Convex Optimization

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Modification of http://stanford.edu/~boyd/cvxbook/bv_cvxslides.pdf

Outline

- **Introduction**
- Convex Sets & Functions
- Convex Optimization Problems
- Duality
- Convex Optimization Methods
- Summary

Mathematical Optimization

Optimization Problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le b_i, \quad i = 1, \dots, m \end{array}$$

- $x = (x_1, \dots, x_n)$: optimization variables
- $f_0 : \mathbf{R}^n \to \mathbf{R}$: objective function
- $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \dots, m$: constraint functions

optimal solution $x^\star$ has smallest value of f_0 among all vectors that satisfy the constraints

Applications

- Dimensionality Reduction (PCA)

$$\max_{\mathbf{w}\in\mathbb{R}^d} \quad \mathbf{w}^\top C\mathbf{w}$$
$$\text{s.t.} \quad \|\mathbf{w}\|_2^2 = 1$$

- Clustering (NMF)

$$\min_{U\in\mathbb{R}^{d\times k}, V\in\mathbb{R}^{n\times k}} \quad \left\|X - UV^\top\right\|_F^2$$
$$\text{s.t.} \quad U \geq 0, V \geq 0$$

- Classification (SVM)

$$\min_{\overline{W}\in\mathbb{R}^d, b\in\mathbb{R}} \quad O = \frac{||\overline{W}||^2}{2} + C\sum_{i=1}^{n} \max\{0, 1 - y_i[\overline{W}\cdot\overline{X_i} + b]\}.$$

Least-squares

The Problem

$$\text{minimize} \quad f_0(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{k} (a_i^T x - b_i)^2$$

- Given $\alpha_i \in \mathbb{R}^d$, predict $b_i \in \mathbb{R}$ by $a_i^\top x$

Properties

- analytical solution: $x^\star = (A^T A)^{-1} A^T b$
- reliable and efficient algorithms and software
- computation time proportional to $n^2 k$ ($A \in \mathbf{R}^{k \times n}$); less if structured
- a mature technology

Linear Programming

The Problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x \leq b_i, \quad i = 1, \ldots, m \end{array}$$

Here the vectors $c, a_1, \ldots, a_m \in \mathbf{R}^n$ and scalars $b_1, \ldots, b_m \in \mathbf{R}$ are problem parameters that specify the objective and constraint functions.

Properties

- no analytical formula for solution
- reliable and efficient algorithms and software
- computation time proportional to $n^2 m$ if $m \geq n$; less with structure
- a mature technology

Convex Optimization Problem

The Problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m \end{array}$$

Conditions

- objective and constraint functions are convex:

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

  if $\alpha + \beta = 1$, $\alpha \geq 0$, $\beta \geq 0$

- includes least-squares problems and linear programs as special cases

Convex Optimization Problem

The Problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m \end{array}$$

Properties

- no analytical solution
- reliable and efficient algorithms
- computation time (roughly) proportional to $\max\{n^3, n^2m, F\}$, where F is cost of evaluating f_i's and their first and second derivatives
- almost a technology

Nonlinear Optimization

- Definition
  - The objective or constraint functions are not linear
  - Could be convex or nonconvex

local optimization methods (nonlinear programming)

- find a point that minimizes f_0 among feasible points near it
- fast, can handle large problems
- require initial guess
- provide no information about distance to (global) optimum

global optimization methods

- find the (global) solution
- worst-case complexity grows exponentially with problem size

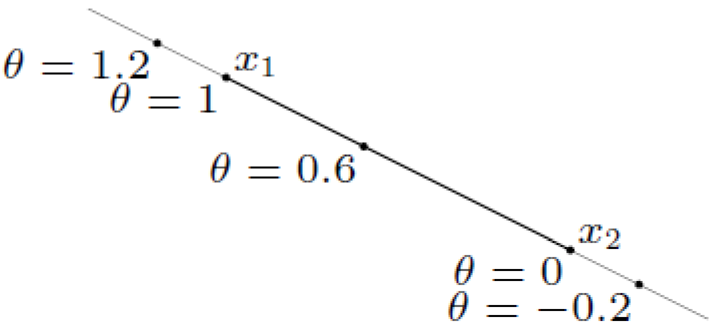
Outline

- Introduction
- **Convex Sets & Functions**
- Convex Optimization Problems
- Duality
- Convex Optimization Methods
- Summary

Affine Set

line through x_1, x_2: all points

$$x = \theta x_1 + (1 - \theta) x_2 \qquad (\theta \in \mathbf{R})$$



affine set: contains the line through any two distinct points in the set

example: solution set of linear equations $\{x \mid Ax = b\}$

(conversely, every affine set can be expressed as solution set of system of linear equations)

Convex Set

line segment between x_1 and x_2: all points

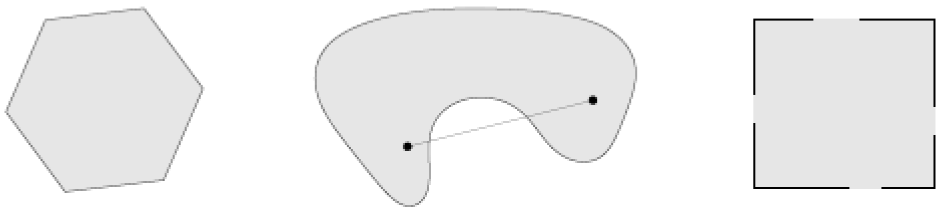
$$x = \theta x_1 + (1 - \theta) x_2$$

with $0 \leq \theta \leq 1$

convex set: contains line segment between any two points in the set

$$x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Longrightarrow \quad \theta x_1 + (1 - \theta) x_2 \in C$$

examples (one convex, two nonconvex sets)

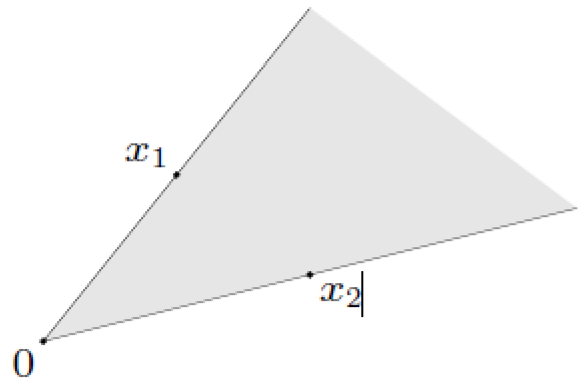


Convex Cone

conic (nonnegative) combination of x_1 and x_2: any point of the form

$$x = \theta_1 x_1 + \theta_2 x_2$$

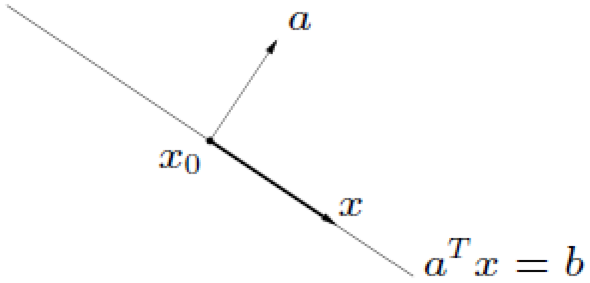
with $\theta_1 \geq 0$, $\theta_2 \geq 0$



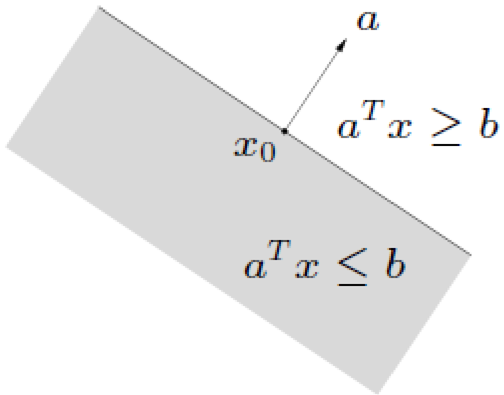
convex cone: set that contains all conic combinations of points in the set

Some Examples (1)

hyperplane: set of the form $\{x \mid a^T x = b\}$ $(a \neq 0)$



halfspace: set of the form $\{x \mid a^T x \leq b\}$ $(a \neq 0)$



- a is the normal vector
- hyperplanes are affine and convex; halfspaces are convex

Some Examples (2)

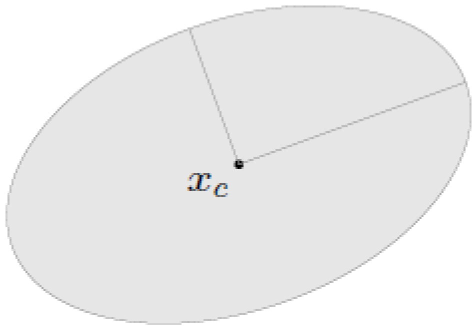
(Euclidean) ball with center x_c and radius r:

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \le r\} = \{x_c + ru \mid \|u\|_2 \le 1\}$$

ellipsoid: set of the form

$$\{x \mid (x - x_c)^T P^{-1}(x - x_c) \le 1\}$$

with $P \in \mathbf{S}^n_{++}$ ($i.e.$, P symmetric positive definite)



other representation: $\{x_c + Au \mid \|u\|_2 \le 1\}$ with A square and nonsingular

Some Examples (3)

norm: a function $\|\cdot\|$ that satisfies

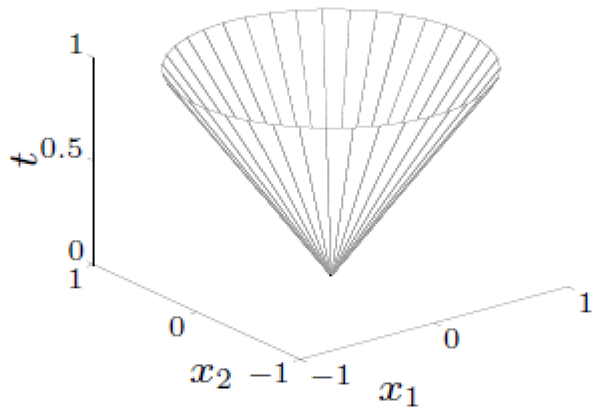
- $\|x\| \geq 0$; $\|x\| = 0$ if and only if $x = 0$
- $\|tx\| = |t| \, \|x\|$ for $t \in \mathbf{R}$
- $\|x + y\| \leq \|x\| + \|y\|$

notation: $\|\cdot\|$ is general (unspecified) norm; $\|\cdot\|_{\text{symb}}$ is particular norm

norm ball with center x_c and radius r: $\{x \mid \|x - x_c\| \leq r\}$

norm cone: $\{(x, t) \mid \|x\| \leq t\}$

Euclidean norm cone is called second-order cone



norm balls and cones are convex

Operations that Preserve Convexity

practical methods for establishing convexity of a set C

1. apply definition

$$x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Longrightarrow \quad \theta x_1 + (1-\theta)x_2 \in C$$

2. show that C is obtained from simple convex sets (hyperplanes, halfspaces, norm balls, . . .) by operations that preserve convexity

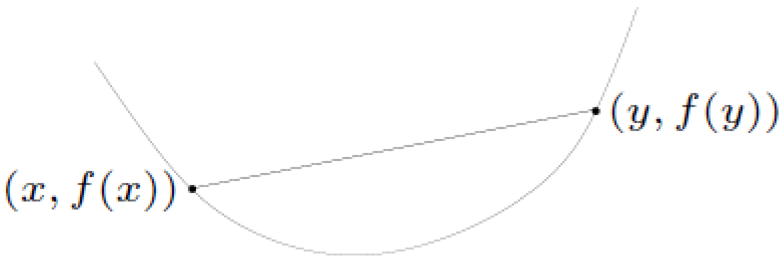
   - intersection
   - affine functions
   - perspective function
   - linear-fractional functions

Convex Functions

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if $\mathbf{dom}\, f$ is a convex set and

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

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



- f is concave if $-f$ is convex
- f is strictly convex if $\mathbf{dom}\, f$ is convex and

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

for $x, y \in \mathbf{dom}\, f$, $x \neq y$, $0 < \theta < 1$

Examples on $\mathbb{R}$

convex:

- affine: $ax + b$ on $\mathbf{R}$, for any $a, b \in \mathbf{R}$
- exponential: e^{ax}, for any $a \in \mathbf{R}$
- powers: x^α on $\mathbf{R}_{++}$, for $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute value: $|x|^p$ on $\mathbf{R}$, for $p \geq 1$
- negative entropy: $x \log x$ on $\mathbf{R}_{++}$

concave:

- affine: $ax + b$ on $\mathbf{R}$, for any $a, b \in \mathbf{R}$
- powers: x^α on $\mathbf{R}_{++}$, for $0 \leq \alpha \leq 1$
- logarithm: $\log x$ on $\mathbf{R}_{++}$

Examples on $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$

affine functions are convex and concave; all norms are convex

examples on R^n

- affine function $f(x) = a^T x + b$
- norms: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p \geq 1$; $\|x\|_\infty = \max_k |x_k|$

examples on R$^{m\times n}$ ($m \times n$ matrices)

- affine function

$$f(X) = \mathbf{tr}(A^T X) + b = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} X_{ij} + b$$

- spectral (maximum singular value) norm

$$f(X) = \|X\|_2 = \sigma_{\max}(X) = (\lambda_{\max}(X^T X))^{1/2}$$

Restriction of a Convex Function to a Line

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if and only if the function $g : \mathbf{R} \to \mathbf{R}$,

$$g(t) = f(x + tv), \qquad \mathbf{dom}\, g = \{t \mid x + tv \in \mathbf{dom}\, f\}$$

is convex (in t) for any $x \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$

can check convexity of f by checking convexity of functions of one variable

example. $f : \mathbf{S}^n \to \mathbf{R}$ with $f(X) = \log\det X$, $\mathbf{dom}\, f = \mathbf{S}^n_{++}$

$$\begin{aligned} g(t) = \log\det(X + tV) &= \log\det X + \log\det(I + tX^{-1/2}VX^{-1/2}) \\ &= \log\det X + \sum_{i=1}^{n} \log(1 + t\lambda_i) \end{aligned}$$

where λ_i are the eigenvalues of $X^{-1/2}VX^{-1/2}$

g is concave in t (for any choice of $X \succ 0$, V); hence f is concave

First-order Conditions

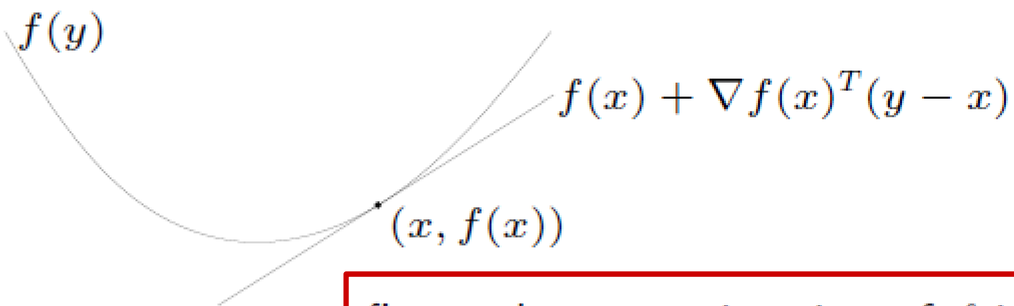
f is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$$

exists at each $x \in \mathbf{dom}\, f$

1st-order condition: differentiable f with convex domain is convex iff

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



first-order approximation of f is global underestimator

Second-order Conditions

f is **twice differentiable** if $\mathbf{dom}\, f$ is open and the Hessian $\nabla^2 f(x) \in \mathbf{S}^n$,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i,j = 1,\ldots,n,$$

exists at each $x \in \mathbf{dom}\, f$

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

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

- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

Examples

quadratic function: $f(x) = (1/2)x^T Px + q^T x + r$ (with $P \in \mathbf{S}^n$)

$$\nabla f(x) = Px + q, \qquad \nabla^2 f(x) = P$$

convex if $P \succeq 0$

least-squares objective: $f(x) = \|Ax - b\|_2^2$

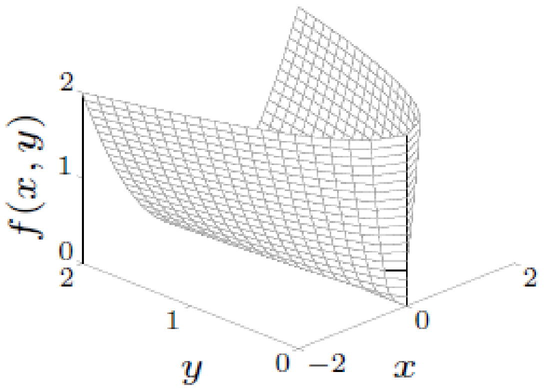
$$\nabla f(x) = 2A^T(Ax - b), \qquad \nabla^2 f(x) = 2A^T A$$

convex (for any A)

quadratic-over-linear: $f(x, y) = x^2/y$

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$

convex for $y > 0$



Operations that Preserve Convexity

practical methods for establishing convexity of a function

1. verify definition (often simplified by restricting to a line)

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
   - composition
   - minimization
   - perspective

Positive Weighted Sum & Composition with Affine Function

nonnegative multiple: αf 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

- log barrier for linear inequalities

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x), \qquad \mathbf{dom}\, f = \{x \mid a_i^T x < b_i, i = 1, \ldots, m\}$$

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

Pointwise Maximum

if $f_1, \ldots, f_m$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex

examples

- piecewise-linear function: $f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$ is convex
- sum of r largest components of $x \in \mathbf{R}^n$:

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

  is convex ($x_{[i]}$ is ith largest component of x)

  proof:

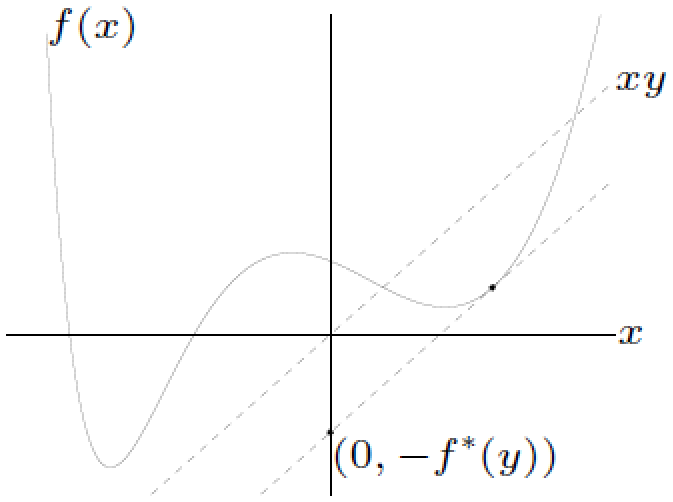
$$f(x) = \max\{x_{i_1} + x_{i_2} + \cdots + x_{i_r} \mid 1 \leq i_1 < i_2 < \cdots < i_r \leq n\}$$

Hinge loss: $\ell(w) = \max(0, 1 - y_i x_i^\top w)$

The Conjugate Function

the **conjugate** of a function f is

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



- f^* is convex (even if f is not)
- will be useful in chapter 5

Examples

- negative logarithm $f(x) = -\log x$

$$
\begin{aligned}
f^*(y) &= \sup_{x>0}(xy + \log x) \\
&= \begin{cases} -1 - \log(-y) & y < 0 \\ \infty & \text{otherwise} \end{cases}
\end{aligned}
$$

- strictly convex quadratic $f(x) = (1/2)x^T Q x$ with $Q \in \mathbf{S}^n_{++}$

$$
\begin{aligned}
f^*(y) &= \sup_{x}(y^T x - (1/2)x^T Q x) \\
&= \frac{1}{2} y^T Q^{-1} y
\end{aligned}
$$

Outline

- Introduction
- Convex Sets & Functions
- **Convex Optimization Problems**
- Duality
- Convex Optimization Methods
- Summary

Optimization Problem in Standard Form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

- $x \in \mathbf{R}^n$ is the optimization variable
- $f_0 : \mathbf{R}^n \to \mathbf{R}$ is the objective or cost function
- $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \ldots, m$, are the inequality constraint functions
- $h_i : \mathbf{R}^n \to \mathbf{R}$ are the equality constraint functions

optimal value:

$$p^\star = \inf\{f_0(x) \mid f_i(x) \leq 0, \ i = 1, \ldots, m, \ h_i(x) = 0, \ i = 1, \ldots, p\}$$

- $p^\star = \infty$ if problem is infeasible (no x satisfies the constraints)
- $p^\star = -\infty$ if problem is unbounded below

Optimal and Locally Optimal Points

x is **feasible** if $x \in \mathbf{dom}\, f_0$ and it satisfies the constraints

a feasible x is **optimal** if $f_0(x) = p^\star$; $X_{\rm opt}$ is the set of optimal points

x is **locally optimal** if there is an $R > 0$ such that x is optimal for

$$\begin{array}{ll} \text{minimize (over } z) & f_0(z) \\ \text{subject to} & f_i(z) \le 0, \quad i = 1, \dots, m, \quad h_i(z) = 0, \quad i = 1, \dots, p \\ & \|z - x\|_2 \le R \end{array}$$

examples (with $n = 1$, $m = p = 0$)

- $f_0(x) = 1/x$, $\mathbf{dom}\, f_0 = \mathbf{R}_{++}$: $p^\star = 0$, no optimal point
- $f_0(x) = -\log x$, $\mathbf{dom}\, f_0 = \mathbf{R}_{++}$: $p^\star = -\infty$
- $f_0(x) = x \log x$, $\mathbf{dom}\, f_0 = \mathbf{R}_{++}$: $p^\star = -1/e$, $x = 1/e$ is optimal
- $f_0(x) = x^3 - 3x$, $p^\star = -\infty$, local optimum at $x = 1$

Implicit Constraints

the standard form optimization problem has an **implicit constraint**

$$x \in \mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \;\cap\; \bigcap_{i=1}^{p} \mathbf{dom}\, h_i,$$

- we call $\mathcal{D}$ the **domain** of the problem
- the constraints $f_i(x) \leq 0$, $h_i(x) = 0$ are the explicit constraints
- a problem is **unconstrained** if it has no explicit constraints ($m = p = 0$)

example:

$$\text{minimize} \quad f_0(x) = -\sum_{i=1}^{k} \log(b_i - a_i^T x)$$

is an unconstrained problem with implicit constraints $a_i^T x < b_i$

Convex Optimization Problem

standard form convex optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \dots, m \\ & a_i^T x = b_i, \quad i = 1, \dots, p \end{array}$$

- $f_0, f_1, \dots, f_m$ are convex; equality constraints are affine
- problem is *quasiconvex* if f_0 is quasiconvex (and $f_1, \dots, f_m$ convex)

often written as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \dots, m \\ & Ax = b \end{array}$$

important property: feasible set of a convex optimization problem is convex

Example

$$\begin{array}{ll} \text{minimize} & f_0(x) = x_1^2 + x_2^2 \\ \text{subject to} & f_1(x) = x_1/(1+x_2^2) \le 0 \\ & h_1(x) = (x_1+x_2)^2 = 0 \end{array}$$

- f_0 is convex; feasible set $\{(x_1, x_2) \mid x_1 = -x_2 \le 0\}$ is convex
- not a convex problem (according to our definition): f_1 is not convex, h_1 is not affine
- equivalent (but not identical) to the convex problem

$$\begin{array}{ll} \text{minimize} & x_1^2 + x_2^2 \\ \text{subject to} & x_1 \le 0 \\ & x_1 + x_2 = 0 \end{array}$$

Local and Global Optima

any locally optimal point of a convex problem is (globally) optimal

proof: suppose x is locally optimal, but there exists a feasible y with $f_0(y) < f_0(x)$

x locally optimal means there is an $R > 0$ such that

$$z \text{ feasible}, \quad \|z - x\|_2 \le R \quad \Longrightarrow \quad f_0(z) \ge f_0(x)$$

consider $z = \theta y + (1-\theta)x$ with $\theta = R/(2\|y - x\|_2)$

- $\|y - x\|_2 > R$, so $0 < \theta < 1/2$
- z is a convex combination of two feasible points, hence also feasible
- $\|z - x\|_2 = R/2$ and

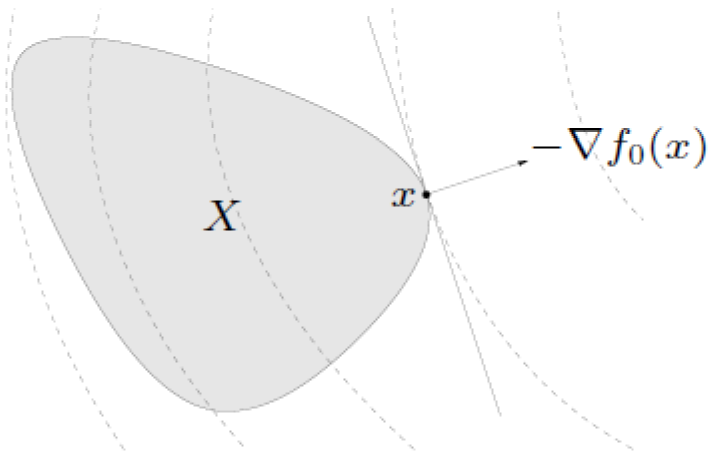
$$f_0(z) \le \theta f_0(y) + (1-\theta) f_0(x) < f_0(x)$$

which contradicts our assumption that x is locally optimal

Optimality Criterion for Differentiable f_0

x is optimal if and only if it is feasible and

$$\nabla f_0(x)^T (y - x) \geq 0 \quad \text{for all feasible } y$$



if nonzero, $\nabla f_0(x)$ defines a supporting hyperplane to feasible set X at x

Examples

- **unconstrained problem**: x is optimal if and only if

$$x \in \textbf{dom}\, f_0, \qquad \nabla f_0(x) = 0$$

- **equality constrained problem**

$$\text{minimize} \quad f_0(x) \quad \text{subject to} \quad Ax = b$$

x is optimal if and only if there exists a ν such that

$$x \in \textbf{dom}\, f_0, \qquad Ax = b, \qquad \nabla f_0(x) + A^T \nu = 0$$

- **minimization over nonnegative orthant**

$$\text{minimize} \quad f_0(x) \quad \text{subject to} \quad x \succeq 0$$

x is optimal if and only if

$$x \in \textbf{dom}\, f_0, \qquad x \succeq 0, \qquad \begin{cases} \nabla f_0(x)_i \geq 0 & x_i = 0 \\ \nabla f_0(x)_i = 0 & x_i > 0 \end{cases}$$

Popular Convex Problems

- Linear Program (LP)
- Linear-fractional Program
- Quadratic Program (QP)
- Quadratically Constrained Quadratic program (QCQP)
- Second-order Cone Programming (SOCP)
- Geometric Programming (GP)
- Semidefinite Program (SDP)

Outline

- Introduction
- Convex Sets & Functions
- Convex Optimization Problems
- **Duality**
- Convex Optimization Methods
- Summary

Lagrangian

standard form problem (not necessarily convex)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value $p^\star$

Lagrangian

standard form problem (not necessarily convex)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p \end{array}$$

variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value $p^\star$

Lagrangian: $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with $\mathbf{dom}\, L = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$,

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(x) \leq 0$
- ν_i is Lagrange multiplier associated with $h_i(x) = 0$

Lagrange Dual Function

Lagrange dual function: $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$
\begin{aligned}
g(\lambda,\nu) &= \inf_{x\in\mathcal{D}} L(x,\lambda,\nu) \\
&= \inf_{x\in\mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\end{aligned}
$$

g is concave, can be $-\infty$ for some λ, ν

Lagrange Dual Function

Lagrange dual function: $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \end{aligned}$$

g is concave, can be $-\infty$ for some λ, ν

lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$

Least-norm Solution of Linear Equations

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b \end{array}$$

dual function

- Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$
- to minimize L over x, set gradient equal to zero:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \quad \Longrightarrow \quad x = -(1/2)A^T \nu$$

- plug in in L to obtain g:

$$g(\nu) = L((-1/2)A^T \nu, \nu) = -\frac{1}{4}\nu^T A A^T \nu - b^T \nu$$

  a concave function of ν

lower bound property: $p^\star \geq -(1/4)\nu^T A A^T \nu - b^T \nu$ for all ν

Lagrange Dual and Conjugate Function

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \preceq b, \quad Cx = d \end{array}$$

dual function

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathbf{dom}\, f_0} \left(f_0(x) + (A^T\lambda + C^T\nu)^T x - b^T\lambda - d^T\nu \right) \\ &= -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu \end{aligned}$$

- recall definition of conjugate $f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$
- simplifies derivation of dual if conjugate of f_0 is known

example: entropy maximization

$$f_0(x) = \sum_{i=1}^n x_i \log x_i, \qquad f_0^*(y) = \sum_{i=1}^n e^{y_i - 1}$$

The Dual Problem

Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on $p^\star$, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted $d^\star$
- λ, ν are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \mathbf{dom}\, g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \mathbf{dom}\, g$ explicit

example: standard form LP and its dual (page 5–5)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0 \end{array} \qquad\qquad \begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \succeq 0 \end{array}$$

Weak and Strong Duality

weak duality: $d^\star \leq p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

  for example, solving the SDP

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array}$$

  gives a lower bound for the two-way partitioning problem on page 5–7

Weak and Strong Duality

weak duality: $d^\star \leq p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

  for example, solving the SDP

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array}$$

  gives a lower bound for the two-way partitioning problem on page 5–7

strong duality: $d^\star = p^\star$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

Slater's Constraint Qualification

strong duality holds for a convex problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

if it is strictly feasible, *i.e.*,

$$\exists x \in \mathbf{int}\, \mathcal{D}: \qquad f_i(x) < 0, \quad i = 1, \ldots, m, \qquad Ax = b$$

- also guarantees that the dual optimum is attained (if $p^\star > -\infty$)
- can be sharpened: *e.g.*, can replace $\mathbf{int}\, \mathcal{D}$ with $\mathbf{relint}\, \mathcal{D}$ (interior relative to affine hull); linear inequalities do not need to hold with strict inequality, . . .
- there exist many other types of constraint qualifications

Complementary Slackness

assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal

$$\begin{aligned} f_0(x^\star) = g(\lambda^\star, \nu^\star) &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\ &\le f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star) \\ &\le f_0(x^\star) \end{aligned}$$

hence, the two inequalities hold with equality

- $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (known as complementary slackness):

$$\lambda_i^\star > 0 \Longrightarrow f_i(x^\star) = 0, \qquad f_i(x^\star) < 0 \Longrightarrow \lambda_i^\star = 0$$

Karush-Kuhn-Tucker (KKT) Conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

1. primal constraints: $f_i(x) \leq 0$, $i = 1, \ldots, m$, $h_i(x) = 0$, $i = 1, \ldots, p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1, \ldots, m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

from page 5–17: if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

KKT Conditions for Convex Problem

if $\tilde{x}$, $\tilde{\lambda}$, $\tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from complementary slackness: $f_0(\tilde{x}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- from 4th condition (and convexity): $g(\tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$

hence, $f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu})$

if **Slater's condition** is satisfied:

x is optimal if and only if there exist λ, ν that satisfy KKT conditions

- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(x) = 0$ for unconstrained problem

An Example—SVM (1)

- The Optimization Problem

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Define the hinge loss as

$$\ell(x) = \max(0, 1-x)$$

- Its Conjugate Function is

$$\ell^*(y) = \sup_x (yx - \ell(x)) = \begin{cases} y, & -1 \leq y \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

An Example—SVM (2)

- The Optimization Problem becomes

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \ell\left(y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- It is Equivalent to

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \mathbf{u}\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ \text{s. t.} \quad & u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b),\ i = 1\ldots, n \end{aligned}$$

- The Lagrangian is

$$\mathcal{L}(\mathbf{w}, b, \mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} v_i\left(u_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right)$$

An Example—SVM (3)

- The Lagrange Dual Function is

$$
\begin{aligned}
g(\mathbf{v}) &= \inf_{\mathbf{w},b,\mathbf{u}} \mathcal{L}(\mathbf{w}, b, \mathbf{u}, \mathbf{v}) \\
&= \inf_{\mathbf{w},b,\mathbf{u}} \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} v_i \left(u_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) \\
&= \inf_{\mathbf{w},b,\mathbf{u}} \sum_{i=1}^{n} (\ell(u_i) + v_i u_i) + \left(\frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \mathbf{w}^\top \sum_{i=1}^{n} v_i y_i \mathbf{x}_i\right) - b \sum_{i=1}^{n} v_i y_i
\end{aligned}
$$

  - Minimize $\mathbf{w}, b, \mathbf{u}$ one by one

$$
\inf_{u_i} (\ell(u_i) + v_i u_i) = -\sup_{u_i} (-v_i u_i - \ell(u_i)) = -\ell^*(-v_i) = v_i, \ \text{if } 0 \leq v_i \leq 1
$$

$$
\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \mathbf{u}, \mathbf{v}) = \lambda \mathbf{w} - \sum_{i=1}^{n} v_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \frac{1}{\lambda} \sum_{i=1}^{n} v_i y_i \mathbf{x}_i
$$

$$
\nabla_b \mathcal{L}(\mathbf{w}, b, \mathbf{u}, \mathbf{v}) = -\sum_{i=1}^{n} v_i y_i = 0
$$

An Example—SVM (4)

- Finally, We Obtain

$$g(\mathbf{v}) = \sum_{i=1}^{n} v_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

- The Dual Problem is

$$\begin{aligned} \max_{\mathbf{v}\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} v_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ \text{s.t.} \quad & 0 \leq v_i \leq 1, \ i = 1\ldots, n \\ \text{s.t.} \quad & \sum_{i=1}^{n} v_i y_i = 0 \end{aligned}$$

An Example—SVM (5)

Karush-Kuhn-Tucker (KKT) Conditions

Let $(\mathbf{w}_*, b_*, \mathbf{u}_*)$ and $\mathbf{v}_*$ are primal and dual solutions.

$$u_{*i} = y_i(\mathbf{w}_*^\top \mathbf{x}_i + b_*)$$

$$\mathbf{w}_* = \frac{1}{\lambda} \sum_{i=1}^{n} v_{*i} y_i \mathbf{x}_i$$

$$\sum_{i=1}^{n} v_{*i} y_i = 0$$

$$u_{*i} = \operatorname*{argmin}_{u_i} \left(\ell(u_i) + v_{*i} u_i \right) = 1 \text{ if } 0 < v_{*i} < 1$$

An Example—SVM (5)

Karush-Kuhn-Tucker (KKT) Conditions

Let $(\mathbf{w}_*, b_*, \mathbf{u}_*)$ and $\mathbf{v}_*$ are primal and dual solutions.

$$u_{*i} = y_i(\mathbf{w}_*^\top \mathbf{x}_i + b_*)$$

$$\mathbf{w}_* = \frac{1}{\lambda} \sum_{i=1}^{n} v_{*i} y_i \mathbf{x}_i$$

Can be used to recover $\mathbf{w}_*$ from $\mathbf{v}_*$

$$\sum_{i=1}^{n} v_{*i} y_i = 0$$

$$u_{*i} = \operatorname*{argmin}_{u_i} \left(\ell(u_i) + v_{*i} u_i\right) = 1 \text{ if } 0 < v_{*i} < 1$$

An Example—SVM (5)

Karush-Kuhn-Tucker (KKT) Conditions

Let $(\mathbf{w}_*, b_*, \mathbf{u}_*)$ and $\mathbf{v}_*$ are primal and dual solutions.

$$u_{*i} = y_i(\mathbf{w}_*^\top \mathbf{x}_i + b_*)$$

$$\mathbf{w}_* = \frac{1}{\lambda}\sum_{i=1}^{n} v_{*i} y_i \mathbf{x}_i$$

$$\sum_{i=1}^{n} v_{*i} y_i = 0$$

$$u_{*i} = \operatorname*{argmin}_{u_i} \left(\ell(u_i) + v_{*i} u_i\right) = 1 \text{ if } 0 < v_{*i} < 1$$

Can be used to recover b_* from $\mathbf{v}_*$

Outline

- Introduction
- Convex Sets & Functions
- Convex Optimization Problems
- Duality
- **Convex Optimization Methods**
- Summary

More Assumptions

Lipschitz continuous

$$\|\nabla f(x)\| \leq G \qquad |f(x) - f(y)| \leq G\|x - y\|$$

Strong Convexity

$$\nabla^2 f(x) \succeq mI$$

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

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m\|x - y\|^2$$

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y) - a(1-a)\frac{m}{2}\|x - y\|^2$$

Smooth

$$\nabla^2 f(x) \preceq MI$$

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

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq M\|x - y\|^2$$

Performance Measure

- The Problem

$$\min_{\mathbf{w} \in \mathcal{W}} f(\mathbf{w})$$

- Convergence Rate
  - After T iterations, the gap between objectives

$$f(\mathbf{w}_T) - f(\mathbf{w}_*) \leq O\left(\frac{1}{\sqrt{T}}\right), O\left(\frac{1}{T}\right), O\left(\frac{1}{T^2}\right), O\left(\frac{1}{\alpha^T}\right)$$

- Iteration Complexity
  - To ensure $f(\mathbf{w}_T) - f(\mathbf{w}_*) \leq \epsilon$, the order of T

$$T \leq O\left(\frac{1}{\epsilon^2}\right), O\left(\frac{1}{\epsilon}\right), O\left(\frac{1}{\sqrt{\epsilon}}\right), O\left(\log \frac{1}{\epsilon}\right)$$

Gradient-based Methods

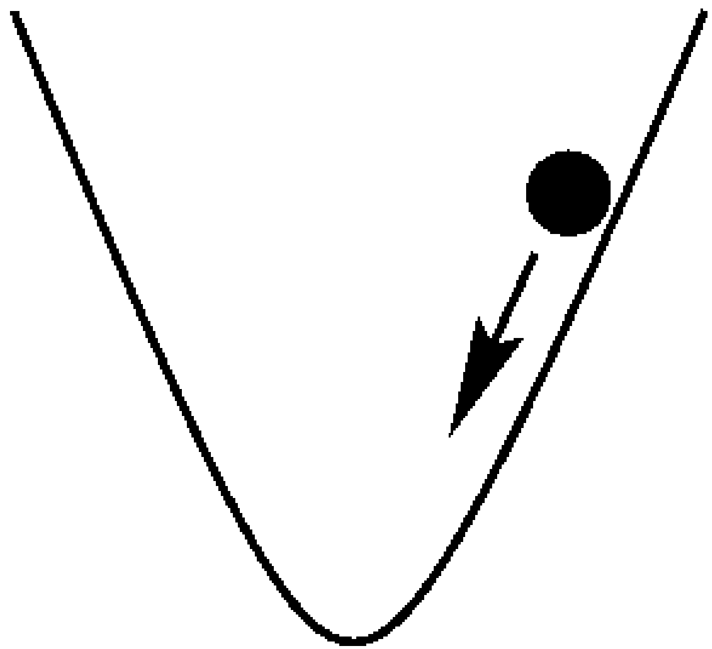
The Convergence Rate

Lipschitz Continuous	Strongly Convex	Smooth	Smooth Strongly Convex
GD $O\left(\frac{1}{\sqrt{T}}\right)$	EGD/SGD$_\alpha$ $O\left(\frac{1}{T}\right)$	AGD $O\left(\frac{1}{T^2}\right)$	GD/AGD $O\left(\frac{1}{\alpha^T}\right)$

- GD—Gradient Descent
- AGD—Nesterov's Accelerated Gradient Descent [Nesterov, 2005, Nesterov, 2007, Tseng, 2008]
- EGD—Epoch Gradient Descent [Hazan and Kale, 2011]
- SGD$_\alpha$—SGD with α-suffix Averaging [Rakhlin et al., 2012]

Gradient Descent (1)

- Move along the opposite direction of gradients



Gradient Descent (2)

Gradient Descent with Projection

for $t = 1, \dots, T$ **do**

$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$$

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$$

end for

return $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

- Projection Operator

$$\Pi_{\mathcal{W}}(\mathbf{y}) = \underset{\mathbf{x}\in\mathcal{W}}{\text{argmin}} \|\mathbf{x} - \mathbf{y}\|_2$$

Analysis (1)

For any $\mathbf{w} \in \mathcal{W}$, we have

$$
\begin{aligned}
& f(\mathbf{w}_t) - f(\mathbf{w}) \\
\leq & \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle \\
= & \frac{1}{\eta_t} \langle \mathbf{w}_t - \mathbf{w}'_{t+1}, \mathbf{w}_t - \mathbf{w} \rangle \\
= & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2 + \|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 \right) \\
= & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2 \right) + \frac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2 \\
\leq & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \right) + \frac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2
\end{aligned}
$$

To simplify the above inequality, we assume

$\eta_t = \eta, \|\nabla f(\mathbf{w})\|_2 \leq G, \ \forall \mathbf{w} \in \mathcal{W}, \ \text{and } \|\mathbf{x} - \mathbf{y}\|_2 \leq D, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{W}$

Analysis (2)

Then, we have

$$f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{1}{2\eta}\left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right) + \frac{\eta}{2}G^2$$

By adding the inequalities of all iterations, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f(\mathbf{w}_t) - Tf(\mathbf{w}) \\
\leq &\frac{1}{2\eta}\left(\|\mathbf{w}_1 - \mathbf{w}\|_2^2 - \|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2\right) + \frac{\eta T}{2}G^2 \\
\leq &\frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta T}{2}G^2 \\
\leq &\frac{1}{2\eta}D^2 + \frac{\eta T}{2}G^2 = GD\sqrt{T}
\end{aligned}$$

where we set

$$\eta = \frac{D}{G\sqrt{T}}$$

Analysis (3)

Then, we have

$$f(\bar{\mathbf{w}}_T) - f(\mathbf{w}) = f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) - f(\mathbf{w})$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{1}{T} GD\sqrt{T} = \frac{GD}{\sqrt{T}}$$

A Key Step (1)

- Evaluate the Gradient or Subgradient
  - Logit loss

$$\ell_i(\mathbf{w}) = \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right)$$

$$\nabla \ell_i(\mathbf{w}) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \exp(-y_i \mathbf{x}_i^\top \mathbf{w})$$

$$= \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla (-y_i \mathbf{x}_i^\top \mathbf{w}) = \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} - y_i \mathbf{x}_i$$

A Key Step (1)

- Evaluate the Gradient or Subgradient
  - Logit loss

$$\ell_i(\mathbf{w}) = \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right)$$

$$\nabla \ell_i(\mathbf{w}) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \exp(-y_i \mathbf{x}_i^\top \mathbf{w})$$

$$= \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla(-y_i \mathbf{x}_i^\top \mathbf{w}) = \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} - y_i \mathbf{x}_i$$

  - Hinge loss

$$\ell_i(\mathbf{w}) = \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$$

A vector λ is a *sub-gradient* of a function f at w if for all u $\in A$ we have that

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{u} - \mathbf{w}, \boldsymbol{\lambda} \rangle .$$

A Key Step (2)

- Evaluate the Gradient or Subgradient
  - Logit loss

$$\ell_i(\mathbf{w}) = \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right)$$

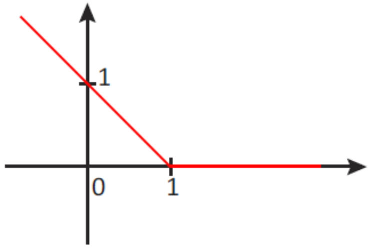
$$\nabla \ell_i(\mathbf{w}) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \exp(-y_i \mathbf{x}_i^\top \mathbf{w})$$

$$= \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla(-y_i \mathbf{x}_i^\top \mathbf{w}) = \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} - y_i \mathbf{x}_i$$

  - Hinge loss

$$\ell_i(\mathbf{w}) = \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$$

$$\partial \max(0, 1 - z) = \begin{cases} -1, & z < 1 \\ 0, & z > 1 \\ [-1, 0], & z = 1 \end{cases}$$



A Key Step (3)

- Evaluate the Gradient or Subgradient
  - Logit loss

$$\ell_i(\mathbf{w}) = \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right)$$

$$\nabla \ell_i(\mathbf{w}) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla \exp(-y_i \mathbf{x}_i^\top \mathbf{w})$$

$$= \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla(-y_i \mathbf{x}_i^\top \mathbf{w}) = \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} - y_i \mathbf{x}_i$$

  - Hinge loss

$$\ell_i(\mathbf{w}) = \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$$

$$\partial \ell_i(\mathbf{w}) = \begin{cases} -y_i \mathbf{x}_i, & y_i \mathbf{x}_i^\top \mathbf{w} < 1 \\ 0, & y_i \mathbf{x}_i^\top \mathbf{w} > 1 \\ \{-\alpha y_i \mathbf{x}_i : \alpha \in [0, 1]\}, & y_i \mathbf{x}_i^\top \mathbf{w} = 1 \end{cases}$$

Outline

- Introduction
- Convex Sets & Functions
- Convex Optimization Problems
- Duality
- Convex Optimization Methods
- **Summary**

Summary

- Convex Sets & Functions
  - Definitions, Operations that Preserve Convexity
- Convex Optimization Problems
  - Definitions, Optimality Criterion
- Duality
  - Lagrange, Dual Problem, KKT Conditions
- Convex Optimization Methods
  - Gradient-based Methods

Reference (1)

- Hazan, E. and Kale, S. (2011)

  Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In Proceedings of the 24th Annual Conference on Learning Theory, pages 421–436.

- Nesterov, Y. (2005)

  Smooth minimization of non-smooth functions. Mathematical Programming, 103(1):127–152.

- Nesterov, Y. (2007).

  Gradient methods for minimizing composite objective function. Core discussion papers.

Reference (2)

- Tseng, P. (2008).

  On acclerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington.

- Boyd, S. and Vandenberghe, L. (2004).

  Convex Optimization. Cambridge University Press.

- Rakhlin, A., Shamir, O., and Sridharan, K. (2012)

  Making gradient descent optimal for strongly convex stochastic optimization. In Proceedings of the 29th International Conference on Machine Learning, pages 449–456.