inequalities

\[
f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)\]

• (any) norm of affine function: $f(x) = \|Ax + b\|$
Pointwise maximum

\[ f(x) = \max\{f_1(x), \ldots, f_m(x)\} \]

is convex if \( f_1, \ldots, f_m \) are convex

Example: sum of \( r \) largest components of \( x \in \mathbb{R}^n \)

\[ f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[r]} \]

is convex (\( x_{[i]} \) is \( i \)th largest component of \( x \))

proof:

\[ f(x) = \max\{x_{i_1} + x_{i_2} + \cdots + x_{i_r} \mid 1 \leq i_1 < i_2 < \cdots < i_r \leq n\} \]
Pointwise supremum

\[ g(x) = \sup_{y \in A} f(x, y) \]

is convex if \( f(x, y) \) is convex in \( x \) for each \( y \in A \)

Examples

- maximum eigenvalue of symmetric matrix

\[ \lambda_{\text{max}}(X) = \sup_{\|y\|_2=1} y^T X y \]

- support function of a set \( C \)

\[ S_C(x) = \sup_{y \in C} y^T x \]
Partial minimization

\[ h(x) = \inf_{y \in C} f(x, y) \]

is convex if \( f(x, y) \) is convex in \((x, y)\) and \( C \) is a convex set

**Examples**

- distance to a convex set \( C \): \( h(x) = \inf_{y \in C} \|x - y\| \)
- optimal value of linear program as function of righthand side

\[ h(x) = \inf_{y : Ay \leq c} c^T y \]

follows by taking

\[ f(x, y) = c^T y, \quad \text{dom } f = \{(x, y) \mid Ay \leq x\} \]
Composition

composition of $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$:

$$f(x) = h(g(x))$$

$f$ is convex if

- $g$ convex, $h$ convex and nondecreasing
- $g$ concave, $h$ convex and nonincreasing

(if we assign $h(x) = \infty$ for $x \in \text{dom } h$)

Examples

- $\exp g(x)$ is convex if $g$ is convex
- $1/g(x)$ is convex if $g$ is concave and positive
Vector composition

composition of $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$:

$$f(x) = h(g(x)) = h(g_1(x), g_2(x), \ldots, g_k(x))$$

$f$ is convex if

- $g_i$ convex, $h$ convex and nondecreasing in each argument
- $g_i$ concave, $h$ convex and nonincreasing in each argument

(if we assign $h(x) = \infty$ for $x \in \text{dom } h$)

Example: $\log \sum_{i=1}^{m} \exp g_i(x)$ is convex if $g_i$ are convex
the **perspective** of a function \( f : \mathbb{R}^n \to \mathbb{R} \) is the function \( g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R} \),

\[
g(x, t) = tf(x/t)
\]

\( g \) is convex if \( f \) is convex on \( \text{dom} \, g = \{(x, t) | x/t \in \text{dom} \, f, \ t > 0\} \)

**Examples**

- perspective of \( f(x) = x^T x \) is quadratic-over-linear function

\[
g(x, t) = \frac{x^T x}{t}
\]

- perspective of negative logarithm \( f(x) = -\log x \) is relative entropy

\[
g(x, t) = t \log t - t \log x
\]
Modeling software

Modeling packages for convex optimization

- CVX, YALMIP (MATLAB)
- CVXPY, CVXMOD (Python)
- MOSEK Fusion (several platforms)

assist the user in formulating convex problems, by automating two tasks:

- verifying convexity from convex calculus rules
- transforming problem in input format required by standard solvers

Related packages

general-purpose optimization modeling: AMPL, GAMS
Example

\[
\begin{align*}
\text{minimize} & \quad \|Ax - b\|_2^2 + \|x\|_1 \\
\text{subject to} & \quad 0 \leq x_k \leq 1, \quad k = 1, \ldots, n \\
& \quad x^T P x \leq 1
\end{align*}
\]

\textbf{CVX code} (Grant and Boyd 2008)

\begin{verbatim}
cvx_begin
  variable x(n);
  minimize( square_pos(norm(A*x - b)) + norm(x,1) )
  subject to
    x >= 0;
    x <= 1;
    quad_form(x, P) <= 1;

cvx_end
\end{verbatim}
Outline

- convex sets and functions
- conic optimization
- duality
Conic linear program

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad b - Ax \in K
\end{align*}
\]

- \( K \) a convex cone (closed, pointed, with nonempty interior)
- if \( K \) is the nonnegative orthant, this is a (regular) linear program
- constraint often written as generalized linear inequality \( Ax \preceq_K b \)

widely used in recent literature on convex optimization

- **modeling:** 3 cones (nonnegative orthant, second-order cone, positive semidefinite cone) are sufficient to represent most convex constraints
- **algorithms:** a convenient problem format when extending interior-point algorithms for linear programming to convex optimization
Norm cone

\[ K = \{(x, y) \in \mathbb{R}^{m-1} \times \mathbb{R} \mid \|x\| \leq y\} \]

for the Euclidean norm this is the second-order cone (notation: \( Q^m \))
Second-order cone program

\[ \text{minimize } \quad c^T x \]
\[ \text{subject to } \quad \|B k_0 x + d k_0\|_2 \leq B k_1 x + d k_1, \quad k = 1, \ldots, r \]

Conic LP formulation: express constraints as \( Ax \preceq_K b \)

\[ K = Q^{m_1} \times \cdots \times Q^{m_r}, \quad A = \begin{bmatrix} -B_{10} \\ -B_{11} \\ \vdots \\ -B_{r0} \\ -B_{r1} \end{bmatrix}, \quad b = \begin{bmatrix} d_{10} \\ d_{11} \\ \vdots \\ d_{r0} \\ d_{r1} \end{bmatrix} \]

(assuming \( B_{k0}, d_{k0} \) have \( m_k - 1 \) rows)
Robust linear program

minimize \( c^T x \)
subject to \( a_i^T x \leq b_i \) for all \( a_i \in \mathcal{E}_i, \ i = 1, \ldots, m \)

- \( a_i \) uncertain but bounded by ellipsoid \( \mathcal{E}_i = \{ \bar{a}_i + P_i u \mid \|u\|_2 \leq 1 \} \)
- we require that \( x \) satisfies each constraint for all possible \( a_i \)

SOCP formulation

minimize \( c^T x \)
subject to \( \bar{a}_i^T x + \|P_i^T x\|_2 \leq b_i, \ i = 1, \ldots, m \)

follows from
\[
\sup_{\|u\|_2 \leq 1} (\bar{a}_i + P_i u)^T x = \bar{a}_i^T x + \|P_i^T x\|_2
\]
Second-order cone representable constraints

**Convex quadratic constraint** \((A = LL^T \text{ positive definite})\)

\[
x^T A x + 2 b^T x + c \leq 0
\]

\[
\uparrow
\]

\[
\|L^T x + L^{-1} b\|_2 \leq (b^T A^{-1} b - c)^{1/2}
\]

extends to positive semidefinite singular \(A\)

**Hyperbolic constraint**

\[
x^T x \leq yz, \quad y, z \geq 0
\]

\[
\uparrow
\]

\[
\left\|\begin{bmatrix} 2x \\ y - z \end{bmatrix}\right\|_2 \leq y + z, \quad y, z \geq 0
\]
Second-order cone representable constraints

Positive powers

\[ x^{1.5} \leq t, \quad x \geq 0 \quad \iff \quad \exists z : \quad x^2 \leq tz, \quad z^2 \leq x, \quad x, z \geq 0 \]

- two hyperbolic constraints can be converted to SOC constraints
- extends to powers \( x^p \) for rational \( p \geq 1 \)
- can be used to represent \( \ell_p \)-norm constraints \( \|x\|_p \leq t \) with rational \( p \)

Negative powers

\[ x^{-3} \leq t, \quad x > 0 \quad \iff \quad \exists z : \quad 1 \leq tz, \quad z^2 \leq tx, \quad x, z \geq 0 \]

- two hyperbolic constraints on r.h.s. can be converted to SOC constraints
- extends to powers \( x^p \) for rational \( p < 0 \)
Example

\[
\begin{align*}
\text{minimize} \quad & \|Ax - b\|_2^2 + \sum_{k=1}^{N} \|B_k x\|_2 \\
\text{arises in total-variation deblurring}
\end{align*}
\]

**SOCP formulation** (auxiliary variables \(t_0, \ldots, t_N\))

\[
\begin{align*}
\text{minimize} \quad & t_0 + \sum_{i=1}^{N} t_i \\
\text{subject to} \quad & \left\| \begin{bmatrix} 2(Ax - b) \\ t_0 - 1 \end{bmatrix} \right\|_2 \leq t_0 + 1 \\
& \|B_k x\|_2 \leq t_k, \quad k = 1, \ldots, N \\
\end{align*}
\]

first constraint is equivalent to \(\|Ax - b\|_2^2 \leq t_0\)
Positive semidefinite cone

\[ S^p = \{ \text{vec}(X) \mid X \in S^p_+ \} \]
\[ = \{ x \in \mathbb{R}^{p(p+1)/2} \mid \text{mat}(x) \succeq 0 \} \]

\text{vec}(\cdot) \text{ converts symmetric matrix to vector; } \text{mat}(\cdot) \text{ is inverse operation}

\[
(x, y, z) \in S^2
\]
\[
\updownarrow
\]
\[
\begin{bmatrix}
  x & y/\sqrt{2} \\
  y/\sqrt{2} & z
\end{bmatrix} \succeq 0
\]
Semidefinite program

\begin{align*}
\text{minimize} \quad & c^T x \\
\text{subject to} \quad & x_1 A_{11} + x_2 A_{12} + \cdots + x_n A_{1n} \preceq B_1 \\
& \quad \quad \cdots \\
& x_1 A_{r1} + x_2 A_{r2} + \cdots + x_n A_{rn} \preceq B_r
\end{align*}

$r$ linear matrix inequalities of order $p_1, \ldots, p_r$

\textbf{Cone LP formulation:} express constraints as $Ax \preceq_K B$

$$K = S^{p_1} \times S^{p_2} \times \cdots \times S^{p_r}$$

$$A = \begin{bmatrix}
\text{vec}(A_{11}) & \text{vec}(A_{12}) & \cdots & \text{vec}(A_{1n}) \\
\text{vec}(A_{21}) & \text{vec}(A_{22}) & \cdots & \text{vec}(A_{2n}) \\
\vdots & \vdots & \ddots & \vdots \\
\text{vec}(A_{r1}) & \text{vec}(A_{r2}) & \cdots & \text{vec}(A_{rn})
\end{bmatrix}, \quad b = \begin{bmatrix}
\text{vec}(B_1) \\
\text{vec}(B_2) \\
\vdots \\
\text{vec}(B_r)
\end{bmatrix}$$
Semidefinite cone representable constraints

Matrix-fractional function

\[ y^T X^{-1} y \leq t, \quad X \succ 0, \quad y \in \text{range}(X) \]

\[ \Leftrightarrow \]

\[
\begin{bmatrix}
X & y \\
y^T & t
\end{bmatrix} \succeq 0
\]

Maximum eigenvalue of symmetric matrix

\[ \lambda_{\text{max}}(X) \leq t \quad \iff \quad X \preceq tI \]
Semidefinite cone representable constraints

**Maximum singular value** \( \|X\|_2 = \sigma_1(X) \)

\[
\|X\|_2 \leq t \iff \begin{bmatrix} tI & X \\ X^T & tI \end{bmatrix} \succeq 0
\]

**Trace norm (nuclear norm)** \( \|X\|_* = \sum_i \sigma_i(X) \)

\[
\|X\|_* \leq t \iff \exists U, V : \begin{bmatrix} U & X \\ X^T & V \end{bmatrix} \succeq 0, \quad \text{tr}U + \text{tr}V \leq 2t
\]
Exponential cone

**Definition:** $K_{\text{exp}}$ is the closure of

$$K = \left\{ (x, y, z) \in \mathbb{R}^3 \mid ye^{x/y} \leq z, \; y > 0 \right\}$$
Power cone

Definition: for $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m) > 0$, $\sum_{i=1}^{m} \alpha_i = 1$

$$K_\alpha = \{ (x, y) \in \mathbb{R}_+^m \times \mathbb{R} \mid |y| \leq x_{\alpha_1}^1 \cdots x_{\alpha_m}^m \}$$

Examples for $m = 2$

$\alpha = \left( \frac{1}{2}, \frac{1}{2} \right)$

$\alpha = \left( \frac{2}{3}, \frac{1}{3} \right)$

$\alpha = \left( \frac{3}{4}, \frac{1}{4} \right)$
Functions representable with exponential and power cone

Exponential cone

- exponential and logarithm
- entropy $f(x) = x \log x$

Power cone

- increasing power of absolute value: $f(x) = |x|^p$ with $p \geq 1$
- decreasing power: $f(x) = x^q$ with $q \leq 0$ and domain $\mathbb{R}_{++}$
- $p$-norm: $f(x) = \|x\|_p$ with $p \geq 1$
Outline

• convex sets and functions

• conic optimization

• duality
Lagrange dual

Convex problem (with linear constraints for simplicity)

\[
\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad Ax = b
\end{align*}
\]

Lagrangian and dual function

\[
L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T (Ax - b)
\]

\[
g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)
\]

(Lagrange) dual problem

\[
\begin{align*}
\text{maximize} & \quad g(\lambda, \nu) \\
\text{subject to} & \quad \lambda \geq 0
\end{align*}
\]

a convex optimization problem in \( \lambda, \nu \)
Duality theorem

let $p^*$ be the primal optimal value, $d^*$ the dual optimal value

Weak duality

$$p^* \geq d^*$$

without exception

Strong duality

$$p^* = d^*$$

if a constraint qualification holds (e.g., primal problem is strictly feasible)
the **conjugate** of a function $f$ is

$$f^*(y) = \sup_{x \in \text{dom } f} (y^T x - f(x))$$

**Properties**

- $f^*$ is convex (even if $f$ is not)
- if $f$ is (closed) convex, $\partial f^* = \partial f^{-1}$:

$$y \in \partial f(x) \iff x \in \partial f^*(y)$$
Examples

Convex quadratic function \((A \succ 0)\)

\[
f(x) = \frac{1}{2} x^T A x + b^T x \quad f^*(y) = \frac{1}{2} (y - b)^T A^{-1} (y - b)
\]

if \(A \succeq 0\), but not necessarily positive definite,

\[
f^*(y) = \begin{cases} \frac{1}{2} (y - b)^T A^\dagger (y - b) & y - b \in \text{range}(A) \\ +\infty & \text{otherwise} \end{cases}
\]

Negative entropy

\[
f(x) = \sum_{i=1}^{n} x_i \log x_i \quad f^*(y) = \sum_{i=1}^{n} e^{y_i} - 1
\]
Examples

**Norm**

\[ f(x) = \|x\| \quad f^*(y) = \begin{cases} 
0 & \|y\|_* \leq 1 \\
+\infty & \text{otherwise}
\end{cases} \]

The conjugate of norm is the indicator function of the unit ball for the dual norm

\[ \|y\|_* = \sup_{\|x\|\leq 1} y^T x \]

**Indicator function** (\( C \) convex)

\[ f(x) = I_C(x) = \begin{cases} 
0 & x \in C \\
+\infty & \text{otherwise}
\end{cases} \]

\[ f^*(y) = \sup_{x\in C} y^T x \]

The conjugate of the indicator function of \( C \) is the support function.
Duality and conjugate functions

Convex problem with composite structure

minimize $f(x) + g(Ax)$

$f$ and $g$ convex

Equivalent problem (auxiliary variable $y$)

minimize $f(x) + g(y)$
subject to $Ax = y$

Dual problem

maximize $-g^*(z) - f^*(-A^Tz)$
Example

Regularized norm approximation

\[
\text{minimize} \quad f(x) + \gamma \|Ax - b\|
\]

a special case with \( g(y) = \gamma \| y - b \| \),

\[
g^*(z) = \begin{cases} 
    b^T z & \|z\|_* \leq \gamma \\
    +\infty & \text{otherwise}
\end{cases}
\]

Dual problem

\[
\text{maximize} \quad -b^T z - f^*(-A^T z)
\]

\[
\text{subject to} \quad \|z\|_* \leq \gamma
\]
2. First-order methods

• (proximal) gradient method

• splitting and alternating minimization methods
the proximal operator (prox-operator) of a convex function $h$ is

$$
\text{prox}_h(x) = \arg\min_u \left( h(u) + \frac{1}{2}\|u - x\|^2 \right)
$$

- $h(x) = 0$: $\text{prox}_h(x) = x$

- $h(x) = I_C(x)$ (indicator function of $C$): $\text{prox}_h$ is projection on $C$

$$
\text{prox}_h(x) = \arg\min_{u \in C} \|u - x\|^2 = P_C(x)
$$

- $h(x) = \|x\|_1$: $\text{prox}_h$ is the ‘soft-threshold’ (shrinkage) operation

$$
\text{prox}_h(x)_i = \begin{cases} 
  x_i - 1 & x_i \geq 1 \\
  0 & |x_i| \leq 1 \\
  x_i + 1 & x_i \leq -1 
\end{cases}
$$
Proximal gradient method

minimize $f(x) = g(x) + h(x)$

- $g$ convex, differentiable, with $\text{dom } g = \mathbb{R}^n$
- $h$ convex, possibly nondifferentiable, with inexpensive prox-operator

Algorithm (update from $x = x^{(k-1)}$ to $x^+ = x^{(k)}$)

$$ x^+ = \text{prox}_{th} (x - t\nabla g(x)) $$

$$ = \arg\min_u \left( g(x) + \nabla g(x)^T (u - x) + \frac{t}{2} \| u - x \|_2^2 + h(x) \right) $$

$t > 0$ is step size, constant or determined by line search
Examples

Gradient method: $h(x) = 0$, i.e., minimize $g(x)$

$$x^+ = x - t\nabla g(x)$$

Gradient projection method: $h(x) = I_C(x)$, i.e., minimize $g(x)$ over $C$

$$x^+ = P_C(x - t\nabla g(x))$$
Iterative soft-thresholding: \( h(x) = \|x\|_1 \)

\[ x^+ = \text{prox}_{th}(x - t\nabla g(x)) \]

where

\[ \text{prox}_{th}(u)_i = \begin{cases} 
  u_i - t & u_i \geq t \\
  0 & -t \leq u_i \leq t \\
  u_i + t & u_i \leq -t 
\end{cases} \]
Properties of proximal operator

\[ \text{prox}_h(x) = \arg\min_u \left( h(u) + \frac{1}{2}\|u - x\|^2 \right) \]

assume \( h \) is closed and convex (i.e., convex with closed epigraph)

- \( \text{prox}_h(x) \) is uniquely defined for all \( x \)

- \( \text{prox}_h \) is nonexpansive

\[ \|\text{prox}_h(x) - \text{prox}_h(y)\|_2 \leq \|x - y\|_2 \]

- Moreau decomposition

\[ x = \text{prox}_h(x) + \text{prox}_{h^*}(x) \]

(surveys in Bauschke & Combettes 2011, Parikh & Boyd 2013)
Examples of inexpensive projections

• hyperplanes and halfspaces

• rectangles

\[ \{ x \mid l \leq x \leq u \} \]

• probability simplex

\[ \{ x \mid 1^T x = 1, x \geq 0 \} \]

• norm ball for many norms (Euclidean, 1-norm, . . . )

• nonnegative orthant, second-order cone, positive semidefinite cone
Examples of inexpensive prox-operators

Euclidean norm: \( h(x) = \|x\|_2 \)

\[
\text{prox}_{th}(x) = \left( 1 - \frac{t}{\|x\|_2} \right) x \quad \text{if} \quad \|x\|_2 \geq t, \quad \text{prox}_{th}(x) = 0 \quad \text{otherwise}
\]

Logarithmic barrier

\[
h(x) = -\sum_{i=1}^{n} \log x_i, \quad \text{prox}_{th}(x)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}, \quad i = 1, \ldots, n
\]

Euclidean distance: \( d(x) = \inf_{y \in C} \|x - y\|_2 \) (\( C \) closed convex)

\[
\text{prox}_{td}(x) = \theta P_C(x) + (1 - \theta)x, \quad \theta = \frac{t}{\max\{d(x), t\}}
\]

generalizes soft-thresholding operator
Prox-operator of conjugate

\[ \text{prox}_{th}(x) = x - t \text{prox}_{h^*/t}(x/t) \]

- follows from Moreau decomposition
- of interest when prox-operator of \( h^* \) is inexpensive

Example: norms

\[ h(x) = \| x \|, \quad h^*(y) = I_C(y) \]

where \( C \) is unit ball for dual norm \( \| \cdot \|_* \)

- \( \text{prox}_{h^*/t} \) is projection on \( C \)
- formula useful for prox-operator of \( \| \cdot \| \) if projection on \( C \) is inexpensive
many convex functions can be expressed as **support functions**

\[ h(x) = S_C(x) = \sup_{y \in C} x^T y \]

with \( C \) closed, convex

- conjugate is indicator function of \( C \): \( h^*(y) = I_C(y) \)
- hence, can compute \( \text{prox}_{th} \) via projection on \( C \)

**Example**: \( h(x) \) is sum of largest \( r \) components of \( x \)

\[ h(x) = x[1] + \cdots + x[r] = S_C(x), \quad C = \{y \mid 0 \leq y \leq 1, 1^T y = r\} \]
Convergence of proximal gradient method

\[
\text{minimize } \quad f(x) = g(x) + h(x)
\]

Assumptions

- \( \nabla g \) is Lipschitz continuous with constant \( L > 0 \)
  \[
  \| \nabla g(x) - \nabla g(y) \|_2 \leq L \| x - y \|_2 \quad \forall x, y
  \]

- optimal value \( f^* \) is finite and attained at \( x^* \) (not necessarily unique)

Result: with fixed step size \( t_k = 1/L \)

\[
f(x^{(k)}) - f^* \leq \frac{L}{2k} \| x^{(0)} - x^* \|_2^2
\]

- compare with \( 1/\sqrt{k} \): rate of subgradient method
- can be extended to include line searches
Fast (proximal) gradient methods

- Beck & Teboulle (2008): FISTA, a proximal gradient version of Nesterov’s 1983 method
- several recent variations and extensions

This lecture: FISTA (Fast Iterative Shrinkage-Thresholding Algorithm)
FISTA

minimize \( f(x) = g(x) + h(x) \)

- \( g \) convex differentiable with \( \text{dom} \, g = \mathbb{R}^n \)
- \( h \) convex with inexpensive prox-operator

Algorithm: choose any \( x^{(0)} = x^{(-1)} \); for \( k \geq 1 \), repeat the steps

\[
y = x^{(k-1)} + \frac{k - 2}{k + 1} (x^{(k-1)} - x^{(k-2)})
\]

\[
x^{(k)} = \text{prox}_{t_k h} (y - t_k \nabla g(y))
\]
Interpretation

- first two iterations \((k = 1, 2)\) are proximal gradient steps at \(x^{(k-1)}\)
- next iterations are proximal gradient steps at extrapolated points \(y\)

\[
x^{(k)} = \text{prox}_{t_k h} (y - t_k \nabla g(y))
\]

sequence \(x^{(k)}\) remains feasible (in \(\text{dom } h\)); \(y\) may be outside \(\text{dom } h\)
Convergence of FISTA

\[
\text{minimize } f(x) = g(x) + h(x)
\]

Assumptions

- \( \text{dom } g = \mathbb{R}^n \) and \( \nabla g \) is Lipschitz continuous with constant \( L > 0 \)
- \( h \) is closed (implies \( \text{prox}_{th}(u) \) exists and is unique for all \( u \))
- optimal value \( f^* \) is finite and attained at \( x^* \) (not necessarily unique)

Result: with fixed step size \( t_k = 1/L \)

\[
f(x^{(k)}) - f^* \leq \frac{2L}{(k + 1)^2} ||x^{(0)} - f^*||_2^2
\]

- compare with \( 1/k \) convergence rate for proximal gradient method
- can be extended to include line searches
Example

\[
\text{minimize } \log \sum_{i=1}^{m} \exp(a_i^T x + b_i)
\]

randomly generated data with \( m = 2000, n = 1000 \), same fixed step size

FISTA is not a descent method
Proximal point algorithm

to minimize $h(x)$, apply fixed-point iteration to $\text{prox}_h(x)$

$$x^+ = \text{prox}_h(x)$$

- proximal gradient method with zero $g$
- implementable if inexact prox-evaluations are used

Convergence

- $O(1/\epsilon)$ iterations to reach $h(x) - h(x^*) \leq \epsilon$ (rate $1/k$)
- $O(1/\sqrt{\epsilon})$ iterations with accelerated $(1/k^2)$ algorithm (Güler 1992)
Smoothing interpretation

Moreau-Yosida regularization of $h$

$$h_t(x) = \inf_u \left( h(u) + \frac{1}{2t} \| u - x \|_2^2 \right)$$

- convex, with full domain
- differentiable with $1/t$-Lipschitz continuous gradient

$$\nabla h_t(x) = \frac{1}{t} (x - \text{prox}_{th}(x)) = \text{prox}_{h^*/t}(x/t)$$

Proximal point algorithm (with constant $t$): gradient method for $h_t$

$$x^+ = \text{prox}_{th}(x) = x - t \nabla h_t(x)$$
**Examples**

**Indicator function** (of closed convex set $C$): squared Euclidean distance

$$h(x) = I_C(x), \quad h_t(x) = \frac{1}{2t} \text{dist}(x)^2$$

**1-Norm: Huber penalty**

$$h(x) = \|x\|_1, \quad h_t(x) = \sum_{k=1}^{n} \phi_t(x_k)$$

\[
\phi_t(z) = \begin{cases} 
  \frac{z^2}{2t} & |z| \leq t \\
  |z| - t/2 & |z| \geq t
\end{cases}
\]
Monotone operator

Monotone (set-valued) operator. $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with

$$(y - \hat{y})^T (x - \hat{x}) \geq 0 \quad \forall x, \hat{x}, y \in F(x), \hat{y} \in F(\hat{x})$$

Examples

- subdifferential $F(x) = \partial f(x)$ of closed convex function
- linear function $F(x) = Bx$ with $B + B^T$ positive semidefinite
Proximal point algorithm for monotone inclusion

to solve $0 \in F(x)$, run fixed-point iteration

$$x^+ = (I + tF)^{-1}(x)$$

the mapping $(I + tF)^{-1}$ is called the **resolvent** of $F$

- $x = (I + tF)^{-1}(\hat{x})$ is (unique) solution of $\hat{x} \in x + tF(x)$
- resolvent of subdifferential $F(x) = \partial h(x)$ is prox-operator:

  $$(I + t\partial h)^{-1}(x) = \text{prox}_{th}(x)$$

- converges if $F$ has a zero and is maximal monotone
Outline

• (proximal) gradient method

• splitting and alternating minimization methods
Convex optimization with composite structure

Primal and dual problems

\[
\begin{align*}
\text{minimize} & \quad f(x) + g(Ax) \\
\text{maximize} & \quad -g^*(z) - f^*(-A^T z)
\end{align*}
\]

\(f\) and \(g\) are ‘simple’ convex functions, with conjugates \(f^*, g^*\)

Optimality conditions

- primal: \(0 \in \partial f(x) + A^T \partial g(Ax)\)
- dual: \(0 \in \partial g^*(z) - A \partial f^*(-A^T z)\)
- primal-dual:

\[
0 \in \begin{bmatrix}
0 & A^T \\
-A & 0
\end{bmatrix}
\begin{bmatrix}
x \\
z
\end{bmatrix}
+ \begin{bmatrix}
\partial f(x) \\
\partial g^*(z)
\end{bmatrix}
\]

First-order methods
Examples

Equality constraints: \( g = I_\{b\}, \) indicator of \( \{b\} \)

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad Ax = b
\end{align*}
\]

maximize \(-b^T z - f^*(-A^T z)\)

Set constraint: \( g = I_C, \) indicator of convex \( C, \) with support function \( S_C \)

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad Ax \in C
\end{align*}
\]

maximize \(-S_C(z) - f^*(-A^T z)\)

Regularized norm approximation: \( g(y) = \gamma \|y - b\| \)

\[
\begin{align*}
\text{minimize} & \quad f(x) + \|Ax - b\| \\
\text{subject to} & \quad \|z\|_* \leq 1
\end{align*}
\]

maximize \(-b^T z - f^*(-A^T z)\)

First-order methods
Augmented Lagrangian method

the proximal-point algorithm applied to the dual

\[
\text{maximize } -g^*(z) - f^*(-A^Tz)
\]

1. minimize augmented Lagrangian

\[
(x^+, y^+) = \arg\min_{\tilde{x}, \tilde{y}} \left( f(\tilde{x}) + g(\tilde{y}) + \frac{t}{2} \| A\tilde{x} - \tilde{y} + z/t \|^2_2 \right)
\]

2. dual update: \( z^+ = z + t(Ax^+ - y^+) \)

- equivalent to gradient method applied to Moreau-Yosida smoothed dual
- also known as Bregman iteration (Yin et al. 2008)
- practical if inexact minimization is used in step 1
Proximal method of multipliers

apply proximal point algorithm to primal-dual optimality condition

\[ 0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix} \]

Algorithm (Rockafellar 1976)

1. minimize generalized augmented Lagrangian

\[(x^+, y^+) = \arg\min_{\tilde{x}, \tilde{y}} \left(f(\tilde{x}) + g(\tilde{y}) + \frac{t}{2} \| A\tilde{x} - \tilde{y} + z/t \|_2^2 + \frac{1}{2t} \| \tilde{x} - x \|_2^2 \right)\]

2. dual update: \( z^+ = z + t(Ax^+ - y^+) \)
Douglas-Rachford splitting algorithm

\[ 0 \in F(x) = F_1(x) + F_2(x) \]

with \( F_1 \) and \( F_2 \) maximal monotone operators

**Algorithm** (Lions and Mercier 1979, Eckstein and Bertsekas 1992)

\[
\begin{align*}
x^+ &= (I + tF_1)^{-1}(z) \\
y^+ &= (I + tF_2)^{-1}(2x^+ - z) \\
z^+ &= z + y^+ - x^+
\end{align*}
\]

- useful when resolvents of \( F_1 \) and \( F_2 \) are inexpensive, but not \( (I + tF)^{-1} \)
- under weak conditions (existence of solution), \( x \) converges to solution
**Alternating direction method of multipliers (ADMM)**

Douglas-Rachford splitting applied to optimality condition for dual

\[
\text{maximize } - g^*(z) - f^*(-A^Tz)
\]

1. alternating minimization of augmented Lagrangian

\[
x^+ = \arg\min_{\tilde{x}} \left( f(\tilde{x}) + \frac{t}{2} \| A\tilde{x} - y + z/t \|_2^2 \right)
\]

\[
y^+ = \arg\min_{\tilde{y}} \left( g(\tilde{y}) + \frac{t}{2} \| Ax^+ - \tilde{y} + z/t \|_2^2 \right)
\]

\[
= \text{prox}_{g/t}(Ax^+ + z/t)
\]

2. dual update \( z^+ = z + t(Ax^+ - y) \)

also known as split Bregman method (Goldstein and Osher 2009)

(recent survey in Boyd, Parikh, Chu, Peleato, Eckstein 2011)
Primal application of Douglas-Rachford method

D-R splitting algorithm applied to optimality condition for primal problem

\[
\begin{align*}
\text{minimize} & \quad f(x) + g(y) & \rightarrow \text{minimize} & \quad f(x) + g(y) + I_{\{0\}}(Ax - y) \\
\text{subject to} & \quad Ax = y & \quad h_1(x, y) & \quad h_2(x, y)
\end{align*}
\]

Main steps

• prox-operator of \( h_1 \): separate evaluations of \( \text{prox}_f \) and \( \text{prox}_g \)
• prox-operator of \( h_2 \): projection on subspace \( H = \{ (x, y) \mid Ax = y \} \)

\[
P_H(x, y) = \begin{bmatrix} I \\ A \end{bmatrix} (I + A^T A)^{-1}(x + A^T y)
\]

also known as \textit{method of partial inverses} (Spingarn 1983, 1985)
Primal-dual application

\[ 0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix} \]

\[ F_2(x,z) \]
\[ F_1(x,z) \]

Main steps

- resolvent of \( F_1 \): prox-operator of \( f, g \)

- resolvent of \( F_2 \):

\[
\begin{bmatrix} I & tA^T \\ -tA & I \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} I \\ tA \end{bmatrix} (I + t^2 A^T A)^{-1} \begin{bmatrix} I \\ -tA \end{bmatrix}^T
\]
Summary: Douglas-Rachford splitting methods

minimize $f(x) + g(Ax)$

Most expensive steps

- **Dual** (ADMM)

  minimize (over $x$) $f(x) + \frac{t}{2} \|Ax - y + z/t\|_2^2$

  if $f$ is quadratic, a linear equation with coefficient $\nabla^2 f(x) + tA^TA$

- **Primal** (Spingarn): equation with coefficient $I + A^TA$

- **Primal-dual**: equation with coefficient $I + t^2 A^TA$
Forward-backward method

\[ 0 \in F(x) = F_1(x) + F_2(x) \]

with \( F_1 \) and \( F_2 \) maximal monotone operators, \( F_1 \) single-valued

**Forward-backward iteration** (for single-valued \( F_1 \))

\[ x^+ = (I + tF_2)^{-1}(I - tF_1(x)) \]

- converges if \( F_1 \) is co-coercive with parameter \( L \) and \( t \in (0, 1/L] \)

\[ (F_1(x) - F_1(\hat{x}))^T(x - \hat{x}) \geq \frac{1}{L} \|F_1(x) - F_1(\hat{x})\|_2^2 \quad \forall x, \hat{x} \]

this is Lipschitz continuity if \( F_1 = \partial f_1 \), a stronger condition otherwise

- Tseng’s modified method (1991) only requires Lipschitz continuous \( F_1 \)

First-order methods
Dual proximal gradient method

\[ 0 \in \partial g^*(z) = F_2(z) - A \nabla f^*(-A^T z) = F_1(z) \]

Proximal gradient iteration

\[ x = \arg\min_{\tilde{x}} \left( f(\tilde{x}) + z^T A \tilde{x} \right) = \nabla f^*(-A^T z) \]

\[ z^+ = \text{prox}_{t g^*}(z + t A x) \]

- does not involve solution of linear equation
- first step is minimization of (unaugmented) Lagrangian
- requires Lipschitz continuous \( \nabla f^* \) (strongly convex \( f \))
- accelerated methods: FISTA, Nesterov’s methods
Primal-dual (Chambolle-Pock) method

\[ 0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix} \]

Algorithm (with parameter \( \theta \in [0, 1] \)) (Chambolle & Pock 2011)

\[
\begin{align*}
    z^+ &= \text{prox}_{tg^*}(z + tA\bar{x}) \\
    x^+ &= \text{prox}_{tf}(x - tA^Tz^+) \\
    \bar{x}^+ &= x^+ + \theta(x^+ - x)
\end{align*}
\]

- widely used in image processing
- step size fixed (\( t \leq 1/\|A\|_2 \)) or adapted by line search
- can be interpreted as pre-conditioned proximal-point algorithm
Summary: Splitting algorithms

\[
\text{minimize } f(x) + g(Ax)
\]

**Douglas-Rachford splitting**

- can be applied to primal (Spingarn’s method), dual (ADMM), primal-dual optimality conditions
- subproblems include quadratic term \(\|Ax\|^2_2\) in cost function

**Forward-backward splitting**

- (accelerated) proximal gradient algorithm applied to dual problem
- Tseng’s FB algorithm applied to primal-dual optimality conditions, semi-implicit primal-dual method (Chambolle-Pock), . . .
- only require application of \(A\) and \(A^T\)

**Extensions:** linearized splitting methods, generalized distances, . . .