19 Lecture 19, Mar 8

Announcements

- HW6 (EM/MM, handwritten digit recognition revisited) this Fri Mar 11 @ 11:59PM.
- Solution sketches for HW1-5 are posted. `http://hua-zhou.github.io/teaching/biostatm280-2016winter/hwXXsol.html`. Substitute XX by 01, 02, ...
- Quiz 4 this Thu in class.
- Don't forget course evaluation: `http://my.ucla.edu`.

Last time

- Nonlinear conjugate gradient.
- Convex optimization: introduction, softwares.
- Linear programming (LP): introduction.

Today

- Linear programming (LP): more examples.
- Quadratic programming (QP).
- Second order cone programming (SOCP).
- Semidefinite programming (SDP).
- Geometric programming (GP).

Quadratic programming (QP)

- A *quadratic program* (QP) has quadratic objective function and affine constraint functions

$$
\begin{array}{ll}
\text{minimize} & (1/2)\boldsymbol{x}^T \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} + r \\
\text{subject to} & \boldsymbol{G}\boldsymbol{x} \preceq \boldsymbol{h} \\
& \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},
\end{array}
$$

  where we require $\boldsymbol{P} \in \mathbf{S}_+^n$ (why?). Apparently LP is a special case of QP with $\boldsymbol{P} = \mathbf{0}_{n\times n}$.

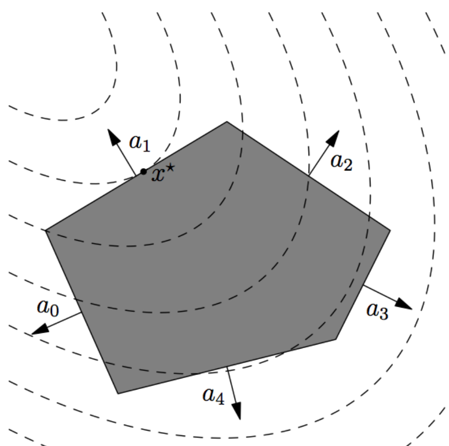


Figure 5.1: Geometric interpretation of quadratic optimization. At the optimal point $x^\star$ the hyperplane $\{x \mid a_1^T x = b\}$ is tangential to an ellipsoidal level curve.

- Example. The *least squares* problem minimizes $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$, which obviously is a QP.

- Example. Least squares with linear constraints. For example, *nonnegative least squares* (NNLS)

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\
\text{subject to} & \boldsymbol{\beta} \succeq \mathbf{0}.
\end{array}
$$

  ☞ In NNMF (nonnegative matrix factorization), the objective $\|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{W}\|_{\mathrm{F}}^2$ can be minimized by alternating NNLS.

- Example. Lasso regression (Tibshirani, 1996; Donoho and Johnstone, 1994) minimizes the least squares loss with ℓ_1 (lasso) penalty

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda \geq 0$ is a tuning parameter. Writing $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$, the equivalent QP is

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)^T\boldsymbol{X}^T\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\boldsymbol{X}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + \\
& \boldsymbol{y}^T\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\boldsymbol{X}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + \lambda\mathbf{1}^T(\boldsymbol{\beta}^+ + \boldsymbol{\beta}^-) \\
\text{subject to} \quad & \boldsymbol{\beta}^+ \succeq \mathbf{0}, \ \boldsymbol{\beta}^- \succeq \mathbf{0}
\end{aligned}$$

in $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$.

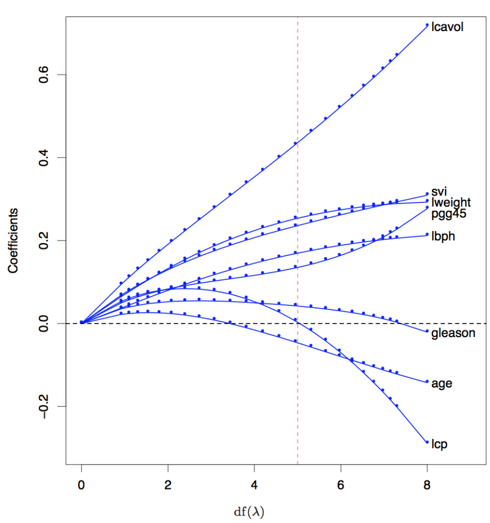


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

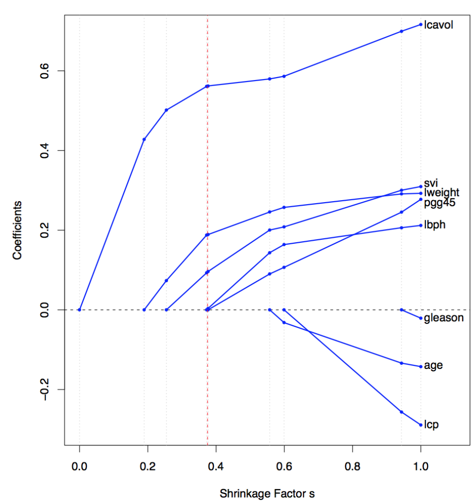


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

- Example: Elastic net (Zou and Hastie, 2005)

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2),$$

where $\lambda \geq 0$ and $\alpha \in [0, 1]$ are tuning parameters.

- Example: Generalized lasso

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1,$$

  where $\lambda \geq 0$ is a tuning parameter $\boldsymbol{D}$ is a fixed regularization matrix. This generates numerous applications (Tibshirani and Taylor, 2011).

- Example: Image denoising by anisotropic penalty. See `http://hua-zhou.github.io/teaching/st790-2015spr/ST790-2015-HW5.pdf`

- Example: (Linearly) constrained lasso

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ \text{subject to} & \boldsymbol{G}\boldsymbol{\beta} \preceq \boldsymbol{h} \\ & \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}, \end{array}$$

  where $\lambda \geq 0$ is a tuning parameter.

- Example: The Huber loss function

$$\phi(r) = \begin{cases} r^2 & |r| \leq M \\ M(2|r| - M) & |r| > M \end{cases}$$

  is commonly used in robust statistics. The robust regression problem

$$\text{minimize} \quad \sum_{i=1}^n \phi(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})$$

  can be transformed to a QP

$$\begin{array}{ll} \text{minimize} & \boldsymbol{u}^T\boldsymbol{u} + 2M\mathbf{1}^T\boldsymbol{v} \\ \text{subject to} & -\boldsymbol{u} - \boldsymbol{v} \preceq \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \preceq \boldsymbol{u} + \boldsymbol{v} \\ & \mathbf{0} \preceq \boldsymbol{u} \preceq M\mathbf{1}, \boldsymbol{v} \succeq \mathbf{0} \end{array}$$

  in $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^n$ and $\boldsymbol{\beta} \in \mathbf{R}^p$. Hint: write $|r_i| = (|r_i| \wedge M) + (|r_i| - M)_+ = u_i + v_i$.

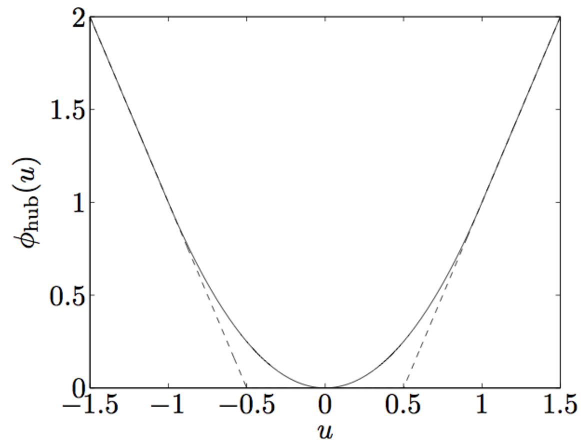


Figure 6.4 The solid line is the robust least-squares or Huber penalty function ϕ_{hub}, with $M = 1$. For $|u| \leq M$ it is quadratic, and for $|u| > M$ it grows linearly.

- Example: Support vector machines (SVM). In two-class classification problems, we are given training data $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i \in \mathbf{R}^n$ are feature vector and $y_i \in \{-1, 1\}$ are class labels. Support vector machine solves the optimization problem

$$\text{minimize} \quad \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \right]_+ + \lambda \|\boldsymbol{\beta}\|_2^2,$$

where $\lambda \geq 0$ is a tuning parameters. This is a QP.

Second-order cone programming (SOCP)

- A *second-order cone program* (SOCP)

$$\begin{array}{ll} \text{minimize} & \boldsymbol{f}^T\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i\|_2 \leq \boldsymbol{c}_i^T\boldsymbol{x} + d_i, \quad i = 1, \ldots, m \\ & \boldsymbol{F}\boldsymbol{x} = \boldsymbol{g} \end{array}$$

over $\boldsymbol{x} \in \mathbf{R}^n$. This says the points $(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i, \boldsymbol{c}_i^T\boldsymbol{x} + d_i)$ live in the second order cone (ice cream cone, Lorentz cone, quadratic cone)

$$\mathbf{Q}^{n+1} = \{(\boldsymbol{x}, t) : \|\boldsymbol{x}\|_2 \leq t\}$$

in $\mathbf{R}^{n+1}$.

☞ QP is a special case of SOCP. Why?

- When $\boldsymbol{c}_i = \mathbf{0}$ for $i = 1, \ldots, m$, SOCP is equivalent to a *quadratically constrained quadratic program* (QCQP)

$$\begin{array}{ll} \text{minimize} & (1/2)\boldsymbol{x}^T \boldsymbol{P}_0 \boldsymbol{x} + \boldsymbol{q}_0^T \boldsymbol{x} \\ \text{subject to} & (1/2)\boldsymbol{x}^T \boldsymbol{P}_i \boldsymbol{x} + \boldsymbol{q}_i^T \boldsymbol{x} + r_i \leq 0, \quad i = 1, \ldots, m \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \end{array}$$

  where $\boldsymbol{P}_i \in \mathbf{S}^n_+$, $i = 0, 1, \ldots, m$. Why?

- A *rotated quadratic cone* in $\mathbf{R}^{n+2}$ is

$$\mathbf{Q}_r^{n+2} = \{(\boldsymbol{x}, t_1, t_2) : \|\boldsymbol{x}\|_2^2 \leq 2t_1 t_2, t_1 \geq 0, t_2 \geq 0\}.$$

  A point $\boldsymbol{x} \in \mathbf{R}^{n+1}$ belongs to the second order cone $\mathbf{Q}^{n+1}$ if and only if

$$\begin{pmatrix} \boldsymbol{I}_{n-2} & 0 & 0 \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \boldsymbol{x}$$

  belongs to the rotated quadratic cone $\mathbf{Q}_r^{n+1}$.

  ☞ Gurobi allows users to input second order cone constraint and quadratic constraints directly.

  ☞ Mosek allows users to input second order cone constraint, quadratic constraints, and rotated quadratic cone constraint directly.

- Following sets are *(rotated) quadratic cone representable sets*:

  - (Absolute values) $|x| \leq t \Leftrightarrow (x, t) \in \mathbf{Q}^2$.
  - (Euclidean norms) $\|\boldsymbol{x}\|_2 \leq t \Leftrightarrow (\boldsymbol{x}, t) \in \mathbf{Q}^{n+1}$.
  - (Sume of squares) $\|\boldsymbol{x}\|_2^2 \leq t \Leftrightarrow (\boldsymbol{x}, t, 1/2) \in \mathbf{Q}_r^{n+2}$.
  - (Ellipsoid) For $\boldsymbol{P} \in \mathbf{S}^n_+$ and if $\boldsymbol{P} = \boldsymbol{F}^T \boldsymbol{F}$, where $\boldsymbol{F} \in \mathbf{R}^{n \times k}$, then

$$\begin{aligned} & (1/2)\boldsymbol{x}^T \boldsymbol{P}\boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x} + r \leq 0 \\ \Leftrightarrow \quad & \boldsymbol{x}^T \boldsymbol{P}\boldsymbol{x} \leq 2t, t + \boldsymbol{c}^T \boldsymbol{x} + r = 0 \\ \Leftrightarrow \quad & (\boldsymbol{F}\boldsymbol{x}, t, 1) \in \mathbf{Q}_r^{k+2}, t + \boldsymbol{c}^T \boldsymbol{x} + r = 0. \end{aligned}$$

Similarly,

$$\|\boldsymbol{F}(\boldsymbol{x}-\boldsymbol{c})\|_2 \le t \Leftrightarrow (\boldsymbol{y}, t) \in \mathbf{Q}^{n+1}, \boldsymbol{y} = \boldsymbol{F}(\boldsymbol{x}-\boldsymbol{c}).$$

☞ This fact shows that QP and QCQP are instances of SOCP.

– (Second order cones) $\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\|_2 \le \boldsymbol{c}^T\boldsymbol{x}+d \Leftrightarrow (\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}, \boldsymbol{c}^T\boldsymbol{x}+d) \in \mathbf{Q}^{m+1}$.

– (Simple polynomial sets)

$$\begin{array}{rcl}
\{(t,x) : |t| \le \sqrt{x}, x \ge 0\} & = & \{(t,x) : (t,x,1/2) \in \mathbf{Q}_r^3\} \\
\{(t,x) : t \ge x^{-1}, x \ge 0\} & = & \{(t,x) : (\sqrt{2}, x, t) \in \mathbf{Q}_r^3\} \\
\{(t,x) : t \ge x^{3/2}, x \ge 0\} & = & \{(t,x) : (x,s,t), (s,x,1/8) \in \mathbf{Q}_r^3\} \\
\{(t,x) : t \ge x^{5/3}, x \ge 0\} & = & \{(t,x) : (x,s,t), (s,1/8,z), (z,s,x) \in \mathbf{Q}_r^3\} \\
\{(t,x) : t \ge x^{(2k-1)/k}, x \ge 0\}, k \ge 2, & \text{can} & \text{be represented similarly} \\
\{(t,x) : t \ge x^{-2}, x \ge 0\} & = & \{(t,x) : (s,t,1/2), (\sqrt{2}, x, s) \in \mathbf{Q}_r^3\} \\
\{(t,x,y) : t \ge |x|^3/y^2, y \ge 0\} & = & \{(t,x,y) : (x,z) \in \mathbf{Q}^2, (z, y/2, s), (s, t/2, z) \in \mathbf{Q}_r^3\}
\end{array}$$

– (Geometric mean) The hypograph of the (concave) geometric mean function

$$\mathbf{K}_{\mathrm{gm}}^n = \{(\boldsymbol{x}, t) \in \mathbf{R}^{n+1} : (x_1 x_2 \cdots x_n)^{1/n} \ge t, \boldsymbol{x} \succeq \mathbf{0}\}$$

can be represented by rotated quadratic cones. See (Lobo et al., 1998) for derivation. For example,

$$\begin{array}{rcl}
\mathbf{K}_{\mathrm{gm}}^2 & = & \{(x_1, x_2, t) : \sqrt{x_1 x_2} \ge t, x_1, x_2 \ge 0\} \\
 & = & \{(x_1, x_2, t) : (\sqrt{2}t, x_1, x_2) \in \mathbf{Q}_r^3\}.
\end{array}$$

– (Harmonic mean) The hypograph of the harmonic mean function $\left(n^{-1} \sum_{i=1}^n x_i^{-1}\right)^{-1}$

can be represented by rotated quadratic cones

$$
\begin{aligned}
& \left(n^{-1}\sum_{i=1}^{n} x_i^{-1}\right)^{-1} \geq t, \boldsymbol{x} \succeq \mathbf{0} \\
\Leftrightarrow \quad & n^{-1}\sum_{i=1}^{n} x_i^{-1} \leq y, \boldsymbol{x} \succeq \mathbf{0} \\
\Leftrightarrow \quad & x_i z_i \geq 1, \sum_{i=1}^{n} z_i = ny, \boldsymbol{x} \succeq \mathbf{0} \\
\Leftrightarrow \quad & 2x_i z_i \geq 2, \sum_{i=1}^{n} z_i = ny, \boldsymbol{x} \succeq \mathbf{0}, \boldsymbol{z} \succeq \mathbf{0} \\
\Leftrightarrow \quad & (\sqrt{2}, x_i, z_i) \in \mathbf{Q}_r^3, \mathbf{1}^T \boldsymbol{z} = ny, \boldsymbol{x} \succeq \mathbf{0}, \boldsymbol{z} \succeq \mathbf{0}.
\end{aligned}
$$

– (Convex increasing rational powers) For $p, q \in \mathbf{Z}_+$ and $p/q \geq 1$,

$$\mathbf{K}^{p/q} = \{(x,t) : x^{p/q} \leq t, x \geq 0\} = \{(x,t) : (t\mathbf{1}_q, \mathbf{1}_{p-q}, x) \in \mathbf{K}_{\text{gm}}^{p}\}.$$

– (Convex decreasing rational powers) For any $p, q \in \mathbf{Z}_+$,

$$\mathbf{K}^{-p/q} = \{(x,t) : x^{-p/q} \leq t, x \geq 0\} = \{(x,t) : (x\mathbf{1}_p, t\mathbf{1}_q, 1) \in \mathbf{K}_{\text{gm}}^{p+q}\}.$$

– (Power cones) The *power cone* with rational powers is

$$\mathbf{K}_{\boldsymbol{\alpha}}^{n+1} = \left\{ (\boldsymbol{x}, y) \in \mathbf{R}_+^n \times \mathbf{R} : |y| \leq \prod_{j=1}^{n} x_j^{p_j/q_j} \right\},$$

where p_j, q_j are integers satisfying $0 < p_j \leq q_j$ and $\sum_{j=1}^{n} p_j/q_j = 1$. Let $\beta = \operatorname{lcm}(q_1, \ldots, q_n)$ and

$$s_j = \beta \sum_{k=1}^{j} \frac{p_k}{q_k}, \quad j = 1, \ldots, n-1.$$

Then it can be represented as

$$
\begin{aligned}
& |y| \leq (z_1 z_2 \cdots z_\beta)^{1/q} \\
& z_1 = \cdots = z_{s_1} = x_1, \quad z_{s_1+1} = \cdots = z_{s_2} = x_2, \quad z_{s_{n-1}+1} = \cdots = z_\beta = x_n.
\end{aligned}
$$

☞ References for above examples: Papers (Lobo et al., 1998; Alizadeh and Goldfarb, 2003) and book (Ben-Tal and Nemirovski, 2001, Lecture 3). Now our catalogue of SOCP terms includes all above terms.

☞ Most of these function are implemented as the built-in function in the convex optimization modeling language `cvx`.

- Example: Group lasso. In many applications, we need to perform variable selection at group level. For instance, in factorial analysis, we want to select or de-select the group of regression coefficients for a factor simultaneously. Yuan and Lin (2006) propose the group lasso that

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} w_g \|\boldsymbol{\beta}_g\|_2,$$

where $\boldsymbol{\beta}_g$ is the subvector of regression coefficients for group g, and w_g are fixed group weights. This is equivalent to the SOCP

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{X}^T\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\boldsymbol{X}\boldsymbol{\beta} + \\ & \quad \boldsymbol{y}^T\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\boldsymbol{X}\boldsymbol{\beta} + \lambda \sum_{g=1}^{G} w_g t_g \\ \text{subject to} & \|\boldsymbol{\beta}_g\|_2 \le t_g, \quad g = 1, \ldots, G, \end{array}$$

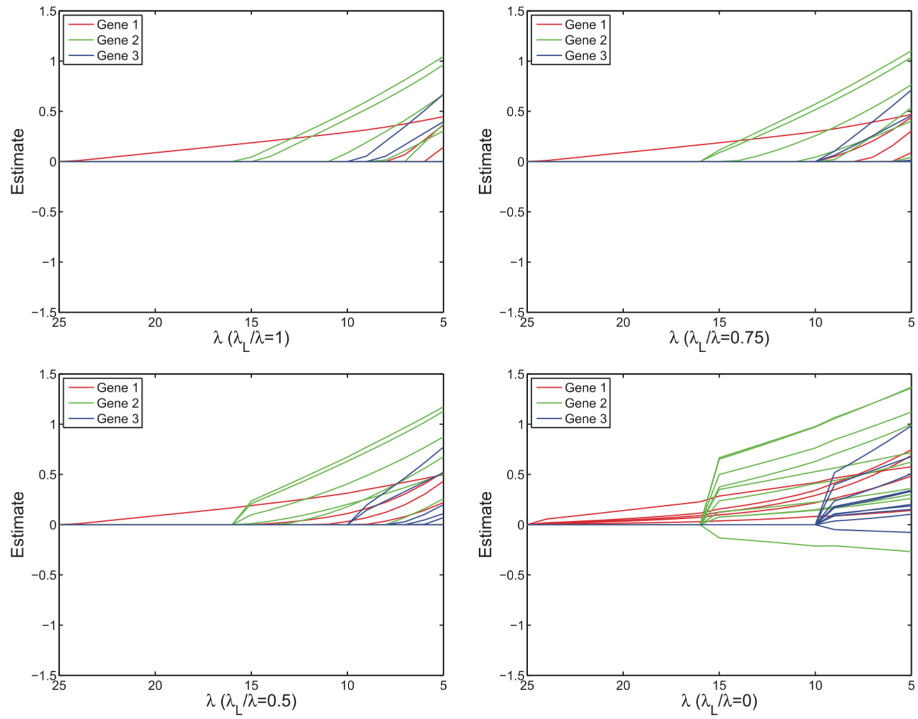
in variables $\boldsymbol{\beta}$ and $t_1, \ldots, t_G$.

☞ Overlapping groups are allowed here.

- Example. Sparse group lasso

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{g=1}^{G} w_g \|\boldsymbol{\beta}_g\|_2$$

achieves sparsity at both group and individual coefficient level and can be solved by SOCP as well.



☞ Apparently we can solve any previous loss functions (quantile, ℓ_1, composite quantile, Huber, multi-response model) plus group or sparse group penalty by SOCP.

- Example. Square-root lasso (Belloni et al., 2011) minimizes

$$\|\boldsymbol{y} - \beta_0 \mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_1$$

by SOCP. This variant generates the same solution path as lasso (why?) but simplifies the choice of λ.

A demo example: `http://hua-zhou.github.io/teaching/biostatm280-2016winter/lasso.html`

- Example: Image denoising by ROF model.

- Example. ℓ_p regression with $p \geq 1$ a rational number

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_p$$

can be formulated as a SOCP. Why? For instance, $\ell_{3/2}$ regression combines advantage of both robust ℓ_1 regression and least squares.

☞ `norm(x, p)` is a built-in function in the convex optimization modeling language `cvx` and `Convex.jl`.

Semidefinite programming (SDP)

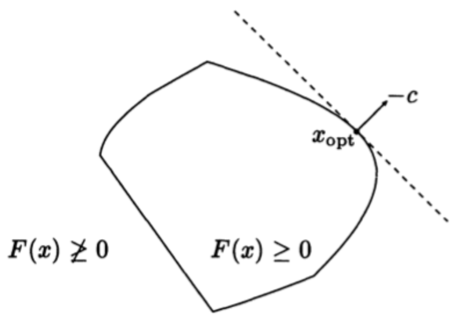


FIG. 1. *A simple semidefinite program with $x \in \mathbf{R}^2$, $F(x) \in \mathbf{R}^{7\times 7}$.*

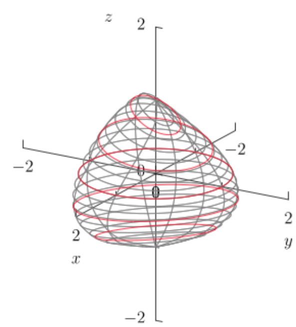


Figure 4.1: Plot of spectrahedron $S = \{(x, y, z) \in \mathbf{R}^3 \mid A(x, y, z) \succeq 0\}$.

- A *semidefinite program* (SDP) has the form

$$
\begin{array}{ll}
\text{minimize} & \boldsymbol{c}^T \boldsymbol{x} \\
\text{subject to} & x_1 \boldsymbol{F}_1 + \cdots + x_n \boldsymbol{F}_n + \boldsymbol{G} \preceq \mathbf{0} \quad \text{(LMI, linear matrix inequality)} \\
& \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},
\end{array}
$$

where $\boldsymbol{G}, \boldsymbol{F}_1, \ldots, \boldsymbol{F}_n \in \mathbf{S}^k$, $\boldsymbol{A} \in \mathbf{R}^{p\times n}$, and $\boldsymbol{b} \in \mathbf{R}^p$.

☞ When $\boldsymbol{G}, \boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$ are all diagonal, SDP reduces to LP.

- The *standard form SDP* has form

$$
\begin{array}{ll}
\text{minimize} & \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}) \\
\text{subject to} & \operatorname{tr}(\boldsymbol{A}_i \boldsymbol{X}) = b_i, \quad i = 1, \ldots, p \\
& \boldsymbol{X} \succeq \mathbf{0},
\end{array}
$$

where $\boldsymbol{C}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_p \in \mathbf{S}^n$.

- An *inequality form SDP* has form

$$\begin{array}{ll}\text{minimize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & x_1\boldsymbol{A}_1 + \cdots + x_n\boldsymbol{A}_n \preceq \boldsymbol{B},\end{array}$$

  with variable $\boldsymbol{x} \in \mathbf{R}^n$, and parameters $\boldsymbol{B}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_n \in \mathbf{S}^n$, $\boldsymbol{c} \in \mathbf{R}^n$.

- Exercise. Write LP, QP, QCQP, and SOCP in form of SDP.

- Example. Nearest correlation matrix. Let $\mathbf{C}^n$ be the convex set of $n \times n$ correlation matrices

$$\mathbf{C} = \{\boldsymbol{X} \in \mathbf{S}_+^n : x_{ii} = 1, i = 1, \ldots, n\}.$$

  Given $\boldsymbol{A} \in \mathbf{S}^n$, often we need to find the closest correlation matrix to $\boldsymbol{A}$

$$\begin{array}{ll}\text{minimize} & \|\boldsymbol{A} - \boldsymbol{X}\|_{\mathrm{F}} \\ \text{subject to} & \boldsymbol{X} \in \mathbf{C}.\end{array}$$

  This projection problem can be solved via an SDP

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & \|\boldsymbol{A} - \boldsymbol{X}\|_{\mathrm{F}} \le t \\ & \boldsymbol{X} = \boldsymbol{X}^T, \operatorname{diag}(\boldsymbol{X}) = \mathbf{1} \\ & \boldsymbol{X} \succeq \mathbf{0}\end{array}$$

  in variables $\boldsymbol{X} \in \mathbf{R}^{n\times n}$ and $t \in \mathbf{R}$. The SOC constraint can be written as an LMI

$$\begin{pmatrix} t\boldsymbol{I} & \operatorname{vec}(\boldsymbol{A} - \boldsymbol{X}) \\ \operatorname{vec}(\boldsymbol{A} - \boldsymbol{X})^T & t \end{pmatrix} \succeq \mathbf{0}$$

  by the Schur complement lemma.

- Eigenvalue problems. Suppose

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}_0 + x_1\boldsymbol{A}_1 + \cdots x_n\boldsymbol{A}_n,$$

  where $\boldsymbol{A}_i \in \mathbf{S}^m$, $i = 0, \ldots, n$. Let $\lambda_1(\boldsymbol{x}) \ge \lambda_2(\boldsymbol{x}) \ge \cdots \ge \lambda_m(\boldsymbol{x})$ be the ordered eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$.

– Minimize the maximal eigenvalue is equivalent to the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \boldsymbol{A}(\boldsymbol{x}) \preceq t\boldsymbol{I} \end{array}$$

in variables $\boldsymbol{x} \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

☞ Minimizing the sum of k largest eigenvalues is an SDP too. How about minimizing the sum of all eigenvalues?

☞ Maximize the minimum eigenvalue is an SDP as well.

– Minimize the spread of the eigenvalues $\lambda_1(\boldsymbol{x}) - \lambda_m(\boldsymbol{x})$ is equivalent to the SDP

$$\begin{array}{ll} \text{minimize} & t_1 - t_m \\ \text{subject to} & t_m\boldsymbol{I} \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq t_1\boldsymbol{I} \end{array}$$

in variables $\boldsymbol{x} \in \mathbf{R}^n$ and $t_1, t_m \in \mathbf{R}$.

– Minimize the *spectral radius* (or *spectral norm*) $\rho(\boldsymbol{x}) = \max_{i=1,\ldots,m} |\lambda_i(\boldsymbol{x})|$ is equivalent to the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t\boldsymbol{I} \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq t\boldsymbol{I} \end{array}$$

in variables $\boldsymbol{x} \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

– To minimize the condition number $\kappa(\boldsymbol{x}) = \lambda_1(\boldsymbol{x})/\lambda_m(\boldsymbol{x})$, note $\lambda_1(\boldsymbol{x})/\lambda_m(\boldsymbol{x}) \leq t$ if and only if there exists a $\mu > 0$ such that $\mu\boldsymbol{I} \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq \mu t\boldsymbol{I}$, or equivalently, $\boldsymbol{I} \preceq \mu^{-1}\boldsymbol{A}(\boldsymbol{x}) \preceq t\boldsymbol{I}$. With change of variables $y_i = x_i/\mu$ and $s = 1/\mu$, we can solve the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \boldsymbol{I} \preceq s\boldsymbol{A}_0 + y_1\boldsymbol{A}_1 + \cdots y_n\boldsymbol{A}_n \preceq t\boldsymbol{I} \\ & s \geq 0, \end{array}$$

in variables $\boldsymbol{y} \in \mathbf{R}^n$ and $s, t \geq 0$. In other words, we normalize the spectrum by the smallest eigenvalue and then minimize the largest eigenvalue of the normalized LMI.

– Minimize the ℓ_1 norm of the eigenvalues $|\lambda_1(\boldsymbol{x})|+\cdots+|\lambda_m(\boldsymbol{x})|$ is equivalent to the SDP

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(\boldsymbol{A}^+) + \operatorname{tr}(\boldsymbol{A}^-) \\ \text{subject to} & \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}^+ - \boldsymbol{A}^- \\ & \boldsymbol{A}^+ \succeq \boldsymbol{0}, \boldsymbol{A}^- \succeq \boldsymbol{0}, \end{array}$$

in variables $\boldsymbol{x} \in \mathbf{R}^n$ and $\boldsymbol{A}^+, \boldsymbol{A}^- \in \mathbf{S}^n_+$.

– Roots of determinant. The determinant of a semidefinite matrix $\det(\boldsymbol{A}(\boldsymbol{x})) = \prod_{i=1}^m \lambda_i(\boldsymbol{x})$ is neither convex or concave, but rational powers of the determinant can be modeled using linear matrix inequalities. For a rational power $0 \le q \le 1/m$, the function $\det(\boldsymbol{A}(\boldsymbol{x}))^q$ is concave and we have

$$\begin{aligned} & t \le \det(\boldsymbol{A}(\boldsymbol{x}))^q \\ \Leftrightarrow & \begin{pmatrix} \boldsymbol{A}(\boldsymbol{x}) & \boldsymbol{Z} \\ \boldsymbol{Z}^T & \operatorname{diag}(\boldsymbol{Z}) \end{pmatrix} \succeq \boldsymbol{0}, \quad (z_{11}z_{22}\cdots z_{mm})^q \ge t, \end{aligned}$$

where $\boldsymbol{Z} \in \mathbf{R}^{m\times m}$ is a lower-triangular matrix. Similarly for any rational $q > 0$, we have

$$\begin{aligned} & t \ge \det(\boldsymbol{A}(\boldsymbol{x}))^{-q} \\ \Leftrightarrow & \begin{pmatrix} \boldsymbol{A}(\boldsymbol{x}) & \boldsymbol{Z} \\ \boldsymbol{Z}^T & \operatorname{diag}(\boldsymbol{Z}) \end{pmatrix} \succeq \boldsymbol{0}, \quad (z_{11}z_{22}\cdots z_{mm})^{-q} \le t \end{aligned}$$

for a lower triangular $\boldsymbol{Z}$.

– Trace of inverse. $\operatorname{tr}\boldsymbol{A}(\boldsymbol{x})^{-1} = \sum_{i=1}^m \lambda_i^{-1}(\boldsymbol{x})$ is a convex function and can be minimized using SDP

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}\boldsymbol{B} \\ \text{subject to} & \begin{pmatrix} \boldsymbol{B} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{A}(\boldsymbol{x}) \end{pmatrix} \succeq \boldsymbol{0}. \end{array}$$

Note $\operatorname{tr}\boldsymbol{A}(\boldsymbol{x})^{-1} = \sum_{i=1}^m \boldsymbol{e}_i^T \boldsymbol{A}(\boldsymbol{x})^{-1}\boldsymbol{e}_i$. Therefore another equivalent formulation is

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m t_i \\ \text{subject to} & \boldsymbol{e}_i^T \boldsymbol{A}(\boldsymbol{x})^{-1}\boldsymbol{e}_i \le t_i. \end{array}$$

Now the constraints can be expressed by LMI

$$\boldsymbol{e}_i^T \boldsymbol{A}(\boldsymbol{x})^{-1} \boldsymbol{e}_i \leq t_i \Leftrightarrow \begin{pmatrix} \boldsymbol{A}(\boldsymbol{x}) & \boldsymbol{e}_i \\ \boldsymbol{e}_i^T & t_i \end{pmatrix} \succeq \boldsymbol{0}.$$

☞ See (Ben-Tal and Nemirovski, 2001, Lecture 4, p146-p151) for the proof of above facts.

☞ `lambda_max`, `lambda_min`, `lambda_sum_largest`, `lambda_sum_smallest`, `det_rootn`, and `trace_inv` are implemented in `cvx` for Matlab.

☞ `lambda_max`, `lambda_min` are implemented in `Convex.jl` package for Julia.

- Example. Experiment design. See HW6 Q1 `http://hua-zhou.github.io/teaching/st790-2015spr/ST790-2015-HW6.pdf`

- Singular value problems. Let $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}_0 + x_1 \boldsymbol{A}_1 + \cdots x_n \boldsymbol{A}_n$, where $\boldsymbol{A}_i \in \mathbf{R}^{p \times q}$ and $\sigma_1(\boldsymbol{x}) \geq \cdots \sigma_{\min\{p,q\}}(\boldsymbol{x}) \geq 0$ be the ordered singular values.

  – *Spectral norm* (or *operator norm* or *matrix-2 norm*) minimization. Consider minimizing the spectral norm $\|\boldsymbol{A}(\boldsymbol{x})\|_2 = \sigma_1(\boldsymbol{x})$. Note $\|\boldsymbol{A}\|_2 \leq t$ if and only if $\boldsymbol{A}^T \boldsymbol{A} \preceq t^2 \boldsymbol{I}$ (and $t \geq 0$) if and only if $\begin{pmatrix} t\boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & t\boldsymbol{I} \end{pmatrix} \succeq \boldsymbol{0}$. This results in the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \begin{pmatrix} t\boldsymbol{I} & \boldsymbol{A}(\boldsymbol{x}) \\ \boldsymbol{A}(\boldsymbol{x})^T & t\boldsymbol{I} \end{pmatrix} \succeq \boldsymbol{0} \end{array}$$

  in variables $\boldsymbol{x} \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

  ☞ Minimizing the sum of k largest singular values is an SDP as well.

  – Nuclear norm minimization. Minimization of the *nuclear norm* (or *trace norm*) $\|\boldsymbol{A}(\boldsymbol{x})\|_* = \sum_i \sigma_i(\boldsymbol{x})$ can be formulated as an SDP.

  Argument 1: Singular values of $\boldsymbol{A}$ coincides with the eigenvalues of the symmetric matrix

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{pmatrix},$$

which has eigenvalues $(\sigma_1, \ldots, \sigma_p, -\sigma_p, \ldots, -\sigma_1)$. Therefore minimizing the nuclear norm of $\boldsymbol{A}$ is same as minimizing the ℓ_1 norm of eigenvalues of the augmented matrix, which we know is an SDP.

Argument 2: An alternative characterization of nuclear norm is $\|\boldsymbol{A}\|_* = \sup_{\|\boldsymbol{Z}\|_2 \le 1} \operatorname{tr}(\boldsymbol{A}^T \boldsymbol{Z})$. That is

$$\begin{array}{ll} \text{maximize} & \operatorname{tr}(\boldsymbol{A}^T \boldsymbol{Z}) \\ \text{subject to} & \begin{pmatrix} \boldsymbol{I} & \boldsymbol{Z}^T \\ \boldsymbol{Z} & \boldsymbol{I} \end{pmatrix} \succeq \boldsymbol{0}, \end{array}$$

with the dual problem

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(\boldsymbol{U} + \boldsymbol{V})/2 \\ \text{subject to} & \begin{pmatrix} \boldsymbol{U} & \boldsymbol{A}(\boldsymbol{x})^T \\ \boldsymbol{A}(\boldsymbol{x}) & \boldsymbol{V} \end{pmatrix} \succeq \boldsymbol{0}. \end{array}$$

Therefore the epigraph of nuclear norm can be represented by LMI

$$\begin{aligned} & \|\boldsymbol{A}(\boldsymbol{x})\|_* \le t \\ \Leftrightarrow \quad & \begin{pmatrix} \boldsymbol{U} & \boldsymbol{A}(\boldsymbol{x})^T \\ \boldsymbol{A}(\boldsymbol{x}) & \boldsymbol{V} \end{pmatrix} \succeq \boldsymbol{0}, \quad \operatorname{tr}(\boldsymbol{U} + \boldsymbol{V})/2 \le t. \end{aligned}$$

Argument 3: See (Ben-Tal and Nemirovski, 2001, Proposition 4.2.2, p154).

☞ See (Ben-Tal and Nemirovski, 2001, Lecture 4, p151-p154) for the proof of above facts.

☞ `sigma_max` and `norm_nuc` are implemented in `cvx` for Matlab.

☞ `operator_norm` and `nuclear_norm` are implemented in `Convex.jl` package for Julia.

- Example. Matrix completion. See HW6 Q2 `http://hua-zhou.github.io/teaching/st790-2015spr/ST790-2015-HW6.pdf`

- Quadratic or quadratic-over-linear matrix inequalities. Suppose

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{x}) &= \boldsymbol{A}_0 + x_1 \boldsymbol{A}_1 + \cdots + x_n \boldsymbol{A}_n \\ \boldsymbol{B}(\boldsymbol{y}) &= \boldsymbol{B}_0 + y_1 \boldsymbol{B}_1 + \cdots + y_r \boldsymbol{B}_r. \end{aligned}$$

Then

$$
\begin{aligned}
& \boldsymbol{A}(\boldsymbol{x})^T \boldsymbol{B}(\boldsymbol{y})^{-1} \boldsymbol{A}(\boldsymbol{x}) \preceq \boldsymbol{C} \\
\Leftrightarrow \quad & \begin{pmatrix} \boldsymbol{B}(\boldsymbol{y}) & \boldsymbol{A}(\boldsymbol{x})^T \\ \boldsymbol{A}(\boldsymbol{x}) & \boldsymbol{C} \end{pmatrix} \succeq \mathbf{0}
\end{aligned}
$$

by the Schur complement lemma.

☞ `matrix_frac()` is implemented in both `cvx` for Matlab and `Convex.jl` package for Julia.

- General quadratic matrix inequality. Let $\boldsymbol{X} \in \mathbf{R}^{m \times n}$ be a rectangular matrix and

$$
F(\boldsymbol{X}) = (\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})^T + \boldsymbol{C}\boldsymbol{X}\boldsymbol{D} + (\boldsymbol{C}\boldsymbol{X}\boldsymbol{D})^T + \boldsymbol{E}
$$

be a quadratic matrix-valued function. Then

$$
\begin{aligned}
& F(\boldsymbol{X}) \preceq \boldsymbol{Y} \\
\Leftrightarrow \quad & \begin{pmatrix} \boldsymbol{I} & (\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})^T \\ \boldsymbol{A}\boldsymbol{X}\boldsymbol{B} & \boldsymbol{Y} - \boldsymbol{E} - \boldsymbol{C}\boldsymbol{X}\boldsymbol{D} - (\boldsymbol{C}\boldsymbol{X}\boldsymbol{D})^T \end{pmatrix} \preceq \mathbf{0}
\end{aligned}
$$

by the Schur complement lemma.

- Another matrix inequality

$$
\begin{aligned}
& \boldsymbol{X} \succeq \mathbf{0}, \boldsymbol{Y} \preceq (\boldsymbol{C}^T \boldsymbol{X}^{-1} \boldsymbol{C})^{-1} \\
\Leftrightarrow \quad & \boldsymbol{Y} \preceq \boldsymbol{Z}, \boldsymbol{Z} \succeq \mathbf{0}, \boldsymbol{X} \succeq \boldsymbol{C}\boldsymbol{Z}\boldsymbol{C}^T.
\end{aligned}
$$

See (Ben-Tal and Nemirovski, 2001, 20.c, p155).

- Cone of nonnegative polynomials. Consider nonnegative polynomial of degree $2n$

$$
f(t) = \boldsymbol{x}^T \boldsymbol{v}(t) = x_0 + x_1 t + \cdots x_{2n} t^{2n} \ge 0, \text{ for all } t.
$$

The cone

$$
\mathbf{K}^n = \{\boldsymbol{x} \in \mathbf{R}^{2n+1} : f(t) = \boldsymbol{x}^T \boldsymbol{v}(t) \ge 0, \text{ for all } t \in \mathbf{R}\}
$$

can be characterized by LMI

$$f(t) \geq 0 \text{ for all } t \Leftrightarrow x_i = \langle \boldsymbol{X}, \boldsymbol{H}_i \rangle, i = 0, \ldots, 2n, \boldsymbol{X} \in \mathbf{S}_+^{n+1},$$

where $\boldsymbol{H}_i \in \mathbf{R}^{(n+1)\times(n+1)}$ are Hankel matrices with entries $(\boldsymbol{H}_i)_{kl} = 1$ if $k+l = i$ or 0 otherwise. Here $k, l \in \{0, 1, \ldots, n\}$.

Similarly the cone of nonnegative polynomials on a finite interval

$$\mathbf{K}_{a,b}^n = \{\boldsymbol{x} \in \mathbf{R}^{n+1} : f(t) = \boldsymbol{x}^T \boldsymbol{v}(t) \geq 0, \text{ for all } t \in [a, b]\}$$

can be characterized by LMI as well.

– (Even degree) Let $n = 2m$. Then

$$\begin{aligned} \mathbf{K}_{a,b}^n &= \{\boldsymbol{x} \in \mathbf{R}^{n+1} : x_i = \langle \boldsymbol{X}_1, \boldsymbol{H}_i^m \rangle + \langle \boldsymbol{X}_2, (a+b)\boldsymbol{H}_{i-1}^{m-1} - ab\boldsymbol{H}_i^{m-1} - \boldsymbol{H}_{i-2}^{m-1} \rangle, \\ &\qquad i = 0, \ldots, n, \boldsymbol{X}_1 \in \mathbf{S}_+^m, \boldsymbol{X}_2 \in \mathbf{S}_+^{m-1}\}. \end{aligned}$$

– (Odd degree) Let $n = 2m+1$. Then

$$\begin{aligned} \mathbf{K}_{a,b}^n &= \{\boldsymbol{x} \in \mathbf{R}^{n+1} : x_i = \langle \boldsymbol{X}_1, \boldsymbol{H}_{i-1}^m - a\boldsymbol{H}_i^m \rangle + \langle \boldsymbol{X}_2, b\boldsymbol{H}_i^m - \boldsymbol{H}_{i-1}^m \rangle, \\ &\qquad i = 0, \ldots, n, \boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbf{S}_+^m\}. \end{aligned}$$

☞ References: paper (Nesterov, 2000) and the book (Ben-Tal and Nemirovski, 2001, Lecture 4, p157-p159).

- Example. Polynomial curve fitting. We want to fit a univariate polynomial of degree n

$$f(t) = x_0 + x_1 t + x_2 t^2 + \cdots x_n t^n$$

to a set of measurements (t_i, y_i), $i = 1, \ldots, m$, such that $f(t_i) \approx y_i$. Define the Vandermonde matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{pmatrix},$$

then we wish $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{y}$. Using least squares criterion, we obtain the optimal solution $\boldsymbol{x}_{\mathrm{LS}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{y}$. With various constraints, it is possible to find optimal $\boldsymbol{x}$ by SDP.

1. Nonnegativity. Then we require $\boldsymbol{x} \in \mathbf{K}_{a,b}^{n}$.

2. Monotonicity. We can ensure monotonicity of $f(t)$ by requiring that $f'(t) \geq 0$ or $f'(t) \leq 0$. That is $(x_1, 2x_2, \ldots, nx_n) \in \mathbf{K}_{a,b}^{n-1}$ or $-(x_1, 2x_2, \ldots, nx_n) \in \mathbf{K}_{a,b}^{n-1}$.

3. Convexity or concavity. Convexity or concavity of $f(t)$ corresponds to $f''(t) \geq 0$ or $f''(t) \leq 0$. That is $(2x_2, 2x_3, \ldots, (n-1)nx_n) \in \mathbf{K}_{a,b}^{n-2}$ or $-(2x_2, 2x_3, \ldots, (n-1)nx_n) \in \mathbf{K}_{a,b}^{n-2}$.

☞ `nonneg_poly_coeffs()` and `convex_poly_coeffs()` are implemented in `cvx`. Not in `Convex.jl` yet.