Second-Order Cone Programming

Julio López L.

Análisis Convexo

Departamento de Ing. Matemática

21 Junio 2011

Outline

What's SOC?

The second-order cone (SOC) in $\mathbb{R}^n$, also called Lorentz cone, of dimension n is defined to be

$$\mathcal{L}_+^n = \{(x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} : \|\bar{x}\| \leq x_1\},$$

where $\|\cdot\|$ denotes the Euclidean norm.

Properties:

- $\mathcal{L}_+^n$ is a convex set in $\mathbb{R}^n$.
- $\mathcal{L}_+^n$ is self-dual, i.e $(\mathcal{L}_+^n)^* = \mathcal{L}_+^n$, where

$$(\mathcal{L}_+^n)^* = \{d \in \mathbb{R} \times \mathbb{R}^{n-1} : z^\top d \geq 0 \; \forall z \in \mathcal{L}_+^n\}.$$

- $\mathcal{L}_{++}^n = \{(x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} : \|\bar{x}\| < x_1\}$ is the interior of the SOC and the set $\partial \mathcal{L}_+^n = \{x \in \mathcal{L}_+^n : \|\bar{x}\| = x_1\}$ denotes its boundary.

If $n = 1$, let $\mathcal{L}_+^n$ denote the set of nonnegative reals $\mathbb{R}_+$.

What's SOC?

The second-order cone (SOC) in $\mathbb{R}^n$, also called Lorentz cone, of dimension n is defined to be

$$\mathcal{L}_+^n = \{(x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} : \|\bar{x}\| \le x_1\},$$

where $\|\cdot\|$ denotes the Euclidean norm.

Properties:

- $\mathcal{L}_+^n$ is a convex set in $\mathbb{R}^n$.
- $\mathcal{L}_+^n$ is self-dual, i.e $(\mathcal{L}_+^n)^* = \mathcal{L}_+^n$, where

$$(\mathcal{L}_+^n)^* = \{d \in \mathbb{R} \times \mathbb{R}^{n-1} : z^\top d \ge 0 \; \forall z \in \mathcal{L}_+^n\}.$$

- $\mathcal{L}_{++}^n = \{(x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} : \|\bar{x}\| < x_1\}$ is the interior of the SOC and the set $\partial\mathcal{L}_+^n = \{x \in \mathcal{L}_+^n : \|\bar{x}\| = x_1\}$ denotes its boundary.

If $n = 1$, let $\mathcal{L}_+^n$ denote the set of nonnegative reals $\mathbb{R}_+$.

What's SOC?

The second-order cone (SOC) in $\mathbb{R}^n$, also called Lorentz cone, of dimension n is defined to be

$$\mathcal{L}_+^n = \{(x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} : \|\bar{x}\| \leq x_1\},$$

where $\|\cdot\|$ denotes the Euclidean norm.

Properties:

- $\mathcal{L}_+^n$ is a convex set in $\mathbb{R}^n$.
- $\mathcal{L}_+^n$ is self-dual, i.e $(\mathcal{L}_+^n)^* = \mathcal{L}_+^n$, where

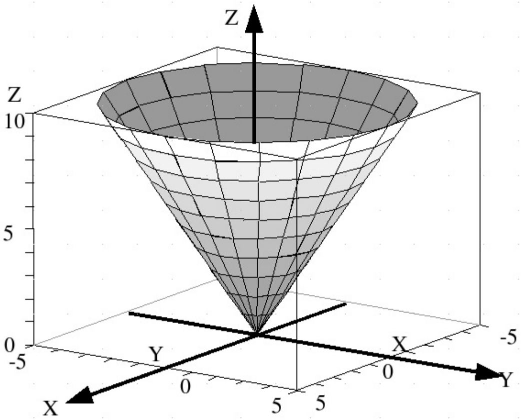
$$(\mathcal{L}_+^n)^* = \{d \in \mathbb{R} \times \mathbb{R}^{n-1} : z^\top d \geq 0 \ \forall z \in \mathcal{L}_+^n\}.$$

- $\mathcal{L}_{++}^n = \{(x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} : \|\bar{x}\| < x_1\}$ is the interior of the SOC and the set $\partial\mathcal{L}_+^n = \{x \in \mathcal{L}_+^n : \|\bar{x}\| = x_1\}$ denotes its boundary.

If $n = 1$, let $\mathcal{L}_+^n$ denote the set of nonnegative reals $\mathbb{R}_+$.

What's SOC?

$$\mathcal{L}_+^3 = \{(x_1, x_2, x_3) \in \mathbb{R} \times \mathbb{R}^2 : \sqrt{x_2^2 + x_3^2} \le x_1\}$$



What's SOCP?

The second-order cone programming (SOCP) problem and its dual are:

$$\begin{aligned} \min \quad & c_1^\top x_1 + \ldots + c_r^\top x_r \\ \text{s.t} \quad & A_1 x_1 + \ldots + A_r x_r = b \\ & x_i \in \mathcal{L}^{n_i},\ i = 1, \ldots, r \end{aligned} \qquad \begin{aligned} \max \quad & b^\top y \\ \text{s.t} \quad & A_i^\top y + s_i = c_i \\ & s_i \in \mathcal{L}^{n_i},\ i = 1, \ldots, r \end{aligned}$$

where $A_i \in \mathbb{R}^{m \times n_i}$.

Let us express the primal and dual problems as

$$\begin{aligned} \min \quad & c^\top x \\ \text{s.t} \quad & \mathbf{A} x = b \\ & x \in \mathcal{K} \end{aligned} \qquad \begin{aligned} \max \quad & b^\top y \\ \text{s.t} \quad & \mathbf{A}^\top y + s = c \\ & s \in \mathcal{K} \end{aligned}$$

where $\mathbf{A} = (A_1, \ldots, A_r) \in \mathbb{R}^{m \times n}$ and $\mathcal{K} = \mathcal{L}^{n_1} \times \ldots \times \mathcal{L}^{n_r}$.

What's SOCP?

The second-order cone programming (SOCP) problem and its dual are:

$$\begin{array}{rl} \min & c_1^\top x_1 + \ldots + c_r^\top x_r \\ \text{s.t} & A_1 x_1 + \ldots + A_r x_r = b \\ & x_i \in \mathcal{L}^{n_i},\ i = 1, \ldots, r \end{array} \qquad \begin{array}{rl} \max & b^\top y \\ \text{s.t} & A_i^\top y + s_i = c_i \\ & s_i \in \mathcal{L}^{n_i},\ i = 1, \ldots, r \end{array}$$

where $A_i \in \mathbb{R}^{m \times n_i}$.

Let us express the primal and dual problems as

$$\begin{array}{rl} \min & c^\top x \\ \text{s.t} & \mathbf{A}x = b \\ & x \in \mathcal{K} \end{array} \qquad \begin{array}{rl} \max & b^\top y \\ \text{s.t} & \mathbf{A}^\top y + s = c \\ & s \in \mathcal{K} \end{array}$$

where $\mathbf{A} = (A_1, \ldots, A_r) \in \mathbb{R}^{m \times n}$ and $\mathcal{K} = \mathcal{L}^{n_1} \times \ldots \times \mathcal{L}^{n_r}$.

KKT conditions and nonlinear SOCP

Under some assumptions (Slater-type constraint qualification), the solutions for the primal-dual SOCP problems satisfy the KKT conditions

$$\begin{aligned} \mathbf{A}^\top y + s &= c \\ \mathbf{A}x &= b \\ x_i, s_i \in \mathcal{L}^n_+,\ x_i^\top s_i &= 0,\ i = 1, \dots, r. \end{aligned}$$

Nonlinear second-order cone program

$$\min f(x) \quad \text{s.t.} \quad Ax = b,\ x \in \mathcal{K},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a proper closed convex function (possibly nonsmooth).

KKT conditions and nonlinear SOCP

Under some assumptions (Slater-type constraint qualification), the solutions for the primal-dual SOCP problems satisfy the KKT conditions

$$\begin{aligned} \mathbf{A}^\top y + s &= c \\ \mathbf{A}x &= b \\ x_i, s_i \in \mathcal{L}_+^n,\ x_i^\top s_i &= 0,\ i = 1, \ldots, r. \end{aligned}$$

Nonlinear second-order cone program

$$\min f(x) \quad \text{s.t.} \quad Ax = b,\ x \in \mathcal{K},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a proper closed convex function (possibly nonsmooth).

Why study SOCP ?

- This problem has wide applications, e.g., Robust linear programming, filter design, structural optimization, support vector machines under uncertainy, etc. (Lobo, Vandenberghe, Boyd, Lebret, 1998)
- It includes a large class of quadratically constrained problems and minimization of sum of Euclidean norms as special cases.
- It also includes as a special case the well-known linear programming (LP): LP$\subset$SOCP

Difficulty: $\mathcal{K}$ is closed and convex, but non-polyhedral.

Why study SOCP ?

- This problem has wide applications, e.g., Robust linear programming, filter design, structural optimization, support vector machines under uncertainy, etc. (Lobo, Vandenberghe, Boyd, Lebret, 1998)
- It includes a large class of quadratically constrained problems and minimization of sum of Euclidean norms as special cases.
- It also includes as a special case the well-known linear programming (LP): LP$\subset$SOCP

Difficulty: $\mathcal{K}$ is closed and convex, but non-polyhedral.

Relation to semidefinite programming (SDP)

Associated with each vector $x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1}$ there is an arrow matrix defined as:

$$\text{Arw}(x) = \begin{pmatrix} x_1 & \bar{x}^\top \\ \bar{x} & x_1 I \end{pmatrix}.$$

Properties:

- $\text{Arw}(x)$ is positive semidefinite if and only if $x \in \mathcal{L}^n_+$.

  $\text{Arw}(x) \succeq 0$ iff either $x = 0$ or $x_1 > 0$ and the Schur complement
  $$x_1 - \bar{x}^\top (x_1 I)^{-1} \bar{x} \geq 0.$$

Relation to semidefinite programming (SDP)

Associated with each vector $x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1}$ there is an arrow matrix defined as:

$$\text{Arw}(x) = \begin{pmatrix} x_1 & \bar{x}^\top \\ \bar{x} & x_1 I \end{pmatrix}.$$

Properties:

- $\text{Arw}(x)$ is positive semidefinite if and only if $x \in \mathcal{L}_+^n$.

  $\text{Arw}(x) \succeq 0$ iff either $x = 0$ or $x_1 > 0$ and the Schur complement $x_1 - \bar{x}^\top (x_1 I)^{-1} \bar{x} \geq 0$.

Relation to semidefinite programming (SDP)

Associated with each vector $x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1}$ there is an arrow matrix defined as:

$$\mathrm{Arw}(x) = \begin{pmatrix} x_1 & \bar{x}^\top \\ \bar{x} & x_1 I \end{pmatrix}.$$

Properties:

- $\mathrm{Arw}(x)$ is positive semidefinite if and only if $x \in \mathcal{L}^n_+$.
- $\mathrm{Arw}(x)$ is positive definite if and only if $x \in \mathcal{L}^n_{++}$.

SOCP is a special of semidefinite programming:

$$\min f(x) \quad \text{s.t.} \quad Ax = b, \ \mathrm{Arw}(x) \succeq 0,$$

Relation to semidefinite programming (SDP)

Associated with each vector $x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1}$ there is an arrow matrix defined as:

$$\mathrm{Arw}(x) = \begin{pmatrix} x_1 & \bar{x}^\top \\ \bar{x} & x_1 I \end{pmatrix}.$$

Properties:

- $\mathrm{Arw}(x)$ is positive semidefinite if and only if $x \in \mathcal{L}^n_+$.
- $\mathrm{Arw}(x)$ is positive definite if and only if $x \in \mathcal{L}^n_{++}$.

SOCP is a special of semidefinite programming:

$$\min f(x) \quad \text{s.t.} \quad Ax = b, \ \mathrm{Arw}(x) \succeq 0,$$

Jordan product

Jordan product: For any $x = (x_1, \bar{x})$, $y = (y_1, \bar{y}) \in \mathbb{R} \times \mathbb{R}^{n-1}$:

$$x \circ y = (x^\top y, x_1 \bar{y} + y_1 \bar{x}).$$

It is easy to verify that

$$x \circ y = \text{Arw}(x) y = \text{Arw}(y) x = y \circ x.$$

Properties:

- The Jordan product is commutative but is not associative.
- $e \circ x = x$ with $e = (1, 0)$, for all $x \in \mathbb{R}^n$.
- $(x + y) \circ z = x \circ z + y \circ z$, for all $x, y, z \in \mathbb{R}^n$.
- $\mathcal{L}_+^n$ is not closed under the Jordan product.
- For any $z \in \mathbb{R}^n$ one has $z^2 = z \circ z \in \mathcal{L}_+^n$.

Jordan product

Jordan product: For any $x = (x_1, \bar{x})$, $y = (y_1, \bar{y}) \in \mathbb{R} \times \mathbb{R}^{n-1}$:

$$x \circ y = (x^\top y, x_1 \bar{y} + y_1 \bar{x}).$$

It is easy to verify that

$$x \circ y = \text{Arw}(x) y = \text{Arw}(y) x = y \circ x.$$

Properties:

- The Jordan product is commutative but is not associative.
- $e \circ x = x$ with $e = (1, 0)$, for all $x \in \mathbb{R}^n$.
- $(x + y) \circ z = x \circ z + y \circ z$, for all $x, y, z \in \mathbb{R}^n$.
- $\mathcal{L}^n_+$ is not closed under the Jordan product.
- For any $z \in \mathbb{R}^n$ one has $z^2 = z \circ z \in \mathcal{L}^n_+$.

Spectral decomposition

Quadratic identity for x:

$$x^2 - 2x_1 x + (x_1^2 - \|\bar{x}\|^2)e = 0.$$

Characteristic polynomial of x:

$$p(\lambda, x) = \lambda^2 - 2x_1\lambda + (x_1^2 - \|\bar{x}\|^2).$$

Roots of characteristic polynomial of x(eigenvalues):

$$\lambda_1(x) = x_1 - \|\bar{x}\|, \quad \lambda_2(x) = x_1 + \|\bar{x}\|.$$

Spectral factorization: $x \in \mathbb{R}^m$ can be decomposed as $(\bar{x} \neq 0)$

$$x = \begin{pmatrix} x_1 \\ \bar{x} \end{pmatrix} = (x_1 - \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix} + (x_1 + \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}.$$

Spectral decomposition

Quadratic identity for x:

$$x^2 - 2x_1 x + (x_1^2 - \|\bar{x}\|^2)e = 0.$$

Characteristic polynomial of x:

$$p(\lambda, x) = \lambda^2 - 2x_1\lambda + (x_1^2 - \|\bar{x}\|^2).$$

Roots of characteristic polynomial of x(eigenvalues):

$$\lambda_1(x) = x_1 - \|\bar{x}\|, \quad \lambda_2(x) = x_1 + \|\bar{x}\|.$$

Spectral factorization: $x \in \mathbb{R}^m$ can be decomposed as $(\bar{x} \neq 0)$

$$x = \begin{pmatrix} x_1 \\ \bar{x} \end{pmatrix} = (x_1 - \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix} + (x_1 + \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}.$$

Spectral decomposition

Quadratic identity for x:

$$x^2 - 2x_1 x + (x_1^2 - \|\bar{x}\|^2)e = 0.$$

Characteristic polynomial of x:

$$p(\lambda, x) = \lambda^2 - 2x_1\lambda + (x_1^2 - \|\bar{x}\|^2).$$

Roots of characteristic polynomial of x(eigenvalues):

$$\lambda_1(x) = x_1 - \|\bar{x}\|, \quad \lambda_2(x) = x_1 + \|\bar{x}\|.$$

Spectral factorization: $x \in \mathbb{R}^m$ can be decomposed as $(\bar{x} \neq 0)$

$$x = \begin{pmatrix} x_1 \\ \bar{x} \end{pmatrix} = (x_1 - \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix} + (x_1 + \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}.$$

Spectral decomposition

Quadratic identity for x:

$$x^2 - 2x_1 x + (x_1^2 - \|\bar{x}\|^2)e = 0.$$

Characteristic polynomial of x:

$$p(\lambda, x) = \lambda^2 - 2x_1\lambda + (x_1^2 - \|\bar{x}\|^2).$$

Roots of characteristic polynomial of x(eigenvalues):

$$\lambda_1(x) = x_1 - \|\bar{x}\|, \quad \lambda_2(x) = x_1 + \|\bar{x}\|.$$

Spectral factorization: $x \in \mathbb{R}^m$ can be decomposed as $(\bar{x} \neq 0)$

$$x = \begin{pmatrix} x_1 \\ \bar{x} \end{pmatrix} = (x_1 - \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix} + (x_1 + \|\bar{x}\|)\frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}.$$

Spectral decomposition

Quadratic identity for x:

$$x^2 - 2x_1 x + (x_1^2 - \|\bar{x}\|^2)e = 0.$$

Characteristic polynomial of x:

$$p(\lambda, x) = \lambda^2 - 2x_1\lambda + (x_1^2 - \|\bar{x}\|^2).$$

Roots of characteristic polynomial of x(eigenvalues):

$$\lambda_1(x) = x_1 - \|\bar{x}\|, \quad \lambda_2(x) = x_1 + \|\bar{x}\|.$$

Spectral factorization: $x \in \mathbb{R}^m$ can be decomposed as $(\bar{x} \neq 0)$

$$x = \begin{pmatrix} x_1 \\ \bar{x} \end{pmatrix} = \underbrace{(x_1 - \|\bar{x}\|)}_{\lambda_1(x)} \frac{1}{2} \begin{pmatrix} 1 \\ -\frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix} + \underbrace{(x_1 + \|\bar{x}\|)}_{\lambda_2(x)} \frac{1}{2} \begin{pmatrix} 1 \\ \frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}.$$

Spectral decomposition

Quadratic identity for x:

$$x^2 - 2x_1 x + (x_1^2 - \|\bar{x}\|^2)e = 0.$$

Characteristic polynomial of x:

$$p(\lambda, x) = \lambda^2 - 2x_1\lambda + (x_1^2 - \|\bar{x}\|^2).$$

Roots of characteristic polynomial of x(eigenvalues):

$$\lambda_1(x) = x_1 - \|\bar{x}\|, \quad \lambda_2(x) = x_1 + \|\bar{x}\|.$$

Spectral factorization: $x \in \mathbb{R}^m$ can be decomposed as $(\bar{x} \neq 0)$

$$x = \begin{pmatrix} x_1 \\ \bar{x} \end{pmatrix} = \underbrace{(x_1 - \|\bar{x}\|)}_{\lambda_1(x)} \underbrace{\frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}}_{u_1(x)} + \underbrace{(x_1 + \|\bar{x}\|)}_{\lambda_2(x)} \underbrace{\frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{x}}{\|\bar{x}\|} \end{pmatrix}}_{u_2(x)}.$$

Spectral decomposition

Case $\bar{x} = 0$:

$$u_1(x) = \frac{1}{2}\begin{pmatrix} 1 \\ -w \end{pmatrix}, \quad u_2(x) = \frac{1}{2}\begin{pmatrix} 1 \\ w \end{pmatrix}, \quad \text{with } w \in \mathbb{R}^{n-1} \text{ s.t. } \|w\| = 1.$$

Properties:

- If $\bar{x} \neq 0$, the decomposition is unique.
- $u_1(x) \circ u_2(x) = 0$.
- $u_i(x) \circ u_i(x) = u_i(x)$ for $i = 1, 2$.
- $x \in \mathcal{L}^n_+$ (resp. $x \in \mathcal{L}^n_{++}$) if and only if $\lambda_1(x), \lambda_2(x) \geq 0$ (resp. > 0).

Trace and determinant of x:

$$\mathrm{tr}(x) = \lambda_1(x) + \lambda_2(x) = 2x_1,$$

$$\det(x) = \lambda_1(x)\lambda_2(x) = x_1^2 - \|\bar{x}\|^2.$$

Spectral decomposition

Case $\bar{x} = 0$:

$$u_1(x) = \frac{1}{2}\begin{pmatrix} 1 \\ -w \end{pmatrix}, \quad u_2(x) = \frac{1}{2}\begin{pmatrix} 1 \\ w \end{pmatrix}, \quad \text{with } w \in \mathbb{R}^{n-1} \text{ s.t. } \|w\| = 1.$$

Properties:

- If $\bar{x} \neq 0$, the decomposition is unique.
- $u_1(x) \circ u_2(x) = 0$.
- $u_i(x) \circ u_i(x) = u_i(x)$ for $i = 1, 2$.
- $x \in \mathcal{L}^n_+$ (resp. $x \in \mathcal{L}^n_{++}$) if and only if $\lambda_1(x), \lambda_2(x) \geq 0$ (resp. > 0).

Trace and determinant of x:

$$\mathrm{tr}(x) = \lambda_1(x) + \lambda_2(x) = 2x_1,$$

$$\det(x) = \lambda_1(x)\lambda_2(x) = x_1^2 - \|\bar{x}\|^2.$$

Spectral decomposition

Case $\bar{x} = 0$:

$$u_1(x) = \frac{1}{2}\begin{pmatrix} 1 \\ -w \end{pmatrix}, \quad u_2(x) = \frac{1}{2}\begin{pmatrix} 1 \\ w \end{pmatrix}, \quad \text{with } w \in \mathbb{R}^{n-1} \text{ s.t. } \|w\| = 1.$$

Properties:

- If $\bar{x} \neq 0$, the decomposition is unique.
- $u_1(x) \circ u_2(x) = 0$.
- $u_i(x) \circ u_i(x) = u_i(x)$ for $i = 1, 2$.
- $x \in \mathcal{L}^n_+$ (resp. $x \in \mathcal{L}^n_{++}$) if and only if $\lambda_1(x), \lambda_2(x) \geq 0$ (resp. > 0).

Trace and determinant of x:

$$\mathrm{tr}(x) = \lambda_1(x) + \lambda_2(x) = 2x_1,$$

$$\det(x) = \lambda_1(x)\lambda_2(x) = x_1^2 - \|\bar{x}\|^2.$$

The SOC-functions

For any function $g : \mathbb{R} \to \mathbb{R}$, we define a corresponding function on $\mathbb{R}^n$ associated with SOC by

$$g^{\mathrm{SOC}}(x) = g(\lambda_1(x))u_1(x) + g(\lambda_2(x))u_2(x), \ \forall x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1}.$$

If g is defined only on a subset of $\mathbb{R}$, then g^{SOC} is defined on the corresponding subset of $\mathbb{R}^n$.

Example

$$g_1(t) = -\ln(t),\ t \in \mathbb{R}_{++} \Rightarrow g_1^{\mathrm{SOC}}(x) = -\ln(\lambda_1)u_1 - \ln(\lambda_2)u_2,\ x \in \mathcal{L}^n_{++}. \\ = -\ln(x),\ x \in \mathcal{L}^n_{++}$$

$$g_2(t) = t\ln(t),\ t \in \mathbb{R}_{+} \Rightarrow g_2^{\mathrm{SOC}}(x) = \lambda_1\ln(\lambda_1)u_1 + \lambda_2\ln(\lambda_2)u_2,\ x \in \mathcal{L}^n_{+} \\ = x \circ \ln(x),\ x \in \mathcal{L}^n_{+}$$

$$g_3(t) = \exp(t),\ t \in \mathbb{R} \Rightarrow g_3^{\mathrm{SOC}}(x) = \exp(\lambda_1)u_1 + \exp(\lambda_2)u_2,\ x \in \mathbb{R}^n.$$

$$g_4(t) = t^{-1},\ t \in \mathbb{R}_{++} \Rightarrow g_4^{\mathrm{SOC}}(x) = \lambda_1^{-1}u_1 + \lambda_2^{-1}u_2 = x^{-1},\ x \in \mathcal{L}^n_{++}.$$

The SOC-functions

For any function $g : \mathbb{R} \to \mathbb{R}$, we define a corresponding function on $\mathbb{R}^n$ associated with SOC by

$$g^{\mathrm{soc}}(x) = g(\lambda_1(x))u_1(x) + g(\lambda_2(x))u_2(x), \ \forall x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1}.$$

If g is defined only on a subset of $\mathbb{R}$, then g^{soc} is defined on the corresponding subset of $\mathbb{R}^n$.

Example

$$g_1(t) = -\ln(t),\ t \in \mathbb{R}_{++} \Rightarrow g_1^{\mathrm{soc}}(x) = -\ln(\lambda_1)u_1 - \ln(\lambda_2)u_2,\ x \in \mathcal{L}^n_{++}. \\ = -\ln(x),\ x \in \mathcal{L}^n_{++}$$

$$g_2(t) = t\ln(t),\ t \in \mathbb{R}_{+} \Rightarrow g_2^{\mathrm{soc}}(x) = \lambda_1\ln(\lambda_1)u_1 + \lambda_2\ln(\lambda_2)u_2,\ x \in \mathcal{L}^n_{+} \\ = x \circ \ln(x),\ x \in \mathcal{L}^n_{+}$$

$$g_3(t) = \exp(t),\ t \in \mathbb{R} \Rightarrow g_3^{\mathrm{soc}}(x) = \exp(\lambda_1)u_1 + \exp(\lambda_2)u_2,\ x \in \mathbb{R}^n.$$

$$g_4(t) = t^{-1},\ t \in \mathbb{R}_{++} \Rightarrow g_4^{\mathrm{soc}}(x) = \lambda_1^{-1}u_1 + \lambda_2^{-1}u_2 = x^{-1},\ x \in \mathcal{L}^n_{++}.$$

Known results about g^{soc}

(a) g^{soc} is continuous iff g is continuous.

(b) g^{soc} is continuously differentiable iff g is continuously differentiable.

(c) g^{soc} is directionally differentiable iff g is directionally differentiable.

(d) g^{soc} is Fréchet-differentiable iff g is Fréchet-differentiable.

(e) g^{soc} is Lipschitz continuous with constant κ iff g is Lipschitz continuous with constant κ.

J.S Chen, X. Chen, P. Teng,
Analysis of nonsmooth vector-valued functions associated with second-order cones,
Math. Program., Ser. B 101: 95Ű117 (2004).

The matrix-valued functions

Let $\mathcal{S}^n$ be the space of $n \times n$ real symmetric matrices. For any $X \in \mathcal{S}^n$, its eigenvalues $\lambda_1, \ldots, \lambda_n$ are real and admits a spectral decomposition:

$$X = P \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} P^\top,$$

where P is orthogonal (i.e., $P^\top P = I$). Then, for any function $g : \mathbb{R} \to \mathbb{R}$, we define a corresponding matrix-valued function $g^{\text{mat}} : \mathcal{S}^n \to \mathcal{S}^n$ by

$$g^{\text{mat}}(X) = P \begin{pmatrix} g(\lambda_1) & & \\ & \ddots & \\ & & g(\lambda_n) \end{pmatrix} P^\top.$$

Parallel results about g^{mat}

(a) g^{mat} is continuous iff g is continuous.

(b) g^{mat} is continuously differentiable iff g is continuously differentiable.

(c) g^{mat} is directionally differentiable iff g is directionally differentiable.

(d) g^{mat} is Fréchet-differentiable iff g is Fréchet-differentiable.

(e) g^{mat} is Lipschitz continuous with constant κ iff g is Lipschitz continuous with constant κ.

A bridge from g^{mat} to g^{soc}

For any $x = (x_1, \bar{x}) \in \mathbb{R} \times \mathbb{R}^n$, let λ_1, λ_2 be its spectral values, then

1. For any $t \in \mathbb{R}$, the matrix $\text{Arw}(x) + tM_{\bar{x}}$ has eigenvalues λ_1, λ_2 and $x_1 + t$ of multiplicity $n - 2$, where

$$M_{\bar{x}} = \begin{pmatrix} 0 & 0 \\ 0 & I - \frac{\bar{x}\bar{x}^\top}{\|\bar{x}\|^2} \end{pmatrix}.$$

2. For any $g : \mathbb{R} \to \mathbb{R}$ and $t \in \mathbb{R}$, we have

$$g^{\text{soc}}(x) = g^{\text{mat}}\left(\text{Arw}(x) + tM_{\bar{x}}\right) e,$$

where $e = (1, 0, \ldots, 0)^\top \in \mathbb{R}^n$.

Spectrally defined function

For any function $g : \mathbb{R} \to \mathbb{R}$, we define a corresponding spectrally defined function $\Psi_g : \mathbb{R}^n \to \mathbb{R}$ by:

$$\Psi_g(x) = \mathrm{tr}(g^{\mathrm{SOC}}(x)) = g(\lambda_1(x)) + g(\lambda_2(x)).$$

Example (Log-barrier)

$$g_1(t) = -\ln(t),\ t \in \mathbb{R}_{++} \quad \Rightarrow \quad \begin{aligned} \Psi_{g_1}(x) &= -\ln(\lambda_1(x)) - \ln(\lambda_2(x)) \\ &= -\ln(\det(x)),\ x \in \mathcal{L}^n_{++} \end{aligned}$$

Example

$$g_2(t) = t\ln(t),\ t \in \mathbb{R}_{+} \quad \Rightarrow \quad \begin{aligned} \Psi_{g_2}(x) &= \lambda_1 \ln(\lambda_1) + \lambda_2 \ln(\lambda_2),\ x \in \mathcal{L}^n_{+} \\ &= \mathrm{tr}(x \circ \ln(x)),\ x \in \mathcal{L}^n_{+}. \end{aligned}$$

Properties:

- The real-valued function $\Psi_g(x) = -\ln(\det(x))$ is convex on $\mathcal{L}^n_{++}$.
- The gradient of $\Psi_g(x)$ is

$$\nabla\Psi_g(x) = -2x^{-1}.$$

- The Hessian of $\Psi_g(x)$ is

$$\nabla^2\Psi_g(x) = 2(Q_x)^{-1} = 2Q_{x^{-1}} = \frac{2}{\det^2(x)}\begin{pmatrix} \|x\|^2 & -2x_1\bar{x}^\top \\ -2x_1\bar{x} & \det(x)I + 2\bar{x}\bar{x}^\top \end{pmatrix}.$$

Here,

$$Q_x = \begin{pmatrix} \|x\|^2 & 2x_1\bar{x}^\top \\ 2x_1\bar{x} & \det(x)I + 2\bar{x}\bar{x}^\top \end{pmatrix}.$$

- The real-valued function $\Psi_g(x) = -\mathrm{tr}(x^{-1}) = \frac{\mathrm{tr}(x)}{\det(x)}$ is convex on $\mathcal{L}^n_{++}$.

Our Problem SOCP

We consider the following convex second-order cone programming

$$\text{(SOCP)}\ f_* = \min_{x\in\mathbb{R}^n} f(x);\ \mathbf{B}x = \mathbf{d},\ w^j(x) = A^j x + b^j \in \mathcal{L}_+^{m_j},\ j = 1,\dots,J$$

where

- $A^j \in \mathbb{R}^{m_j\times n}$ full rank, $b^j \in \mathbb{R}^{m_j}$, $j = 1,\dots,J$
- $\mathbf{B} \in \mathbb{R}^{r\times n}$ full rank with $r \leq n$, $d \in \mathbb{R}^r$
- $f : \mathbb{R}^n \to \mathbb{R}$ convex (possibly nonsmooth) and defined everywhere

Relative interior of the feasible set:

$$C = \{x \in \mathbb{R}^n : \mathbf{B}x = \mathbf{d},\ w^j(x) \in \mathcal{L}_{++}^{m_j}, j = 1,\dots,J\}$$

Algorithm with Bregman distance

Step 0: Start with $x^0 \in C$. Set $k = 0$.

Step 1: Given $x^k \in C$, and $\gamma_k > 0$, find x^{k+1} solution of

$$\min_x \{ f(x) + \gamma_k \sum_{j=1}^{J} D_\psi(w^j(x), w^j(x^k)) \, ; \, \mathbf{B}x = \mathbf{d} \}.$$

$$(\text{with } D_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle)$$

Step 2: If x^{k+1} satisfies a given criterium (KKT, etc.), then stop.

Step 3: Replace k by $k+1$ and go to step 1.

Assumptions and Strategy

Assumptions

(A1) $f_* > -\infty$

(A2) Slater's condition: dom $f \cap C \neq \emptyset$.

Strategy

Introduce the induced norm:

$$\|u\|_{\mathbf{M}} := (u, u)_{\mathbf{M}}^{\frac{1}{2}}$$

where

$$(u, v)_{\mathbf{M}} = \langle \mathbf{A}^\top \mathbf{M} \mathbf{A} u, v \rangle,$$

and $\mathbf{M} = \mathrm{Diag}(M^1, \ldots, M^J)$ a block diagonal matrix with $M^j \in \mathcal{S}_{++}^{m_j}$ for $j = 1, \ldots, J$ and $\mathbf{A} = (A^1; \ldots; A^J) \in \mathbb{R}^{q \times n}$ with $q = \sum_{j=1}^{J} m_j$.

Algorithm PAVM

Step 0: Start with $x^0 \in C$ and $\mathbf{M}_0 \in \mathcal{S}^q_{++}$ ($q = \sum_{j=1}^J m_j$).
Set $k = 0$.

Step 1: Given $x^k \in C$, $\mathbf{M}_k \in \mathcal{S}^q_{++}$ and $\gamma_k > 0$,
find x^{k+1} solution of

$$\min_x \{ f(x) + \frac{\gamma_k}{2} \|x - x^k\|^2_{\mathbf{M}_k} \; ; \; \mathbf{B}x = \mathbf{d} \}.$$

Go bundle

Step 2: If x^{k+1} satisfies a given criterium (KKT, etc.), then stop.

Step 3: Update $\mathbf{M}_{k+1}$. Replace k by $k+1$ and go to step 1.

Algorithm PAVM

Step 0: Start with $x^0 \in C$, $g^0 \in \partial f(x^0)$ and $\mathbf{M}_0 \in \mathcal{S}^q_{++}$ $(q = \sum_{j=1}^J m_j)$.
Set $k = 0$.

Step 1: Given $x^k \in C$, $g^k \in \partial f(x^k)$, $\mathbf{M}_k \in \mathcal{S}^q_{++}$ and $\gamma_k > 0$,
find x^{k+1}, $g^{k+1} \in \mathbb{R}^n$ and $\omega^{k+1} \in \mathbb{R}^r$ such that

$$g^{k+1} \in \partial f(x^{k+1}),$$

$$g^{k+1} + \gamma_k \mathbf{A}^\top \mathbf{M}_k \mathbf{A}(x^{k+1} - x^k) + \mathbf{B}^\top \omega^{k+1} = 0.$$

$$\mathbf{B}x^{k+1} = \mathbf{d}.$$

(if f is linear then $x^{k+1} = x^k + \gamma_k^{-1} \Delta x^k$)

Step 2: If x^{k+1} satisfies a given criterium (KKT, etc.), then stop.

Step 3: Update $\mathbf{M}_{k+1}$. Replace k by $k + 1$ and go to step 1.

Hessian Log-barrier function

The Hessian of Ψ_g:

$$\nabla^2 \Psi_g(w) = 2(Q_w)^{-1},$$

where

$$(Q_w)^{-1} = \frac{1}{\det^2(w)} \begin{pmatrix} \|w\|^2 & -2w_1 \bar{w}^\top \\ -2w_1 \bar{w} & \det(w) I + 2\bar{w}\bar{w}^\top \end{pmatrix}$$

We consider the norm induced by the Hessian of the Log-barrier:

$$\mathbf{M}_k = 2\mathbf{Q}^{-1}_{\mathbf{w}(x^k)} = 2\text{diag}(Q^{-1}_{w^1(x^k)}, \ldots, Q^{-1}_{w^J(x^k)}).$$

Algorithm PAVM-Hessian

Step 0: Start with $x^0 \in C$, $g^0 \in \partial f(x^0)$ and compute $\mathbf{Q}^{-1}_{\mathbf{w}(x^0)}$. Set $k = 0$.

Step 1: Given $x^k \in C$, $g^k \in \partial f(x^k)$ and $\gamma_k > 0$,
find x^{k+1}, $g^{k+1} \in \mathbb{R}^n$ and $\omega^{k+1} \in \mathbb{R}^r$ such that

$$g^{k+1} \in \partial f(x^{k+1}),$$

$$g^{k+1} + 2\gamma_k \mathbf{A}^\top \mathbf{Q}^{-1}_{\mathbf{w}(x^k)} \mathbf{A}(x^{k+1} - x^k) + \mathbf{B}^\top \omega^{k+1} = 0.$$

$$\mathbf{B}x^{k+1} = \mathbf{d}.$$

(if f is linear then $x^{k+1} = x^k + \gamma_k^{-1} \Delta x^k$)

Step 2: If x^{k+1} satisfies a given criterium (KKT, etc.), then stop.

Step 3: Replace k by $k+1$ and go to step 1.

Interior Point Iterates and Boundedness

Proposition

Suppose that:

$$\gamma_k > \bar{\gamma}_k \quad \text{for every } k = 0, 1, \ldots$$

(that is, the "step length" γ_k^{-1} should be small enough) where

$$\bar{\gamma}_k = \frac{\sqrt{2}}{2}(\sigma_{min}(A))^{-1} \boldsymbol{\lambda}_{\max}(\mathbf{Q}_{\mathbf{w}(x^k)})^{1/2}(\|g^k\| + \delta_k)$$

Then the sequence $\{x^k\}$ generated by PAVM is contained in C.

Proposition

(i) $\{f(x^k)\}$ converges.

(ii) If $\mathcal{X}^*$ is nonempty and bounded, then $\{x^k\}$ is bounded.

Interior Point Iterates and Boundedness

Proposition

Suppose that:

$$\gamma_k > \bar{\gamma}_k \quad \text{for every } k = 0, 1, \ldots$$

(that is, the "step length" γ_k^{-1} should be small enough) where

$$\bar{\gamma}_k = \frac{\sqrt{2}}{2}(\sigma_{min}(A))^{-1}\lambda_{\max}(\mathbf{Q}_{\mathbf{w}(x^k)})^{1/2}(\|g^k\| + \delta_k)$$

Then the sequence $\{x^k\}$ generated by PAVM is contained in C.

Proposition

(i) $\{f(x^k)\}$ converges.

(ii) If $\mathcal{X}^*$ is nonempty and bounded, then $\{x^k\}$ is bounded.

Convergence results

KKT conditions:

$$g + \mathbf{B}^\top \omega = \mathbf{A}^\top \mathbf{s}, \quad \mathbf{B}x = \mathbf{d}, \quad \mathbf{w}(x) \in \mathcal{K}, \quad \mathbf{s} \in \mathcal{K}, \quad \mathbf{w}(x) \circ \mathbf{s} = 0,$$

where $\mathcal{K} = \mathcal{L}_+^{m_1} \times \ldots \times \mathcal{L}_+^{m_J}$, $\omega \in \mathbb{R}^r$, $g \in \partial f(x)$.

Proposition

Assume that $\mathcal{X}^*$ is nonempty and bounded. Then any limit point $(\tilde{x}, \tilde{\mathbf{s}}, \tilde{g}, \tilde{\omega})$ of $\{(x^k, \mathbf{s}^k, g^k, \omega^k)\}$ satisfy:

$$\begin{cases} \tilde{g} + \mathbf{B}^\top \tilde{\omega} = \mathbf{A}^\top \tilde{\mathbf{s}}, \quad \mathbf{B}\tilde{x} = \mathbf{d}, \quad \mathbf{w}(\tilde{x}) \in \mathcal{K}, \\ \lambda_{\max}(\tilde{s}^j) \geq 0 \text{ and } w^j(\tilde{x})^\top \tilde{s}^j = 0, \ j = 1, \ldots, J, \end{cases}$$

where the dual sequence $\{\mathbf{s}^{k+1}\}$ defined by

$$\mathbf{s}^{k+1} := 2\gamma_k \mathbf{Q}^{-1}_{\mathbf{w}(x^k)} (\mathbf{w}(x^k) - \mathbf{w}(x^{k+1})).$$

Convergence results

KKT conditions:

$$g + \mathbf{B}^\top \omega = \mathbf{A}^\top \mathbf{s}, \quad \mathbf{B}x = \mathbf{d}, \quad \mathbf{w}(x) \in \mathcal{K}, \quad \mathbf{s} \in \mathcal{K}, \quad \mathbf{w}(x) \circ \mathbf{s} = 0,$$

where $\mathcal{K} = \mathcal{L}_+^{m_1} \times \ldots \times \mathcal{L}_+^{m_J}$, $\omega \in \mathbb{R}^r$, $g \in \partial f(x)$.
A complete different approach based on recession analysis leads to a fully convergence result for the linear SOCP.

Proposition

Assume that f is linear and that $\mathcal{X}^*$ is nonempty and bounded. If the following inclusion holds for each $j = 1, \ldots, J$

$$A^j(\mathrm{Ker}\,\mathbf{B}) \supseteq \mathcal{L}_+^{m_j},$$

then $\tilde{\mathbf{s}} \in \mathcal{K}$. In consequence any limit point of $\{x^k\}$ satisfies the KKT conditions.

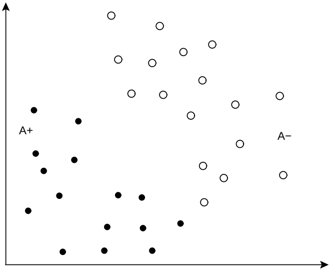
Motivation:Example

Suppose we have 50 photographs of elephants and 50 photos of tigers.

We digitize them into 100×100 pixel images, so we have $x \in \mathbb{R}^n$ where $n = 10000$.

Now, given a new (different) photograph we want to answer the question: **is it an elephant or a tiger?**

Classification Problem



Main goal:

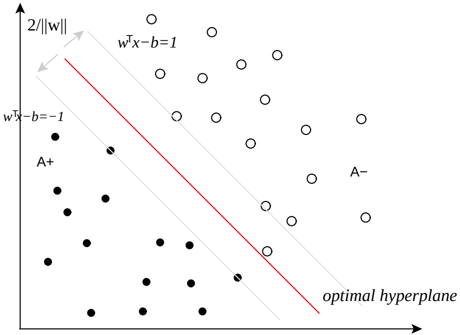
Predict the unseen class label for new data

Find a function $h : \mathbb{R}^n \to \mathbb{R}$ by learning from data

$$h(x) > 0 \Rightarrow x \in A_+ \text{ and } h(x) < 0 \Rightarrow x \in A_-$$

The simplest function is linear: $h(x) = w^\top x - b$.

Maximizing the Margin between Bounding Planes



Margin: Distance between hyperplanes defined by support vectors $\{x_i : |w^\top x_i - b| = 1\}$.

Distance between hyperplanes

Distance of a point x to hyperplane $H(w, b)$:

$$d(w, b; x) = \frac{|w^\top x - b|}{\|w\|}.$$

The margin is given by:

$$\begin{aligned}
\rho(w, b) &= \min_{x_i : y_i = -1} d(w, b; x_i) + \min_{x_i : y_i = 1} d(w, b; x_i) \\
&= \min_{x_i : y_i = -1} \frac{|w^\top x_i - b|}{\|w\|} + \min_{x_i : y_i = 1} \frac{|w^\top x_i - b|}{\|w\|} \\
&= \frac{1}{\|w\|} \left(\min_{x_i : y_i = -1} |w^\top x_i - b| + \min_{x_i : y_i = 1} |w^\top x_i - b| \right) \\
&= \frac{2}{\|w\|}.
\end{aligned}$$

Classification under certainty (Linearly separable)

Let us consider a training dataset

$$\mathcal{T} = \{(x_i, y_i) : x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \dots, m\}.$$

$$x_i \in A_+ \Leftrightarrow y_i = 1 \quad \& \quad x_i \in A_- \Leftrightarrow y_i = -1.$$

Optimal hyperplane $H(w, b)$:

$$\begin{aligned} \min_{w,b \in \mathbb{R}^{n+1}} \quad & \|w\| \\ \text{s.t.} \quad & y_i(w^\top x_i - b) \geq 1, \; i = 1, \dots, m. \end{aligned}$$

Classification under certainty (Linearly separable)

Let us consider a training dataset

$$\mathcal{T} = \{(x_i, y_i) : x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \ldots, m\}.$$

$$x_i \in A_+ \Leftrightarrow y_i = 1 \quad \& \quad x_i \in A_- \Leftrightarrow y_i = -1.$$

Optimal hyperplane $H(w, b)$:

$$\begin{array}{cl} \min\limits_{w,b \in \mathbb{R}^{n+1}} & \frac{1}{2}\|w\|^2 \\ \text{s.t.} & y_i(w^\top x_i - b) \geq 1, \; i = 1, \ldots, m. \end{array}$$

Soft-margin SVM (Nonseparable case)

- If data are not linearly separable
  - Primal problem is infeasible
  - Dual problem is unbounded above
- Introduce the slack variable for each training point

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i = 1, \dots, m.$$

An error occurs if $\xi_i > 1$ (Misclassified).

- The inequality system is always feasible, e.g.

$$w = 0, \ b = 0, \ \xi = \mathbf{1}.$$

Soft-margin SVM (Nonseparable case)

- If data are not linearly separable
  - Primal problem is infeasible
  - Dual problem is unbounded above
- Introduce the slack variable for each training point

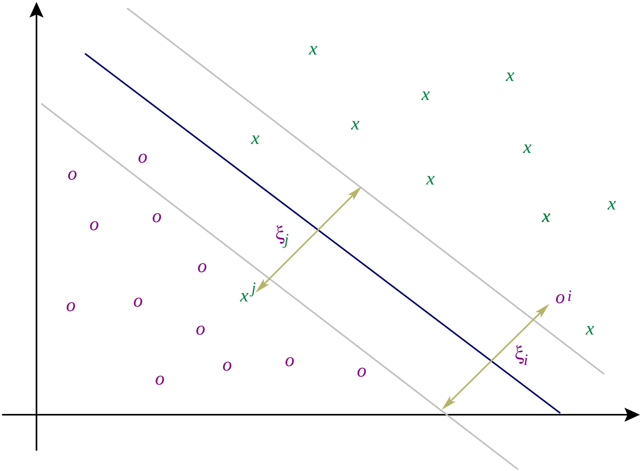
$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,\ \forall i = 1, \ldots, m.$$

An error occurs if $\xi_i > 1$ (Misclassified).

- The inequality system is always feasible, e.g.

$$w = 0,\ b = 0,\ \xi = \mathbf{1}.$$

Soft-margin SVM (Nonseparable case)



Soft-margin SVM (Nonseparable case)

Optimal hyperplane $H(w, b)$:

$$(QP)\quad \begin{array}{cl} \min\limits_{w,b\in\mathbb{R}^{n+1}} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} & y_i(w^\top x_i + b) \geq 1 - \xi_i,\ i = 1, \dots, m, \\ & \xi_i \geq 0,\ i = 1, \dots, m. \end{array}$$

The parameter $C > 0$ is the penalty parameter of the error term.

Unconstrained formulation (Nonsmooth SVM):

$$\min_{w,b\in\mathbb{R}^{n+1}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(1 - y_i(w^\top x_i - b))_+,$$

where $(\cdot)_+ = \max\{0, \cdot\}$.

- Change (QP) into an unscontrained minimization problem.
- Reduce $(n + m + 1)$ variables to $(n + 1)$ variables

Soft-margin SVM (Nonseparable case): Insensitive

Unconstrained insensitive formulation (Nonsmooth SVM):

$$\min_{w,b\in\mathbb{R}^{n+1}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(1 - y_i(w^\top x_i - b))_\epsilon,$$

where $(\cdot)_\epsilon = \max\{\epsilon, \cdot\}$ with $\epsilon > 0$ given.

Algorithms for solving nonsmooth problems:

- Cutting planes.
- Bundle methods.
- ...

J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal and C. Sagastizábal,
Numerical Optimization: Theoretical and Practical Aspects,
Universitext, Springer-Verlag, Berlin, 2003.

Classification under uncertainty

- In many classifications tasks the cost of misclassification is different for each class.
- For instance, in case of medical diagnosis of cancer, the cost of misclassifying a normal patient is far less than that of misclassifying a cancer patient.
- Also, the number of patients with cancer is far less than those who are normal (training data are highly unbalanced).
- Traditional classification methods like SVM do not address these issues satisfactory.
- Hence, this problem is studied in other context.

False positive: Is when there is no disease but the results come back as positive.

False negative: Is when there actually is disease but the results come back as negative.

Classification under uncertainty

- In many classifications tasks the cost of misclassification is different for each class.
- For instance, in case of medical diagnosis of cancer, the cost of misclassifying a normal patient is far less than that of misclassifying a cancer patient.
- Also, the number of patients with cancer is far less than those who are normal (training data are highly unbalanced).
- Traditional classification methods like SVM do not address these issues satisfactory.
- Hence, this problem is studied in other context.

False positive: Is when there is no disease but the results come back as positive.

False negative: Is when there actually is disease but the results come back as negative.

Classification under uncertainty

- Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be random vector variables that generate samples of class A_+ and A_-, resp.
- $\mu_i \in \mathbb{R}^n$ and $\Sigma_i \in \mathbb{R}^{n\times n}$ mean and covariance matrix of $\mathbf{X}_i$, $i = 1, 2$.

Goal: To construct a maximum margin linear classifier s.t. false-positive and false-negative error rates do not exceed $\eta_1 \in (0, 1]$ and $\eta_2 \in (0, 1]$ (Saketha PhD thesis, 2007).

Quadratic Chance-constrained programming:

$$\begin{array}{ll} \min\limits_{w,b} & \frac{1}{2}\|w\|^2 \\ & \text{Prob}\{w^\top \mathbf{X}_1 - b < 0\} \le \eta_1, \\ & \text{Prob}\{w^\top \mathbf{X}_2 - b > 0\} \le \eta_2. \end{array}$$

(Require that $\mathbf{X}_i$ lies on the correct side with probability greater than $1 - \eta_i$).

Case: Normal distribution

Assume that $\mathbf{X}_i$ are distributed according to a normal distribution, the above constraints becomes:

$$\sup_{\mathbf{X}_i \sim \mathcal{N}(\mu_i, \Sigma_i)} \text{Prob}\{y_i(w^\top \mathbf{X}_i - b) < 0\} \leq \eta_i, \quad i = 1, 2.$$

Then,

$$1 - \eta_i \quad \leq \quad \inf_{\mathbf{X}_i \sim \mathcal{N}(\mu_i, \Sigma_i)} \text{Prob}\{y_i(w^\top \mathbf{X}_i - b) > 0\} = \Phi\left(\frac{y_i(w^\top \mathbf{X}_i - b)}{\sqrt{w^\top \Sigma_i w}}\right),$$

where $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-s^2/2)ds$.

Since that Φ is monotone increasing:

$$y_i(w^\top \mathbf{X}_i - b) \geq \kappa_i \sqrt{w^\top \Sigma_i w}, \quad i = 1, 2,$$

where $\kappa_i = \Phi^{-1}(1 - \eta_i)$

Case: Robust formulation

Assume that only know the mean and covariance matrix of X_i. In this case, we want to able to classify correctly even for the *worst distribution*.

Replacing the probability constraints with their robust counterparts:

$$(*) \qquad \sup_{\mathbf{X}_i \sim (\mu_i, \Sigma_i)} \mathrm{Prob}\{y_i(w^\top \mathbf{X}_i - b) < 0\} \leq \eta_i, \quad i = 1, 2,$$

where $\mathbf{X}_i \sim (\mu_i, \Sigma_i)$ denotes a family of distributions which have a common mean and covariance.

Multivariate Chebyshev-Cantelli inequality transform $(*)$ to:

$$y_i(w^\top \mathbf{X}_i - b) \geq \kappa_i \sqrt{w^\top \Sigma_i w}, \quad i = 1, 2,$$

where $\kappa_i = \sqrt{\frac{1-\eta_i}{\eta_i}}$.

Formulation as a SOCP problem

Quadratic Chance-constrained programming:

$$\min_{(w,b)\in\mathbb{R}^{n+1}} \quad \frac{1}{2}\|w\|^2$$
$$\text{Prob}\{w^\top \mathbf{X}_1 - b < 0\} \leq \eta_1,$$
$$\text{Prob}\{w^\top \mathbf{X}_2 - b > 0\} \leq \eta_2.$$

As the constraints are positively homogenous, we consider $\text{Prob}\{y_i(w^\top \mathbf{X}_i - b) \leq 1\} \leq \eta_i$. Hence:

Determinist optimization problem:

$$(Psvm) \qquad \min_{(w,b)\in\mathbb{R}^{n+1}} \frac{1}{2}\|w\|^2$$
$$w^\top \mu_1 - b \geq 1 + \kappa_1\|S_1^\top w\|,$$
$$b - w^\top \mu_2 \geq 1 + \kappa_2\|S_2^\top w\|,$$

where $\Sigma_i = S_i S_i^\top$ and $\kappa_i > 0$.

Formulation as a SOCP problem

Quadratic Chance-constrained programming:

$$\min_{(w,b)\in\mathbb{R}^{n+1}} \quad \frac{1}{2}\|w\|^2$$
$$\text{Prob}\{w^\top \mathbf{X}_1 - b < 0\} \leq \eta_1,$$
$$\text{Prob}\{w^\top \mathbf{X}_2 - b > 0\} \leq \eta_2.$$

As the constraints are positively homogenous, we consider $\text{Prob}\{y_i(w^\top \mathbf{X}_i - b) \leq 1\} \leq \eta_i$. Hence:

Determinist optimization problem:

$$(Psvm) \qquad \min_{(w,b)\in\mathbb{R}^{n+1}} \frac{1}{2}\|w\|^2$$
$$w^\top \mu_1 - b \geq 1 + \kappa_1\|S_1^\top w\|,$$
$$b - w^\top \mu_2 \geq 1 + \kappa_2\|S_2^\top w\|,$$

where $\Sigma_i = S_i S_i^\top$ and $\kappa_i > 0$.

Formulation as a SOCP problem

Quadratic Chance-constrained programming:

$$\min_{(w,b)\in\mathbb{R}^{n+1}} \quad \tfrac{1}{2}\|w\|^2$$
$$\text{Prob}\{w^\top \mathbf{X}_1 - b < 0\} \leq \eta_1,$$
$$\text{Prob}\{w^\top \mathbf{X}_2 - b > 0\} \leq \eta_2.$$

Second order cone programming:

$$\min_{z\in\mathbb{R}^{n+1}} \frac{1}{2}\|w\|^2;\ g_i(z) = A^i z + d_i \in \mathcal{L}^{n+1},\ i = 1, 2.$$

where

$$A^1 = \begin{pmatrix} \mu_1^\top & -1 \\ \kappa_1 S_1^\top & 0 \end{pmatrix},\ A^2 = \begin{pmatrix} -\mu_2^\top & 1 \\ \kappa_2 S_2^\top & 0 \end{pmatrix},\ d_1 = d_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Formulation as a SOCP problem

Quadratic Chance-constrained programming:

$$\min_{(w,b)\in\mathbb{R}^{n+1}} \quad \frac{1}{2}\|w\|^2$$
$$\text{Prob}\{w^\top \mathbf{X}_1 - b < 0\} \leq \eta_1,$$
$$\text{Prob}\{w^\top \mathbf{X}_2 - b > 0\} \leq \eta_2.$$

Linear SOCP problem:

$$\min_{(w,b,t)\in\mathbb{R}^{n+2}} \quad t$$
$$t \geq \|w\|,$$
$$w^\top \mu_1 - b \geq 1 + \kappa_1\|S_1^\top w\|,$$
$$b - w^\top \mu_2 \geq 1 + \kappa_2\|S_2^\top w\|.$$

where $\Sigma_i = S_i S_i^\top$ and $\kappa_i > 0$.

Numerical experience

Dataset: Customers lost.
A portfolio of clients ($m = 1248$-training data) with $n = 19$ descriptions of each one.

The descriptor were divided into four categories:

- banking behavior variables: average monthly balances, number of monthly transactions, ...
- socio-demographic variables: age, salary, ...
- variables perceptions of service quality: number of complaints, ...
- environment variables: antiquity customer, ...

We use the linear classifier.

Numerical experience (certainty)

Dataset: Customers lost.
A portfolio of clients ($m = 1248$) with $n = 19$ descriptions of each one.

Customers	Num. training data	Num. test data
closed	619	67
not closed	629	71

Customers	closed	not closed	Total	Classification rate
closed	64	3	67	95.5%
not closed	18	53	71	74.7%

Numerical experience (uncertainty)

Dataset: Customers lost.
A portfolio of clients ($m = 1248$) with $n = 19$ descriptions of each one.

Customers	**Num. training data**	**Num. test data**	η_i
closed	619	67	0.9
not closed	629	71	0.7

Customers	closed	not closed	Total	Classification rate
closed	44	23	67	65.67%
not closed	11	60	71	84.51%

Numerical experience (uncertainty)

Dataset: Customers lost.
A portfolio of clients ($m = 1248$) with $n = 19$ descriptions of each one.

Customers	Num. training data	Num. test data	η_i
closed	619	67	0.7
not closed	629	71	0.7

Customers	closed	not closed	Total	Classification rate
closed	55	12	67	82.09%
not closed	13	58	71	81.69%

Numerical experience (uncertainty)

Dataset: Customers lost.
A portfolio of clients ($m = 1248$) with $n = 19$ descriptions of each one.

Customers	Num. training data	Num. test data	η_i
closed	619	67	0.5
not closed	629	71	0.7

Customers	closed	not closed	Total	Classification rate
closed	26	41	67	38.81%
not closed	11	60	71	84.51%

Nonsmooth case: Bundle Method

- Let $J^\ell = \{0, 1, \ldots, \ell\} \subset \mathbb{N}$ be a finite index set.
- Bundle: $\mathcal{B}_\ell = \{(y^j, f(y^j), g^j) : j \in J^\ell\}$ with $g^j \in \partial f(y^j)$.
- Cutting-planes model $\varphi_\ell(y) = \max_{j \in J^\ell}\{f(y^j) + \langle g^j, y - y^j\rangle\}$.
- Replacing f by φ_ℓ in $(prox)$

$$\min_{y \in \mathbb{R}^p}\{\varphi_\ell(y) + \frac{1}{2}\gamma_k\|y - x^k\|^2_{\mathbf{M}_k} : \mathbf{B}y = \mathbf{d}\}, \qquad (*)$$

Equivalent problem:

$$\begin{array}{rl} \min\limits_{(r,y) \in \mathbb{R}^{p+1}} & \{r + \frac{1}{2}\gamma_k\|y - x^k\|^2_{\mathbf{M}_k}\} \\ s.t. & \mathbf{B}y = \mathbf{d} \\ & f(x^k) - e_j + \langle g^j, y - x^k\rangle \leq r, \ \forall j \in J^\ell, \end{array}$$

with e^j the linearization error at x^k.

Bundle Method

- Let $J^\ell = \{0, 1, \dots, \ell\} \subset \mathbb{N}$ be a finite index set.
- Bundle: $\mathcal{B}_\ell = \{(y^j, f(y^j), g^j) : j \in J^\ell\}$ with $g^j \in \partial f(y^j)$.
- Cutting-planes model $\varphi_\ell(y) = \max_{j\in J^\ell}\{f(y^j) + \langle g^j, y - y^j\rangle\}$.
- Replacing f by φ_ℓ in (*prox*)

$$\min_{y\in\mathbb{R}^p}\{\varphi_\ell(y) + \frac{1}{2}\gamma_k\|y - x^k\|^2_{\mathbf{M}_k} : \mathbf{B}y = \mathbf{d}\}.$$

Dual problem (DP):

$$\begin{aligned}\min_{(\alpha,w)\in\mathbb{R}^{|J^\ell|}\times\mathbb{R}^r} \quad & \{\frac{1}{2}\left\|\mathbf{B}^\top w - \sum_{j\in J^\ell}\alpha_j g^j\right\|^{*2}_{\mathbf{M}_k} + \gamma_k\sum_{j\in J^\ell}\alpha_j e_j\} \\ s.t. \quad & \sum_{j\in J^\ell}\alpha_j = 1,\ \alpha_j \geq 0, \quad \forall j \in J^\ell.\end{aligned}$$

Bundle PAVM Algorithm

Step 0: Choose parameters $tol \geq 0$ and $m \in (0,1)$. Select $x^0 \in C$, $g^0 \in \partial f(x^0)$, $\mathbf{M}_0 \in \mathcal{S}^q_{++}$ and suitable parameter $\gamma_0 > 0$. Set $y^0 = x^0$, $J^0 = \{0\}$, $e_0 = 0$, and set the counter $\ell = 0$, $k = 0$.

Step 1: Find multipliers (α_j^k, w^k) $(j \in J^\ell)$ that solve the dual problem (DP). Set $\hat{J}^\ell = \{j \in J^\ell : \alpha_j^k \neq 0\}$. Calculate

$$\tilde{g}^\ell = \sum_{j \in \hat{J}^\ell} \alpha_j^k g^j;$$

$$\varepsilon_\ell = \sum_{j \in \hat{J}^\ell} \alpha_j^k e_j; \qquad \text{(aggregate error)}$$

$$\delta_\ell = \varepsilon_\ell + \frac{1}{2\gamma_n} \|\tilde{g}^\ell\|^{*2}_{\mathbf{M}_k}, \qquad \text{(predicted decrease).}$$

Step 2: Set $y^{\ell+1} = x^k + \gamma_k^{-1} (\mathbf{A}^\top \mathbf{M}_k \mathbf{A})^{-1} (\mathbf{B}^\top w^k - \tilde{g}^\ell)$.

Bundle PAVM Algorithm

Step 3: IF (Descent test) $f(y^{\ell+1}) \leq f(x^k) - m\delta_\ell$,
THEN (Serious step)
set $x^{k+1} = y^{\ell+1}$. If x^{k+1} satisfies a given stopping rule, then stop.
Else, choose $g^{\ell+1} \in \partial f(x^{k+1})$.
Linearization error update

$$\begin{aligned} e_j &= e_j + f(x^{k+1}) - f(x^k) - \langle g^j, x^{k+1} - x^k \rangle, \quad \forall j \in J^\ell, \\ e_{\ell+1} &= 0. \end{aligned}$$

Update $\gamma_{k+1} > 0$ and $\mathbf{M}_{k+1}$. Replace k by $k+1$.
ELSE (Null step)
choose $g^{\ell+1} \in \partial f(y^{\ell+1})$.
Linearization error update

$$\begin{aligned} e_j &= e_j, \quad \forall j \in J^\ell, \\ e_{\ell+1} &= f(x^k) - f(y^{\ell+1}) + \langle g^{\ell+1}, y^{\ell+1} - x^k \rangle, \end{aligned}$$

Step 4: $J^{\ell+1} := \hat{J}^\ell \cup \{\ell+1\}$, increase ℓ by 1 and go to step 1.

References

F. Alizadeh and D. Goldfarb,
Second-order cone programming,
Math. Prog., vol. 95 (2003), pp. 3-51.

F. Alvarez, J. López and H. Ramírez C,
Interior proximal algorithm with variable metric for convex SOCP: Application to structural optimization and support vector machines,
Optimization Methods and Software, vol. 25 (2010), no 6, pp. 859-881.

R. Correa and C. Lemaréchal,
Convergence of some algorithms for convex minimization,
Math. Program., vol. 62 (1993), pp. 261-275.

O. Güler,
On the convergence of the proximal point algorithm for convex minimization,
SIAM J. Control Optim. 29(2) (1991), pp. 403Ű419.

G.L. Oliveira, S.S. Souza, J.X. da Cruz Neto and P.R. Oliveira,
Interior proximal methods for optimization over the positive orthant,
preprint, 2009.

J. Saketha Nath and C. Bhattacharyya,
Maximum margin classifiers with specified false positive and false negative error rates,
Proceedings of the Seventh SIAM International Conference on Data Mining, 2007.

THE END

THANKS FOR YOUR ATTENTION

Classification under uncertainty

Theorem (Multivariate Chebyshev Inequality)

Let x be a n-dimensional random variable with mean and covariance (μ, σ), where σ is a positive semidefinite symmetric matrix. Given $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ and $\eta \in [0, 1)$, the condition

$$\sup_{x \sim (\mu, \sigma)} \mathrm{Prob}\{a^\top x - b \geq 0\} \leq \eta$$

holds if and only if

$$b - a^\top \mu \geq \kappa(\eta)\sqrt{a^\top \sigma a},$$

where $\kappa(\eta) = \sqrt{\frac{1-\eta}{\eta}}$.