The Power of Duality

Aharon Ben-Tal

MINERVA Optimization Center
Technion – Israel Institute of Technology

WHAT IS “DUALITY THEORY”?

Kuhn (SIAM-AMS Proc., Vol. 9, 1976):

“A **Duality Theory** is made of the following elements:

(a) A pair of optimization problems, one a minimization problem (problem (P)) and the other a maximization problem (program (D)) based on the same data.

(b) Weak duality holds: $\min(P) \geq \max(D)$.

(c) Necessary and sufficient conditions for optimality of a feasible pair is the equality of the corresponding objective function.”

My addition

(d) A computable (tractable) relation between the optimal solutions of (P) and (D).

(e) Relation between: primal feasibility and dual boundedness and dual attainability.

LAGRANGIAN DUALITY

$$(P) \qquad \boxed{\inf_{x \in \mathbb{R}^n} \{f(x) \mid g_i(x) \le 0\,, \quad i = 1, \ldots, m\}}$$

f, g_i convex

Lagrangian: $L(x, y) = f(x) + \sum_{i=1}^{n} y_i g_i(x)$

Dual objective function:

$$\begin{aligned} g(y) &= \inf_{x \in \mathbb{R}^n} L(x, y) \quad \text{a concave function} \\ \mathrm{dom} g &= \{y \mid g(y) > -\infty\} \end{aligned}$$

$$(D) \qquad \boxed{\sup_{y \in \mathbb{R}^m} \{g(y) \mid y \in \mathrm{dom} g\,, \quad y \ge 0\}}$$

Duality Theorem If (P) satisfies

$$\exists\ \bar{x} : g_i(\bar{x}) < 0\ \forall\, i = 1, \ldots, m \quad \text{(Slater condition)}$$

then

$$\inf(P) = \max(D)$$

Extensions to ∞-dimensional spaces

Fenchel Duality

We define the primal problem:

$$(P) \qquad \inf\{f(x) - g(x) \mid x \in \mathrm{dom}(f) \cap \mathrm{dom}(g)\}\,.$$

The **Fenchel dual** of (P) is given by:

$$(D) \qquad \sup\{g_*(y) - f^*(y) \mid y \in \mathrm{dom}(g_*) \cap \mathrm{dom}(f^*)\}\,.$$

Now the **Fenchel duality theorem**:

Theorem *If ri dom$(f) \cap$ ri dom$(g) \neq \emptyset$, then the optimal values of (P) and (D) are equal and the maximal value of (D) is attained.*

Recall:

$$f^*(y) = \sup_x (y^T x - f(x)) \quad g_*(y) = \inf_x (y^T x - g(x))$$

CONIC DUALITY

K = closed convex pointed cone, int $K \neq \emptyset$

$$(P) \qquad \inf_{x \in \mathbb{R}^n} \left\{ c^T x \mid Ax - b \in K \right\} \qquad A = m \times n$$

(P) is ***strictly feasible*** if $\exists\, \bar{x} : A\bar{x} - b \in$ int K

The conic-dual of (P) is

$$(D) \qquad \sup_{y \in \mathbb{R}^m} \left\{ b^T y \mid A^T y = c,\ y \in K_* \right\}$$

K_* is the ***dual cone*** of K

$$K_* = \left\{ y \mid y^T x \geq 0, \quad \forall\, x \in K \right\}$$

(D) is strictly feasible if $\exists\, \bar{y} \in$ int $K_* : A^T \bar{y} = c$

Examples of dual cones

$$K = \mathbb{R}^n_+ , \qquad K_* = K$$

$$K = L_n = \left\{ (x_1, \ldots, x_{n-1}, x_n) \mid x_n \geq \sqrt{x_1^2 + \cdots + x_{n-1}^2} \right\}, \qquad K_* = K$$

$$K = S^n_+ = \left\{ A \in \mathbb{R}^{n \times n} \mid A^T = A,\ A \succeq 0 \right\}, \quad K_* = K$$

$$K = \{ x \mid Ax \geq 0 \} \quad K_* = \left\{ A^T y \mid y \geq 0 \right\}$$

SDP DUALITY

$$\boxed{(P) \qquad \min_{x \in \mathbb{R}^n} \{\langle c, x \rangle \mid \mathcal{A}x - B \succeq 0\}}$$

$$\mathcal{A}x = x_1 A_1 + \cdots + x_n A_n \quad \mathcal{A} : \mathbb{R}^n \to \mathbf{S}^m$$

$$A_i,\ B \in \mathbf{S}^m,$$

$$\text{Dual of } (P) : \max_{Y \in \mathbf{S}^m} \{\langle B, Y \rangle \mid \mathcal{A}^* Y = c,\ Y \succeq 0\}$$

$$\langle B, Y \rangle = \ \text{Trace}(BY)$$

$$\mathcal{A}^* : \mathbf{S}^m \to \mathbb{R}^n,\ \mathcal{A}^* Y = (\text{Tr}(YA_1), \ldots, \text{Tr}(YA_n))^T$$

$\Longrightarrow$

$$(D) \boxed{\max_{Y \in \mathbf{S}^m} \{\text{Tr}(BY) \mid \text{Tr}(YA_i) = c_i,\ i = 1, \ldots, n,\ Y \succeq 0\}}$$

CONIC DUALITY THEOREM

$$(P) \quad \inf\{c^T x \mid Ax - b \in K\}$$

$$(D) \quad \sup\{b^T y \mid A^T y = c,\ y \in K_*\}$$

1. Dual of $(D) = (P)$
2. Weak duality: $\inf(P) \geq \sup(D)$
3. Strong duality: If one of the (P) (D) is strictly feasible and bounded, then its dual is solvable and $\inf(P) = \sup(D)$. If both are strictly feasible, then $\min(P) = \max(D)$.
4. Complementarity: If (P) or (D) is strictly feasible and bounded, then a feasible pair x, y is an optimal pair, if and only if

$$y^T(Ax - b) = 0$$

EXAMPLE: A NONCONVEX QUADRATIC PROBLEM

$$(A) \qquad \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} z^T Q z + c^T z : \quad z^T z < 1 \right\}$$

$Q = n \times n$ symmetric indefinite

eigenvalues $\quad \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \ (\lambda_1 < 0)$

orthonormal e-vectors $\quad u_1 \leq u_2 \leq \cdots \leq u_n$

$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n); \quad P = (u_1, \ldots, u_n)$

change of variables: $x = Pz$ and using $Q = P^T \Lambda P$, $P^T P = I$,

problem (A) converted to:

$$(P) \qquad \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \, \Sigma \, \lambda_i x_i^2 + \bar{c}^T x : \ x^T x \leq 1 \right\}, \qquad \bar{c} = Pc\,.$$

Dual of (P) via lagrangian duality:

$$(D) \qquad \boxed{\max_{\mu \geq 0} \left\{ -\frac{1}{2} \left(\Sigma \, \frac{\bar{c}_i^2}{\lambda_i + u} + \mu \right) : \ \lambda_i + \mu \geq 0, \ \forall \, i \right\}}$$

a concave program

computing the *dual of the dual*.

Rewriting (D) as

$$(\hat{D}) \quad \begin{cases} \max\limits_{y \in \mathbb{R}^n, \mu \in \mathbb{R}} \left\{ -\frac{1}{2} \left(\Sigma \; \frac{\bar{c}_i^2}{y_i} + \mu \right) \right\} \\ \text{subject to} \\ -y_i + \lambda_i + \mu = 0 \qquad \leftarrow \; \text{multiplier } \frac{u_i}{2} \\ y_i \geq 0, \;\; \mu \geq 0 \end{cases}$$

The (lagrangian) dual of $(\hat{D})$ is

$$(DD) \quad \boxed{\begin{array}{l} \min \left\{ \frac{1}{2} \Sigma \lambda_i u_i - \Sigma |\bar{c}_i| \sqrt{u_i} \right\} \\ \text{subject to} \\ \qquad \Sigma u_i \leq 1, \;\; u_i \geq 0 \end{array}}$$

(DD) is, of course, a convex program.

What is the relation between (P) and (DD)?

$$\boxed{(P) \quad \begin{array}{c} \min \Sigma \left(\frac{1}{2} \lambda_i x_i^2 + \bar{c}_i^T x_i \right) \\ \Sigma \, x_i^2 \leq 1 \end{array}}$$

Theorem 1 *The nonconvex program (P) is equivalent to the convex program (DD).*

$$\{u_i^* : \ i = 1, \ldots, n\} \ \textit{solves} \ (DD) \ \textit{iff}$$
$$\left\{ x_i^* = -(\text{sign } \bar{c}_i)\sqrt{u_i^*}, \ i = 1, \ldots, n \right\} \ \textit{solves} \ (P)\,.$$

Proof:

$$\min(P) \leq \max(D) = \min(DD)$$

but $x_i^* = -(\text{sign } \bar{c}_i)\sqrt{u_i^*}$ is feasible to (P) and objective function of (P) evaluated at x^* is equal to $\min(DD)$.

EXAMPLE: STATISTICAL INFORMATION THEORY

X – random variable (nondegenerate)

B – support of X

f_X – density of X

D – class of density functions with support B which are absolutely contin. w.r.t. a nonnegative measure dt

$A_i(t)$ – summable functions

$$\boxed{\begin{array}{ll} (P) & \displaystyle\inf_{f\in D}\left\{ I(f,f_X) = \int_B f(t)\log\frac{f(t)}{f_X(t)}dt \right\} \\ & \text{s.t.} \\ & \qquad \int f(t)A_i(t)dt = \theta_i, \quad i=1,\ldots,m \end{array}}$$

∞-dimensional problem*

Fundamental Problem in *Statistical Information Theory Application* in Traffic Engineering, Accounting, Marketing, Signal Processing, Statistics ...

---

*$D \in L^p(\Omega, \mathcal{F}, P)$ linear space of measurable real-valued function $f:\Omega \to \mathbb{R}$, $\|f\|_p < \infty$
$\|f\|_p = \left(\int_\Omega |f(\omega)|^p dP(\omega)\right)^{1/p}$, $(1 < p < \infty)$

The Dual Problem

$$(D) \qquad \sup_{\pi \in \mathbb{R}^m} \left\{ \Sigma \theta_i \pi_i - \log \int_B f_X(t) e^{\Sigma A_i(t) \pi_i} dt \right\}$$

Unconstrained! **Finite Dimensional!**

Duality Theorem

(i) $\inf(P) = \sup(D)$

(ii) $\inf(P) = \min(P)$

(iii) $\sup(D) = \max(D)$ iff (P) is superconsistent

(iv) $\sup(D) < \infty \Leftrightarrow (P)$ is feasible
$\Rightarrow$

(v) If π^* solves (D), then

$$f^*(t) = \frac{f_X(t) e^{\Sigma A_i(t) \pi_i^*}}{\int_B f_X(t) e^{\Sigma A_i(t) \pi_i^*} dt}$$

solves (P)

* If $f_X(t)$ is not a density function, but just a positive summable function, then dual problem:

$$\sup_{\pi \in \mathbb{R}^m} \left\{ \Sigma \theta_i \pi_i - \int_B f_X(t) e^{\Sigma A_i(t)\pi_i - 1} dt \right\}$$

Examples

(1) $$\inf \left\{ \int_a^b f(t) \log f(t) dt : \int f(t) dt = 1 \right\}$$

sol. $$f^*(t) = \begin{cases} 1/(b-a) & a \leq t \leq b \\ 0 & \text{otherwise} \end{cases}$$ **UNIFORM**

(2) $$\inf_{f \in D} \left\{ \int_{-\infty}^{\infty} f(t) \log f(t) dt : \int_{-\infty}^{\infty} t^2 f(t) dt = \sigma^2 \right\}$$

sol. $$f^*(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$ **NORMAL**

(3) $$\inf_{f \in D} \int_0^{\infty} f(t) \log \frac{f(t)}{t^{K-1}} dt$$

$$\int t f(t) dt = K/\lambda$$

sol. $$f^*(t) = \frac{t^{K-1} e^{-\lambda t}}{\Gamma(K)}$$ **GAMMA**

LIST OF P.D.F.'s DERIVED FROM THE ABOVE DUALITY

DISCRETE R.V.	CONTINUOUS R.V.
uniform	Normal
geometric	Laplace
binomial	Generalized Cauchy
Poisson	exponential
log Series	gamma
Truncated Geometric	beta
⋮	log normal
	⋮

MULTIVARIATE R.V.

Multi Normal

Multi log Normal

Dirichlet

Multivariate Beta of 2nd kind

(Generalized) multivariate logistic

⋮

Fenchel Duality

We define the primal problem:

$$(P) \qquad \inf\{f(x) - g(x) \mid x \in \text{dom}(f) \cap \text{dom}(g)\}\,.$$

The **Fenchel dual** of (P) is given by:

$$(D) \qquad \sup\{g_*(y) - f^*(y) \mid y \in \text{dom}(g_*) \cap \text{dom}(f^*)\}\,.$$

Now the **Fenchel duality theorem**:

Theorem *If ri dom$(f) \cap$ ri dom$(g) \neq \emptyset$, then the optimal values of (P) and (D) are equal and the maximal value of (D) is attained.*

Recall:

$$f^*(y) = \sup_x (y^T x - f(x)) \quad g_*(y) = \inf_x (y^T x - g(x))$$

Indicator and Support Function

A special conjugate function is the support function, which is the conjugate of the indicator function.

The **indicator function** on the set S is defined as:

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

Then the conjugate function of $\delta(x|S)$,

$$\delta^*(y|S) = \sup_x \left\{ y^T x - \delta(x|S) \right\} = \sup_{x \in S} y^T x \,,$$

is the so-called **support function** of the set S.

The Essential Role of Duality in Robust Optimization

Consider the Robust Counterparts of an uncertain constraint

$$(RC) \qquad f(a,x) \le 0\,, \quad \forall\, a \in U, \quad f \text{ concave in } a$$

where

$$U = \left\{a = a^0 + A\zeta \,|\, \zeta \in Z \subset IR^L\right\}, \; Z \text{ a convex uncertainty set}$$

Theorem *The vector $x \in IR^n$ satisfies (RC) if and only if $x \in IR^n$ and $v \in IR^m$ satisfy the single inequality*

$$(FRC) \qquad (a^0)^T v + \delta^*(A^T v|Z) - f_*(v,x) \le 0\,.$$

$(\boldsymbol{FRC})$ $\boxed{\boldsymbol{(a^0)^T v + \delta^*(A^T v|Z) - f_*(v,x) \le 0}}$

we have that x is robust iff $f(a,x) \le 0\,, \quad \forall\, a \in U$,

or equivalently $F(x) \le 0$

$$\text{Now, } F(x) := \max_{a \in U} f(a,x)$$

$$F(x) \;=\; \max_{a \in R^m} \{f(a,x) - \delta(a|U)\}\,.$$

By **Fenchel duality**

$$F(x) = \min_{v \in R^m} \{\delta^*(v|U) - f_*(v,x)\}\,, \qquad (1)$$

where f_* is the concave conjugate w.r.t. $f(\cdot,x)$ and δ^* is the support function of U, i.e.:

$$\begin{aligned}\delta^*(v|U) &\sup_{\zeta \in Z}\{a^T v \,|\, a = a_0 + A\zeta\}\\ &= (a^0)^T v + \sup_{\zeta \in Z} v^T A\zeta\\ &= (a^0)^T v + \delta^*(A^T v|Z)\,.\end{aligned}$$

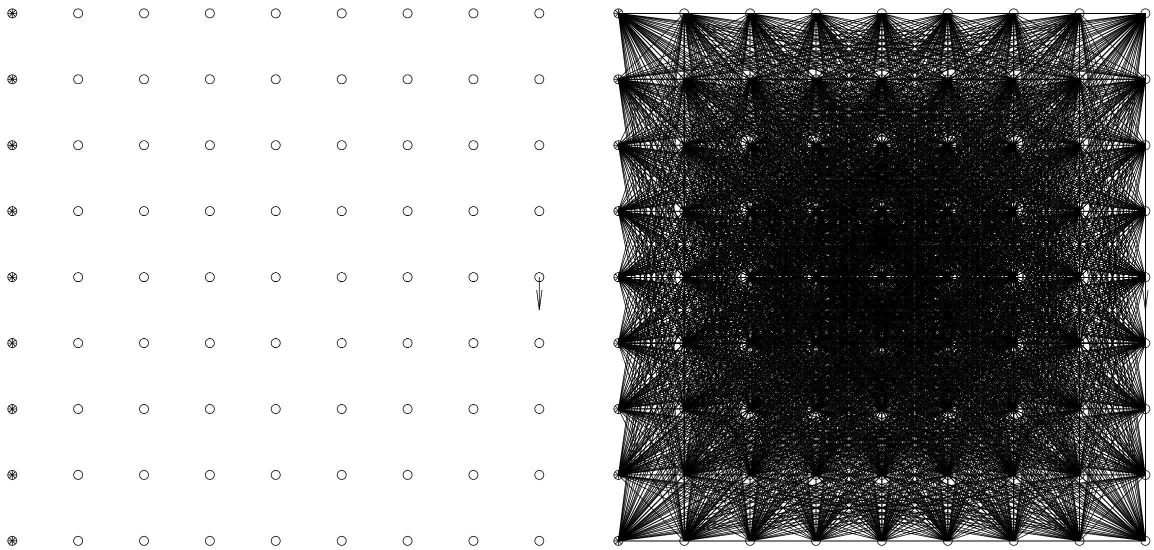
Putting things together in (1)

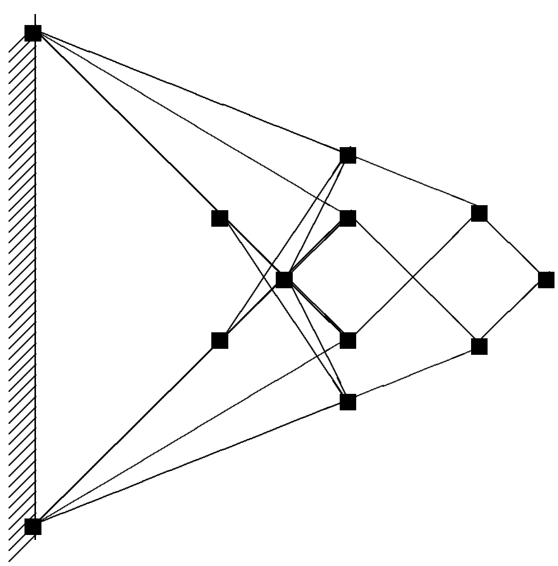
$$F(x) \le 0 \Longleftrightarrow \min_{v} (a^0)^T v + \delta^*(A^T v|Z) - f_*(v,x) \le 0$$

which is just the inequality (FRC) in the theorem.

THE POWER OF DUALITY IN STRUCTURAL DESIGN

Truss Topology Design





The simplest TTD problem is

$$\text{Compliance} = \tfrac{1}{2} f^T x \quad \to \quad \min$$

s.t.

$$\begin{array}{rcl}
\underbrace{\left[\sum_{i=1}^{m} t_i b_i b_i^T\right]}_{A(t)\succeq 0} x & = & f \\
\sum_{i=1}^{m} t_i & \leq & w \\
t & \geq & 0
\end{array}$$

- **Data:**

— $b_i \in \mathbb{R}^n$, n – **# of nodal degrees of freedom (for a** $10 \times 10 \times 10$ **ground structure,** $n \approx 3,000$**)**

— m – **# of tentative bars (for** $10 \times 10 \times 10$ **ground structure,** $m \approx 500,000$**)**

- **Design variables:** $t \in \mathbb{R}^m$, $x \in \mathbb{R}^n$

Multi-Load TTD

$$
\text{(P1)} \quad \left\{
\begin{array}{l}
\min\limits_{t \in R^m, x_1,\ldots,x_k \in R^n} \quad \max\limits_{j=1,\ldots,k} \{f_j^T x_j\} \\
\text{subject to} \\
\qquad\qquad A(t)x_j = f_j \qquad j = 1,\ldots,k \\
\qquad\qquad \sum_{i=1}^m t_i = v \ , \qquad t \geq 0
\end{array}
\right.
$$

$$
(A(t) = \sum t_i A_i, \ \ A_i \text{ sym. psd})
$$

Observation 1 *Let* $A \succeq 0$*, then*

$$
\min_x \{f^T x \mid Ax = f\} = \max_x \{2f^T x - x^T A x\} \ (= f^T A^{-1} f)
$$

Conclusion

$$
\text{(P1)} \Leftrightarrow \min_{\substack{\Sigma t_i = v \\ t \geq 0}} \ \max_j \left\{ \max_{x_j} (2f_j^T x_j - x_j^T A(t) x_j) \right\}
$$

Observation 2

$$
\left\{ \max_j \alpha_j = \max \left\{ \sum \lambda_j \alpha_j \mid \sum \lambda_j = 1, \lambda \geq 0 \right\} \right\}
$$

Conclusion

$$
\text{(P1)} \Leftrightarrow \min_{\substack{\Sigma t_i = v \\ t \geq 0}} \ \max_{\substack{\Sigma \lambda_j = 1 \\ \lambda_j \geq 0}} \left\{ \sum \lambda_j \max_{x_j} (2f_j^T x_j - x_j^T A(t) x_j) \right\}
$$

$$
= \min_{\substack{\Sigma t_i = v \\ t \geq 0}} \ \sup_{\substack{\Sigma \lambda_j = 1 \\ \lambda_j > 0}} \ \max_{x_j} \left\{ \sum 2f_j^T (\lambda_j x_j) - \lambda_j^{-1} (\lambda_j x_j)^T A(t) (\lambda_j x_j) \right\}
$$

Change of variables $\lambda_j x_j \to x_j$

$$(\text{P1}) \Leftrightarrow \underbrace{\min_{\substack{t\geq 0\\ \Sigma t_i = v}}}_{\substack{\uparrow\\ \text{compact}\\ \text{convex}}} \sup_{\substack{\lambda>0,\Sigma\lambda_j=1\\ \{x_j\}}} \underbrace{\left\{\sum 2f_j^T x_j - \lambda_j^{-1} x_j^T A(t) x_j\right\}}_{\substack{\uparrow\\ \text{linear in } t\\ \text{concave in}\\ \lambda_1,\ldots,\lambda_k,x_1,\ldots,x_k}}$$

$(f(\alpha_1,\alpha_2) = \alpha_1^2/\alpha_2$ is convex in $R \times R_+)$

MINMAX = MAX MIN

$$(\text{P1}) \Leftrightarrow \text{(dual sense)} \left[x_j A(t) x_j = \sum t_j x_j^T A_j x_j\right]$$

$$\sup_{\substack{\lambda>0,\Sigma\lambda_j=1\\ \{x_j\}}} \min_{\substack{t\geq 0\\ \Sigma t_j = v}} \left\{\sum 2f_j^T x_j - \sum t_j(\lambda_j^{-1} x_j^T A_j x_j)\right\}$$

$$t_i^* = \begin{cases} v & \text{if} \qquad i = \arg\max_j\{\lambda_j^{-1} x_j^T A_j x_j\} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{P1}) \Leftrightarrow \sup_{\substack{\lambda>0,\Sigma\lambda_j=1\\ \{x_j\}}} \left(2\sum f_j^T x_j - v \max_j\{\lambda_j^{-1} x_j A_j x_j\}\right)$$

summarizing

Multi-Load TTD

$$(\mathrm{P}) \qquad \min_{x_j, t} \max_{j=1,\ldots,k} \{f_j^T x_j\}$$

s.t.

$$\sum_{i=1}^{m} t_i A_i x_j = f_j$$

$$\sum_{i=1}^{m} t_i = v$$

$$t_i = 0\,, \qquad i = 1, \ldots, m$$

$$x^j \in I\!R^n, \quad j = 1, \ldots, k$$

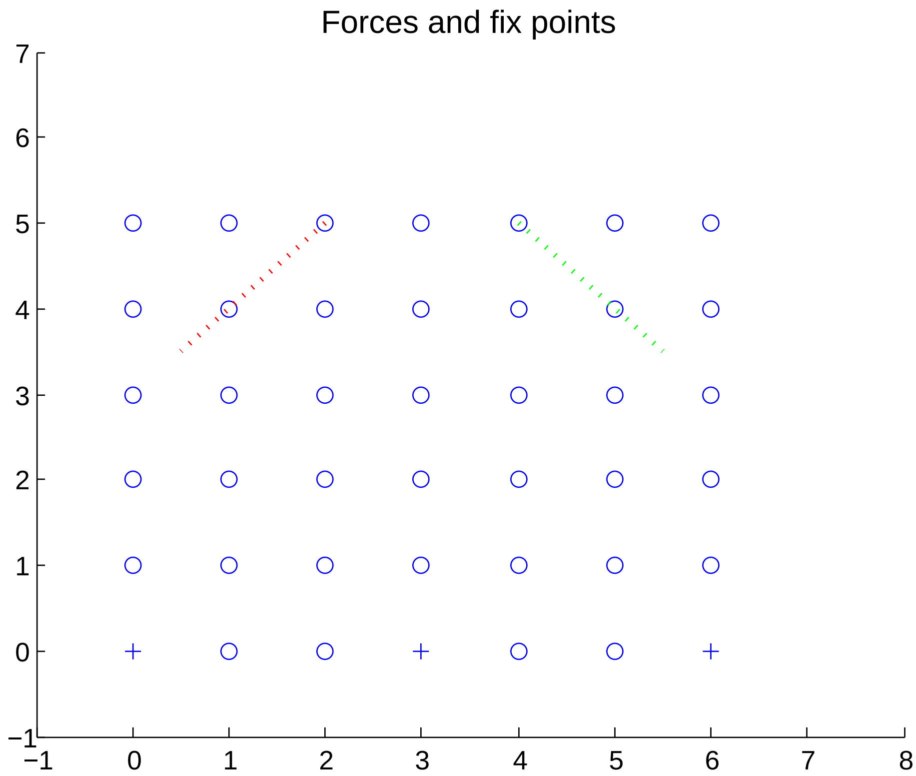
$$\boxed{\begin{array}{l} \displaystyle\min_{\substack{x_j,\ldots,x_k \\ \lambda \in R^k}} \left\{ \sum_{j=1}^{k} f_j^T x_j + v \cdot \max_{i=1,\ldots,m} \left\{ \sum_{j=1}^{K} \frac{x_j^T A_i x_j}{\lambda_j} \right\} \right\} \\ \text{s.t.} \\ \displaystyle\qquad\qquad \sum_{j=1}^{k} \lambda_j = 1, \quad \lambda \geq 0\,. \end{array}}$$

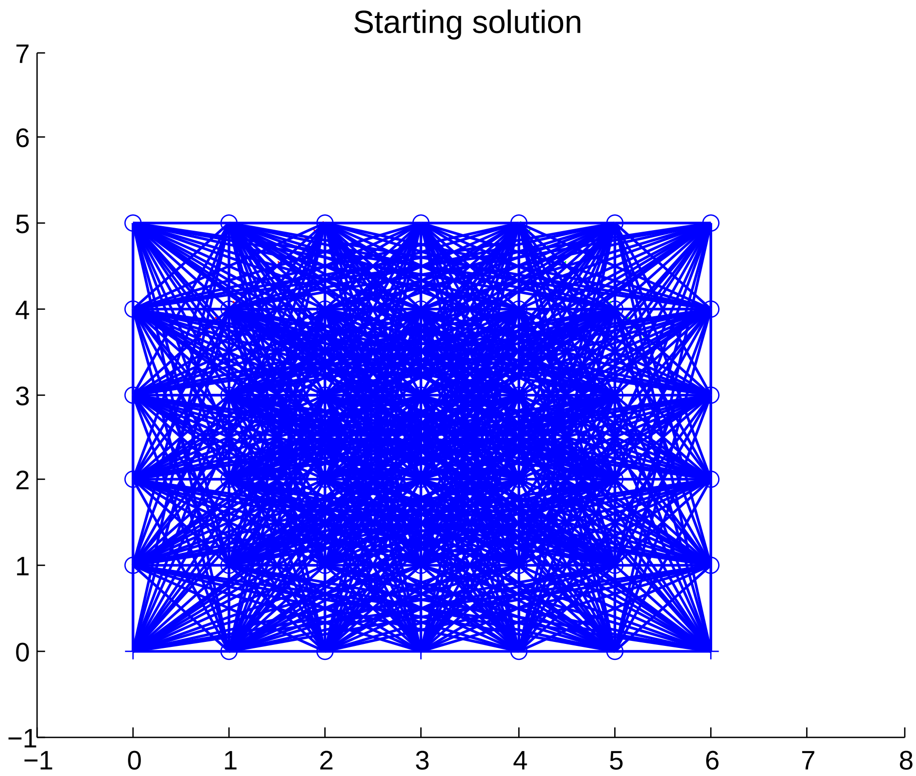
CONVEX PROGRAM

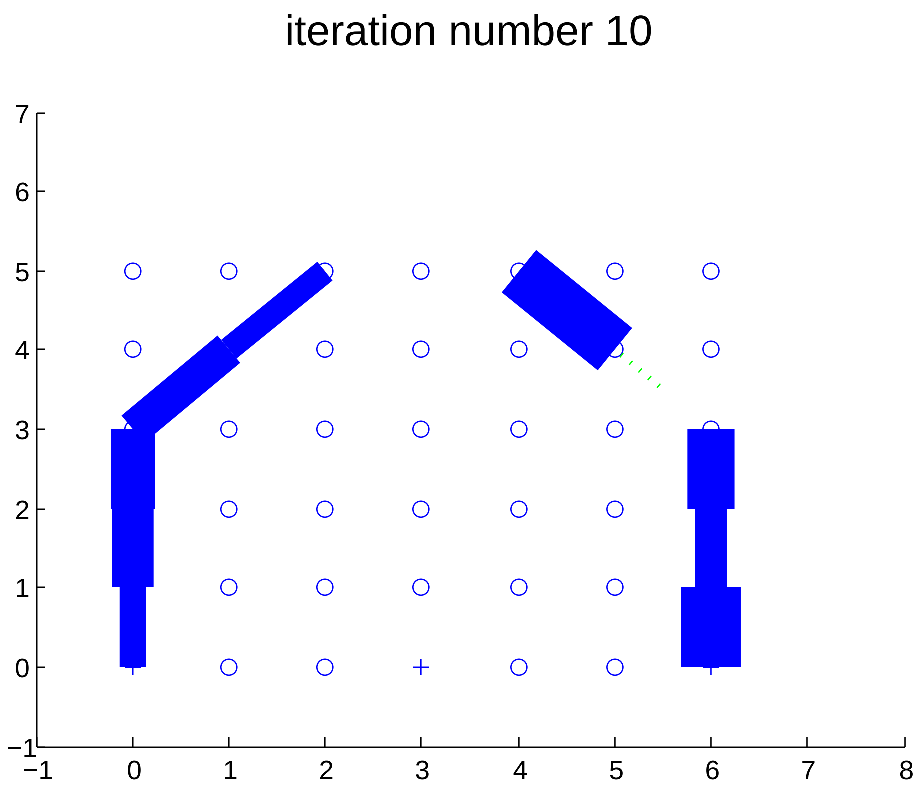
$$\min_{\substack{x_1,\ldots,x_k \\ \lambda \in R^k \\ \tau \in R}} \left\{ \sum f_j^T x_j + v\tau \right\}$$

$$\sum \lambda_j = 1, \quad \lambda \geq 0$$

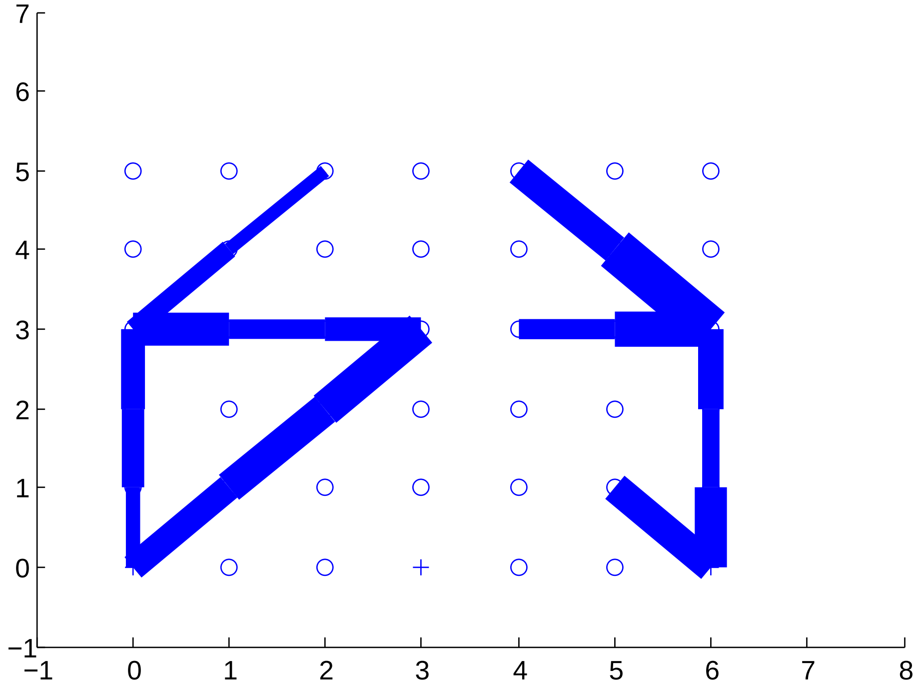
$$\begin{array}{c}\text{dual var.} \\ t_j^*\end{array} \quad \Rightarrow \quad \underbrace{\sum_j \left(\frac{x_j^T A_j x_j}{\lambda_j} \right) \leq \tau}_{\text{a conic quadratic constraint}} \qquad \forall\, i$$

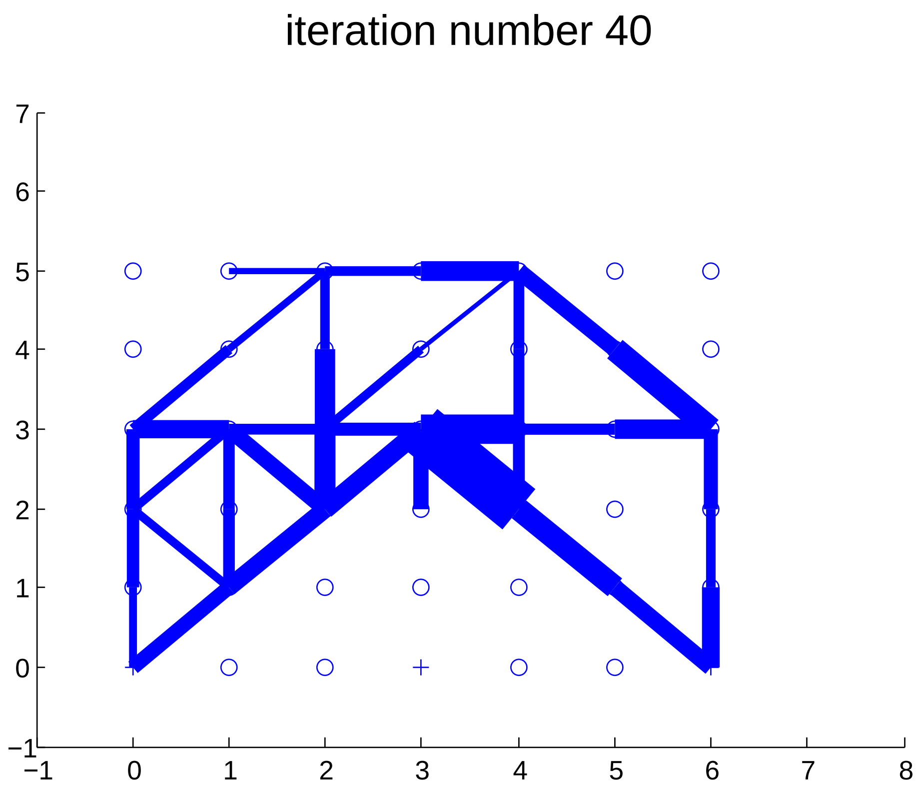

Forces and fix points
7
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
7
8


Starting solution
7
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
7
8

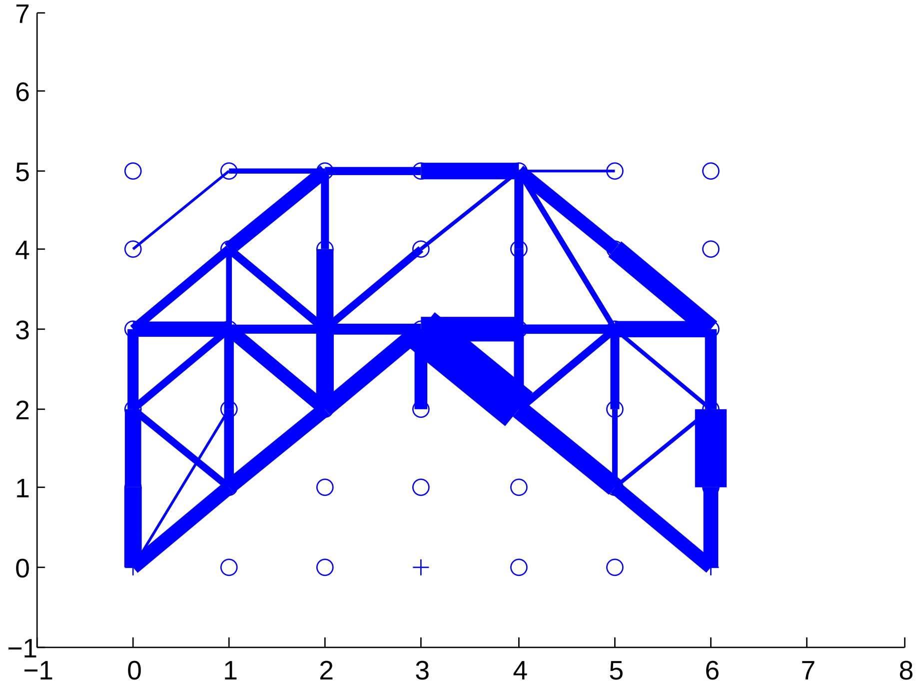

iteration number 10
7
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
7
8

iteration number 20

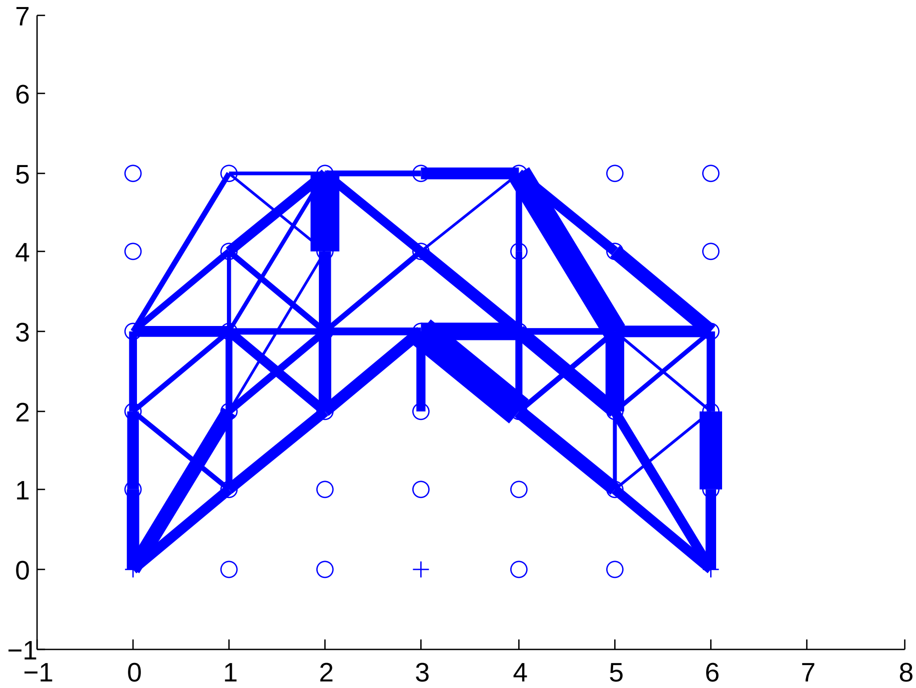



iteration number 40
7
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
7
8

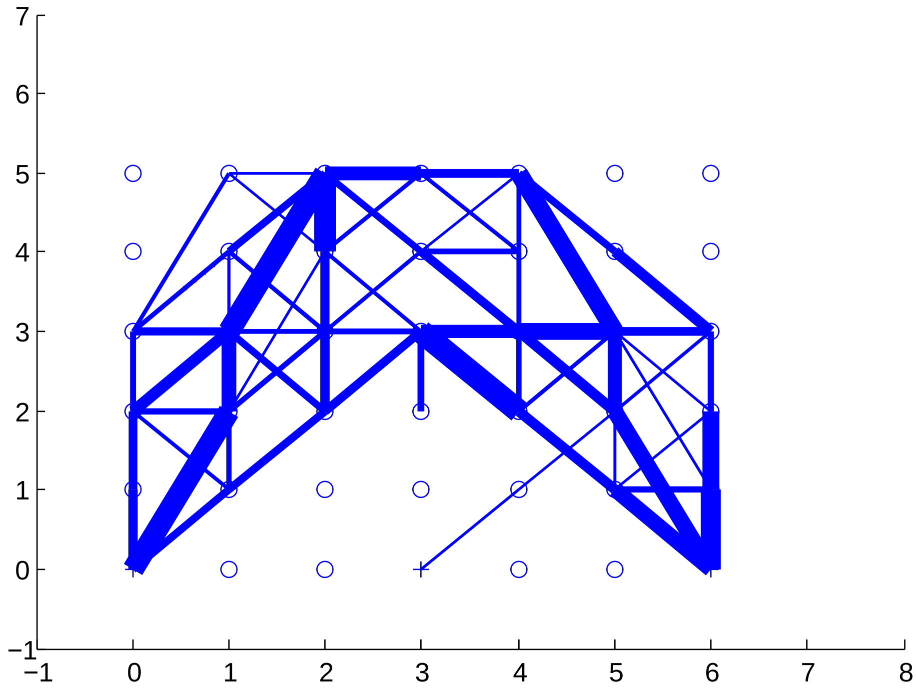
iteration number 60



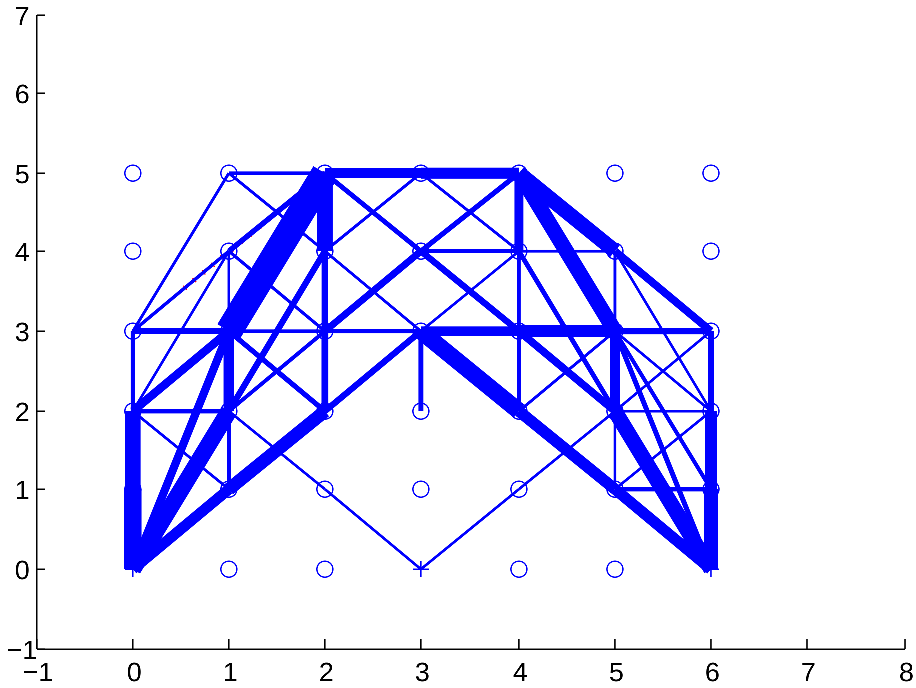
iteration number 80



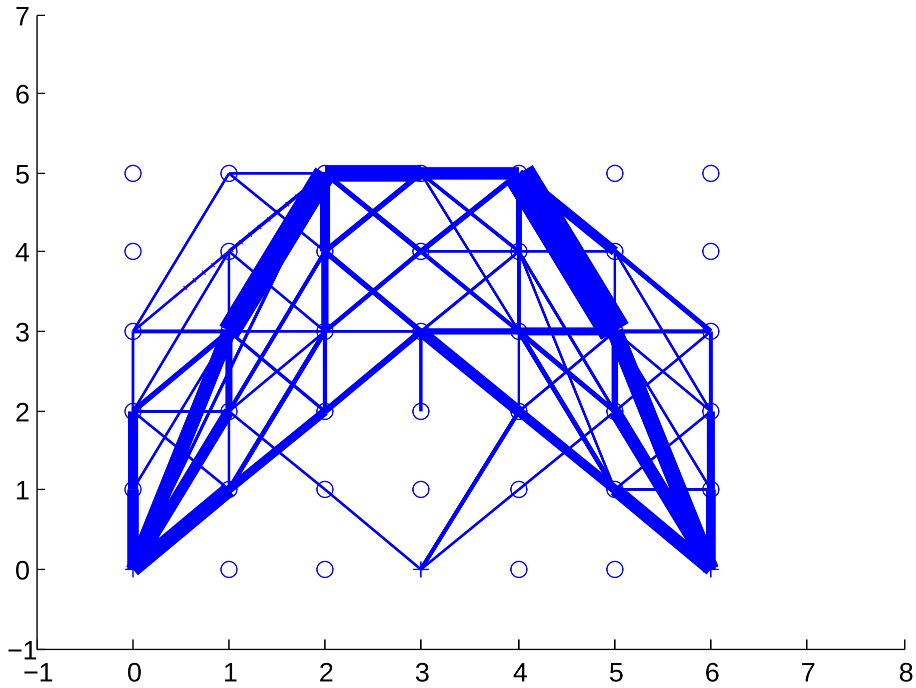
iteration number 100



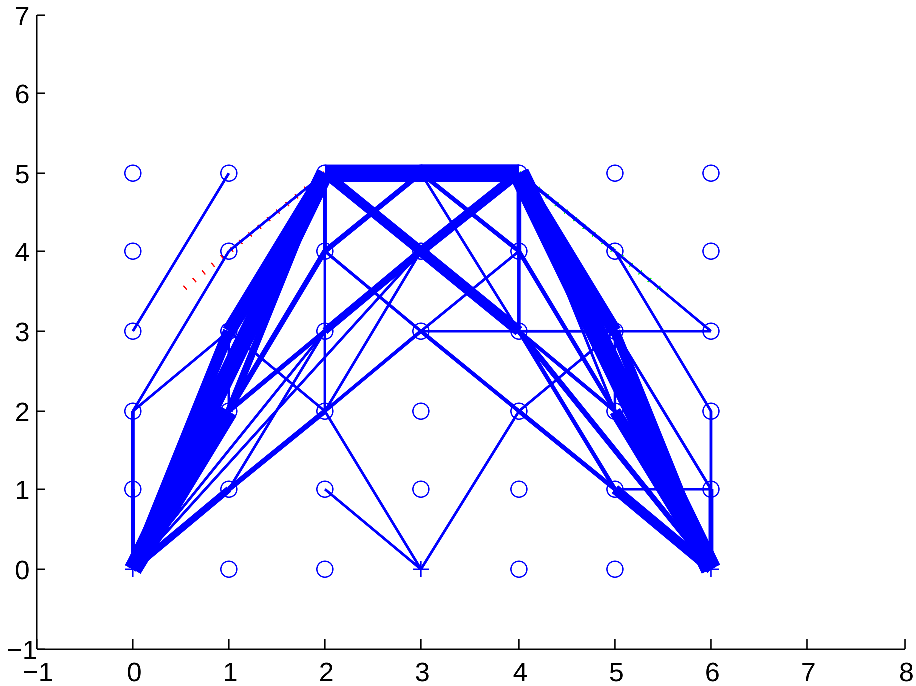
iteration number 150



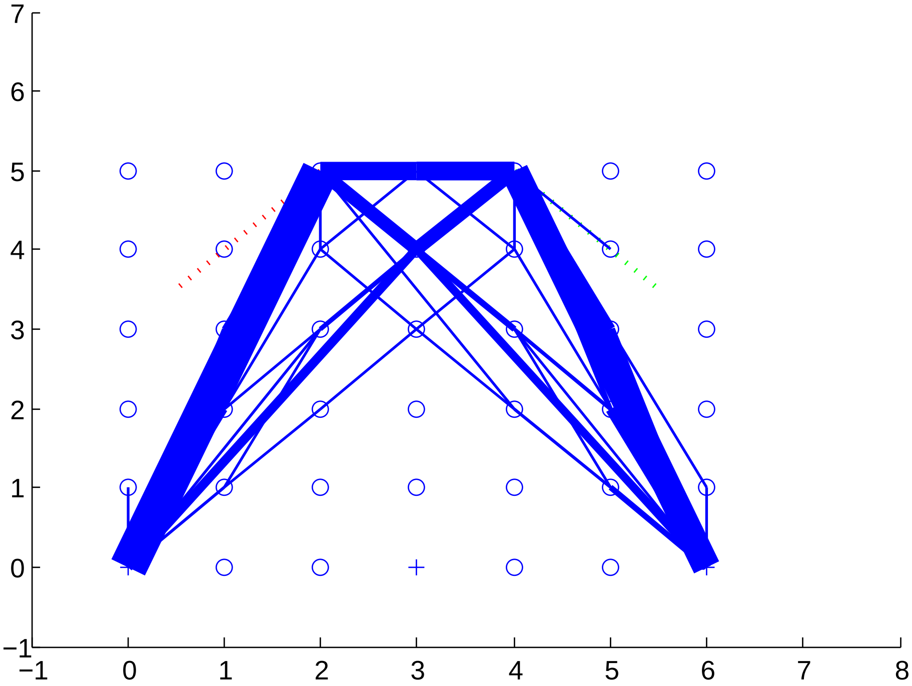
iteration number 200



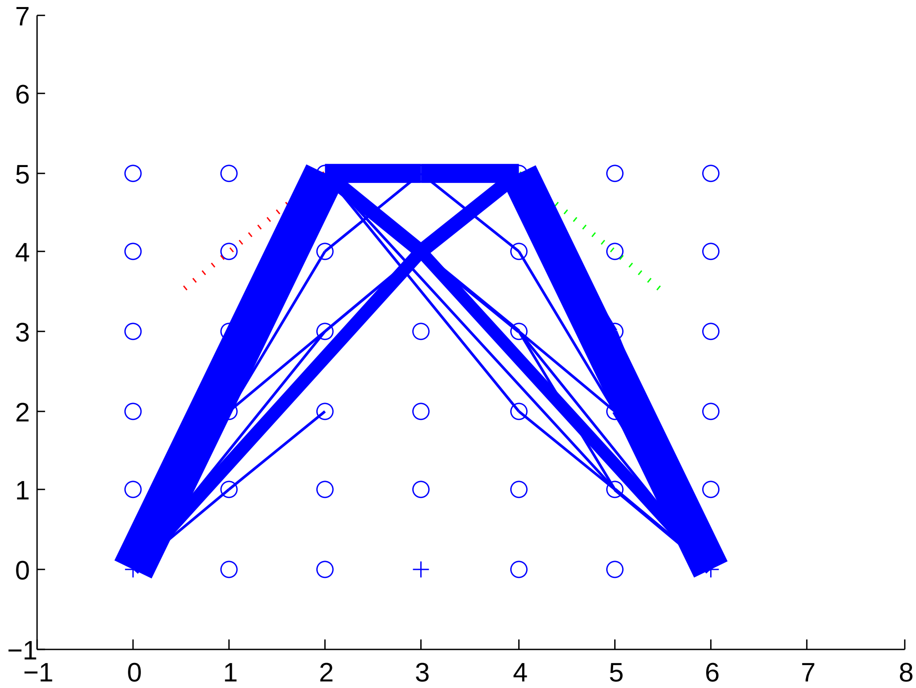
iteration number 400



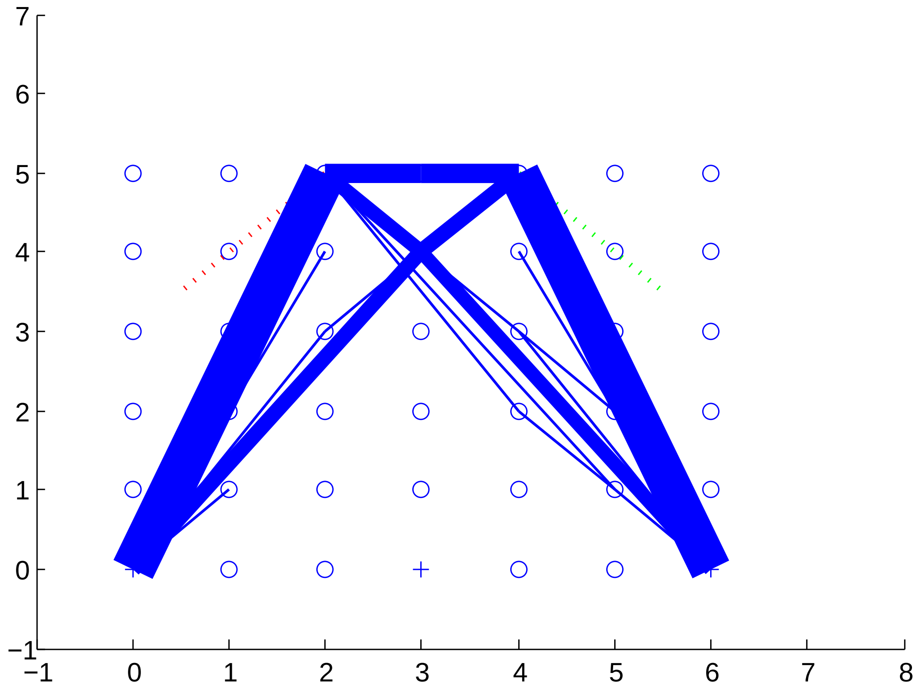
iteration number 600

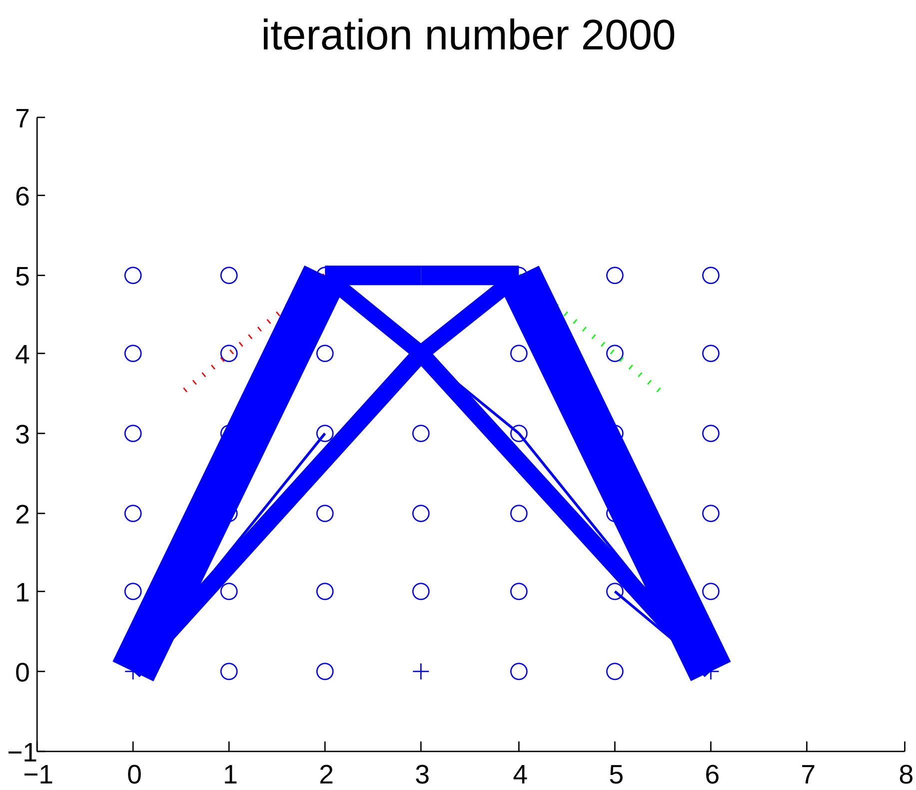


iteration number 800



iteration number 1000





iteration number 2000
7
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
7
8

ANALYSIS BASED ON II (Conic Optimization)

$$\boxed{\min_{t\in T} \tfrac{1}{2} f^T A(t)^{-1} f} \qquad A(t) = \Sigma t_i A_i \quad A_i = b_i b_i^T$$

$$T = \{t \geq 0,\ \Sigma t_i \leq w\}$$

$$\Leftrightarrow \quad \begin{array}{ll} \min\limits_{t\in T,\tau} & \tfrac{1}{2}\tau \\ & f^T A(t)^{-1} f \leq \tau \end{array}$$

$$\left[\begin{array}{l} \underline{\textbf{Schur Complement}}: \\ \exists D^{-1}, \begin{pmatrix} B & C^T \\ C & D \end{pmatrix} \succcurlyeq 0 \ \Leftrightarrow\ B - CD^{-1}C^T \succcurlyeq 0 \end{array}\right]$$

$$\Leftrightarrow \quad \boxed{\begin{array}{l} \min\limits_{t,\tau\in R} \tfrac{1}{2}\tau \\ \quad \begin{pmatrix} \tau & f^T \\ f & A(t) \end{pmatrix} \succcurlyeq 0 \\ \quad \Sigma t_i \leq w \\ \qquad t_i \geq 0 \end{array}}$$

SEMIDEFINITE PROGRAM

(SDP)

- In the case of Shape design, the advantages of the dual setting also can be quite significant.

  Consider, e.g., the obstacle-free planar Shape problem with rectangular cells and with simple bounds.

- The primal problem is:

$$
\begin{array}{rcl}
\tau & \to & \min \\
\left(\begin{array}{cc} 2\tau & -f_\ell^T \\ -f_\ell & \sum_{i=1}^n \sum_{s=1}^4 b_{is} t_i b_{is}^T \end{array}\right) & \succeq & 0,\ \ell = 1,\ldots,k\,; \\
t_i & \succeq & 0,\ i = 1,\ldots,n\,; \\
\sum\limits_{i=1}^n \mathrm{Tr}(t_i) & \leq & w \\
\left[\tau \in \mathbb{R},\, t_i \in \mathbf{S}^3\right] & &
\end{array}
$$

- The dual problem is

$$-2\sum_{\ell=1}^{k} f_\ell^T v_\ell + w\gamma \to \min$$

$$\left(\begin{array}{ccc|c|ccc|c}
\alpha_1 & & & & & & & v_1^T b_{i1} \\
 & \ddots & & & & & & \cdots \\
 & & \alpha_1 & & & & & v_1^T b_{iS} \\
\hline
 & & & \ddots & & & & \cdots \\
\hline
 & & & & \alpha_k & & & v_k^T b_{i1} \\
 & & & & & \ddots & & \cdots \\
 & & & & & & \alpha_k & v_k^T b_{iS} \\
b_{i1}^T v_1 & \cdots & b_{iS}^T v_1 & \cdots & b_{i1}^T v_k & \cdots & b_{iS}^T v_k & \gamma I_3
\end{array}\right) \succeq 0,\ i = 1, \ldots, n;$$

$$2\sum_{\ell=1}^{k} \alpha_\ell = 1\,.$$

$$[\alpha_\ell,\ \gamma \in \mathbb{R},\ v_\ell \in \mathbb{R}^m]$$

- E.g., for planar shape with 14×14 cells and 3 loads:

Setting	Design dimension	Effort of analyzing LMI's at a point, a.o.
(Pr)	1,177	37,309,230
$(D1)$	1,264	71,608

Optimization in Flight...

Application Example: Free Material Optimization

♣ FMO is a methodology for design of mechanical structures. In FMO, one seeks how to distribute a given amount of elastic material over a given domain in order to get a structure capable to withstand best of all a given collection of external loads. It is assumed that

- the mechanical properties of the material (its rigidity tensor) may vary, in an arbitrary fashion, from point to point;
- the rigidity of a construction w.r.t. a given external load is measured by the compliance – potential energy capacitated by the construction at the static equilibrium corresponding to the load;

♠ The goal is, given the weight of the construction, to minimize its largest, over a given set of loading scenarios, compliance.

♣ Usually it is technically impossible or too expensive to implement an FMO design "as it is". The role of FMO is in providing a good guess for the structure of the would-be construction. After the structure is guessed, the construction is designed from traditional materials via standard engineering techniques.

♣ **With Finite Element discretization, the Multi-Load FMO problem is**

$$\min_t \left\{ \max_{\ell=1,\dots,K} f_\ell^T S^{-1}(t) f_\ell : t_i \succeq 0, \sum_i \mathrm{Tr}(t_i) \le 1 \right\} \qquad \textbf{(FMO)}$$

where

- t_i, $i = 1, \dots, N$, **are symmetric** 3×3 **(in 2D) or** 6×6 **(in 3D) variable matrices (rigidity tensors of the material in Finite Element cells),**
- f_ℓ, $\ell = 1, \dots, K$, **are** M**-dimensional data vectors representing loading scenarios,**
- $S(t) = \sum_{i,s} b_{is}^T t_i b_{is}$ **is the** $M \times M$ **<u>stiffness matrix</u> of the construction.**

♣ **In a realistic 2D FMO problem,**

- **the number** N **of Finite Element cells is tens of thousands**

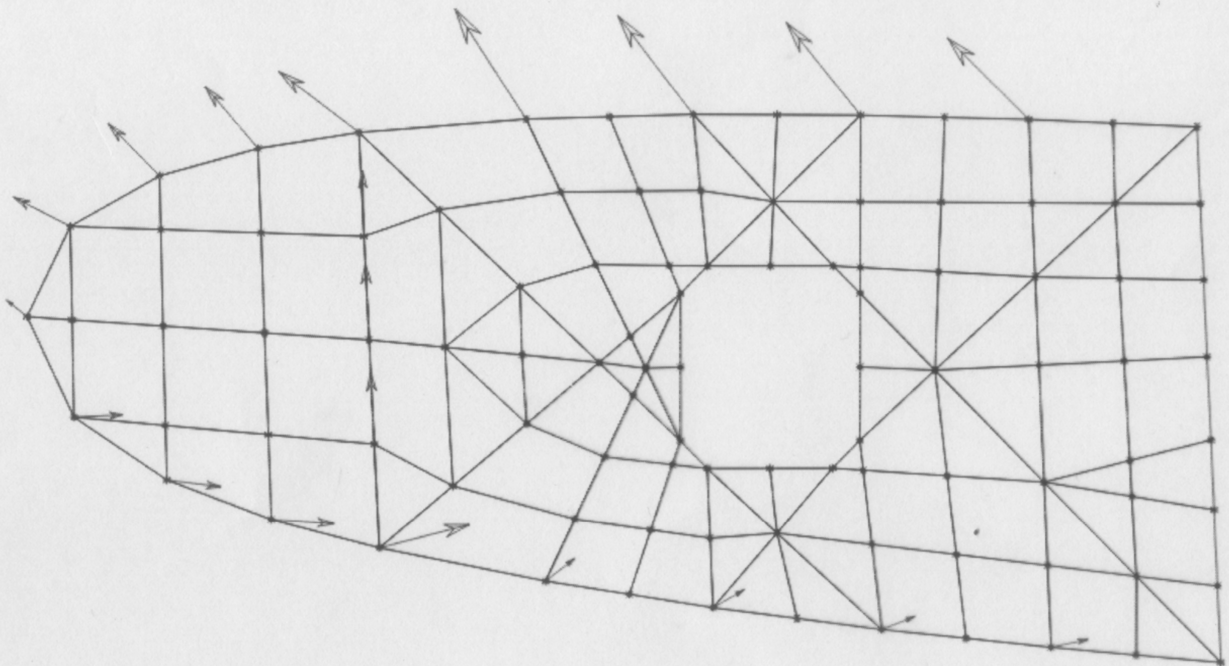
$\Rightarrow$ **design dimension of (FMO) is of order of 50,000 – 200,000**

- **the size** M **of the stiffness matrix is** $\approx 2N$

$\Rightarrow$ **it is a nontrivial problem just to compute the objective!**

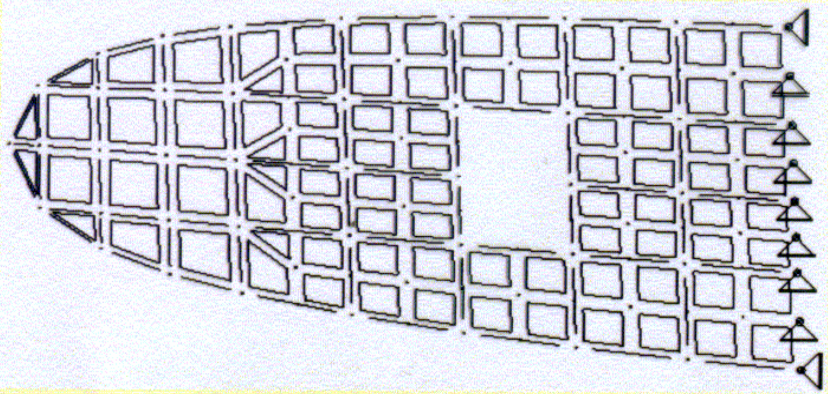
Design of stiffeners: MOPED & MBB-LAGRANGE

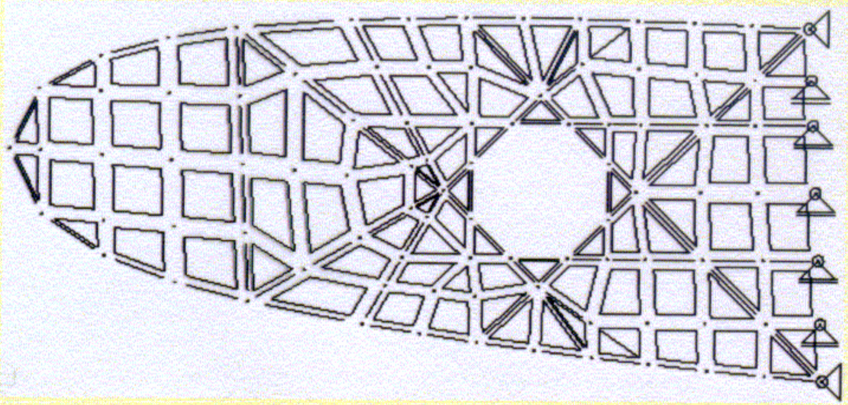
Design of stiffeners: MOPED & MBB-LAGRANGE

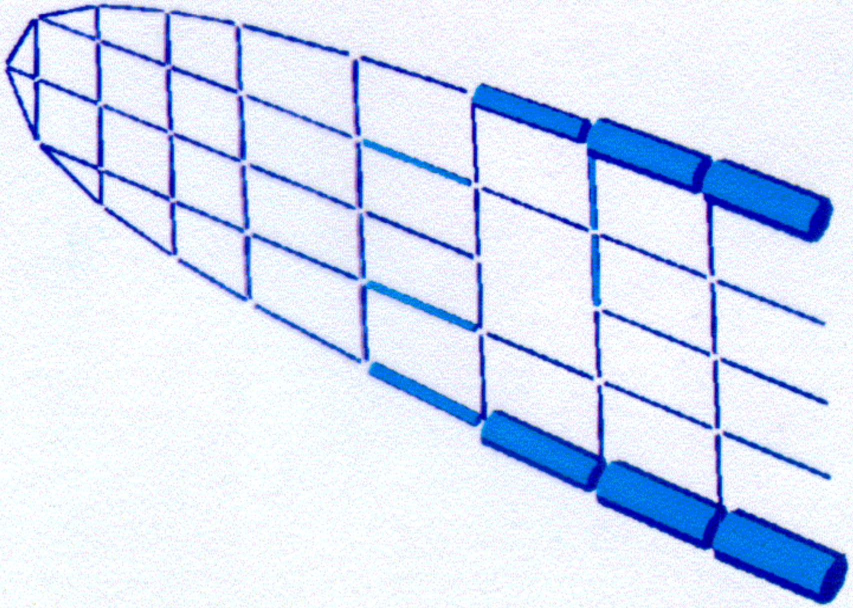


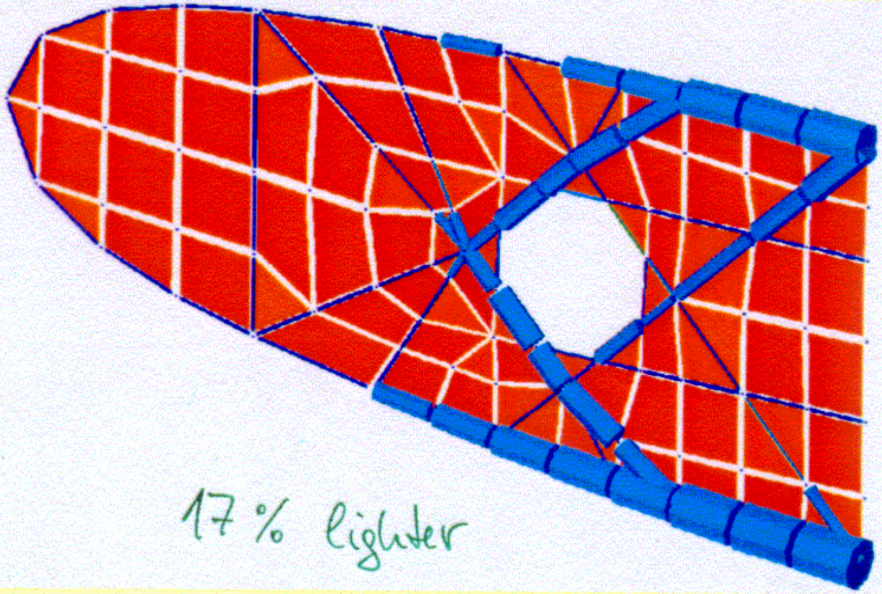
Reference design

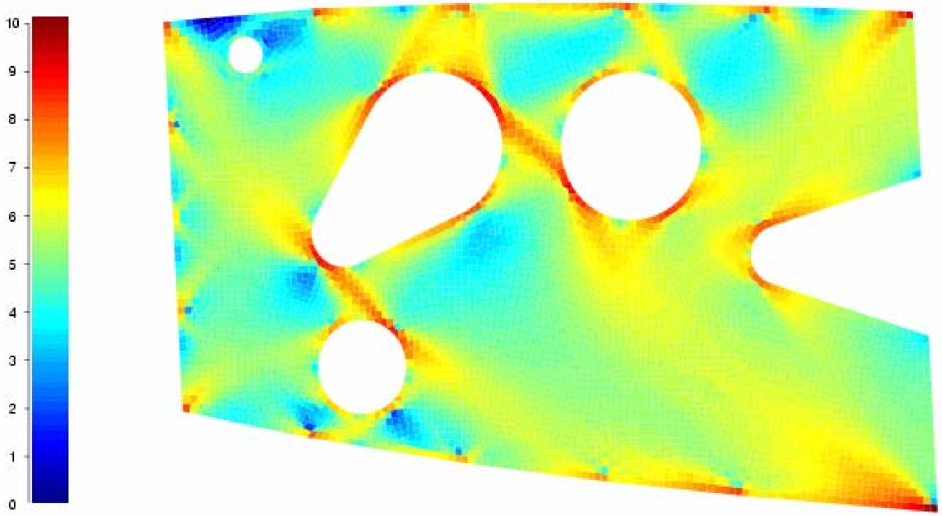
FMO based design











Wing element for Aerobus A380
FMO design by NERML
80 iterations

Implementation, Erlangen University and European Aero Defence and Space Co.

$n = 39,780, \ N = 6,630, \ M = 13,824$

Free Material Optimization: element of aircraft wing
FMO allowed for 17% reduction in element's weight

Based on computational results for maximum stiffness and quite a bit of engineering interpretation a new type of structure was devised for the ribs which gave a weight benefit against traditional and competitive honeycomb/ composite designs (up to 40% !)

A total weight saving of more than 500 kg per wing was obtained by optimizing the ribs in the area shown. These are now — since April 27, 2005 - the first topology optimized parts in flight.

Recovery of signals from noisy outputs

The Estimation Problem

$$y = Hx + w$$

Given y, find an estimator $\hat{x}$, which is as "close" as possible to x.

w random vector

$E(w) = 0, \;\; \mathrm{cov}(w) = C$ positive definite

CLASSICAL METHODS are based on minimizing data error $\|y - Hx\|$.

CLASSICAL APPROACH (Gauss,...)

Closeness measured by (standardized) data error

$$\|C^{-1/2}(y - H\hat{x})\|_2$$

Least Squares Estimator

$$\hat{x}_{LS} = \arg\min_x \|C^{-1/2}(y - Hx)\|_2$$

convex
optimization

SOLUTION (H full column rank)

$$\boxed{\hat{x}_{LS} = (H^T C^{-1} H)^{-1} H^T C^{-1} y}$$

a *linear estimator*

$$\hat{x} = Gy$$

CLASSICAL MODIFICATION (Tikhonov,...)

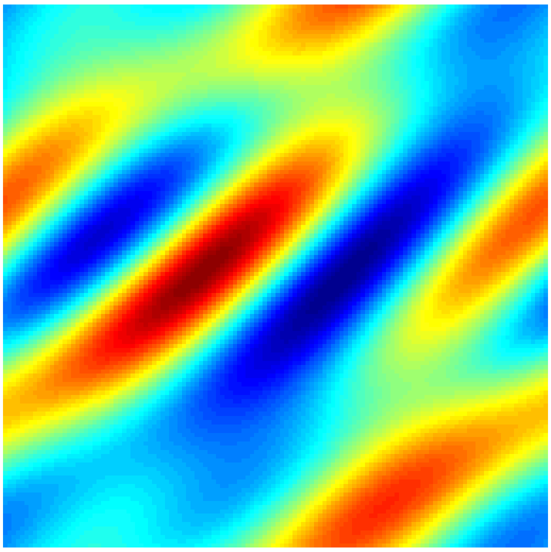
$$\boxed{\hat{x}_T = \arg\min_x \left\{ \|C^{-1/2}(y - Hx)\|^2 + \lambda \|x\|^2 \right\}}$$

still
convex
optimization

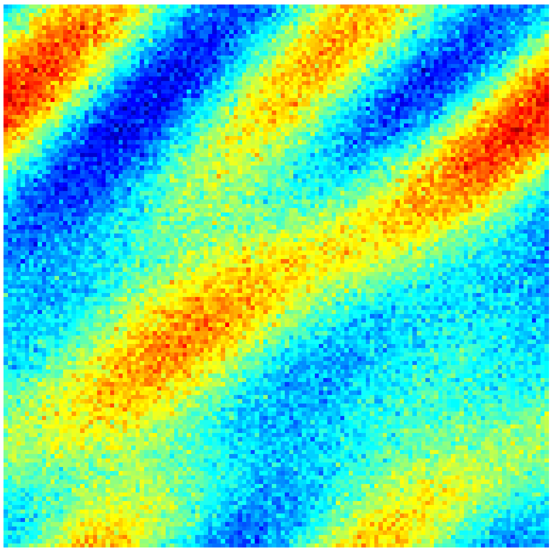
SOLUTION

$$\hat{x}_T = (H^T C^{-1} H + \lambda I)^{-1} H^T C^{-1} y$$

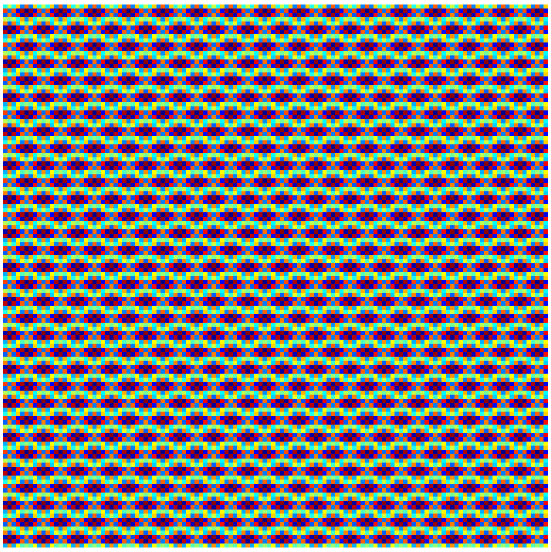
also a linear estimator.



True signal



Observations



LS

MSE estimator

$$\min_{\hat{x}} \; E\|x-\hat{x}\|^2$$

With a *linear estimator* $\hat{x} = Gy$ problem becomes

$$\min_G \left\{ \underbrace{x^T(I-GH)^T(I-GH)x}_{\text{bias}} + \underbrace{Tr(GCG)}_{\text{variance}} \right\}$$

but x unknown!

"Solution": minimal variance unbiased estimator
$GH = I$

Solution: Same as $\hat{x}_{L_S} \cdots$

Our approach: minmax MSE linear estimator:
$\hat{x} = Gy$, where:

$$\min_G \max_{\|x\|_T \le L} \left\{ x^T(I-GH)^T(I-GH)x + Tr(GCG) \right\}$$

$$\min_{G}\{L^2\lambda \max(T^{-1/2}(I-GH)^T(I-GH)T^{-1/2}) + Tr(GCG^T)\} \tag{7}$$

$$\Updownarrow$$

$$\boxed{\begin{array}{ll} \min\limits_{s.t} & L^2\lambda + t \\ & T^{-1/2}(I-GH)^T(I-GH)T^{-1/2} \leq \lambda I \\ & Tr(GCG^T) \leq t \end{array}} \tag{8}$$

Not an SDP ... yet.

Schur's complement:

$$T^{-1/2}(I-GH)^T(I-GH)T^{-1/2} \preceq \lambda I$$

$$\Leftrightarrow \underbrace{\begin{pmatrix} \lambda I & T^{-1/2}(I-GH)^T \\ (I-GH)T^{-1/2} & I \end{pmatrix} \succeq 0}_{LMI}$$

$$Tr(GCG^T) \leq t \Leftrightarrow \underbrace{\begin{pmatrix} t & g^T \\ g & I \end{pmatrix} \succeq 0}_{LMI}$$

where $g = \text{vec}\,(GC^{1/2})$.

Theorem I: Original MinMax MSE problem (1) is equivalent to the SDP problem:

$$\boxed{\begin{array}{l} \min\limits_{s.t.} L^2\lambda + t \\ \begin{pmatrix} \lambda I & T^{-1/2}(I-GH)^T \\ (I-GH)T^{-1/2} & I \end{pmatrix} \succeq 0 \\ \begin{pmatrix} t & g^T \\ g & I \end{pmatrix} \succeq 0 \end{array}}$$

Theorem II: For the special case $T = I$, SDP can be solved explicitly. The optimal MMX MSE estimator is

$$\hat{x}_{\text{mmx}} = \alpha \underbrace{(H^T C^{-1} H)^{-1} H^T C^{-1} y}_{\hat{x}_{LS}}$$

where $$\alpha = \frac{L^2}{L^2 + Tr((H^T C^{-1} H)^{-1})}$$

$\boxed{\text{Proof Structure}}$

(I) Establish the structure of the optimal solution

$$\boxed{G = VDV^T(H^TC^{-1}H)^{-1}H^TC^{-1}}$$

where V is the orthogonal matrix diagonalizing $H^TC^{-1}H$, i.e.,

$$H^TC^{-1}H = V\Sigma V^*$$
$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$$

This is obtained by optimality condition. Using this, we end up with an equivalent problem in variable (matrix) D (Problem B below).

(II) Show that $\exists$ an optimal matrix D which is <u>diagonal</u>.

(III) Find the diagonal elements of D.

The First Part of the Proof

The optimization problem

$$\text{(A)} \qquad \boxed{\begin{array}{l} \min\limits_{s.t.} \quad L^2\lambda + Tr(GCG^T) \\ \begin{pmatrix} \lambda I & (I-GH)^T \\ I-GH & I \end{pmatrix} \succeq 0 \end{array}}$$

Form the Lagrangian:

$$\begin{aligned} L(G,\lambda,U) &= L^2\lambda + Tr(GCG^T) \\ &- \quad Tr\left\{ \begin{pmatrix} U_1 & U_2^T \\ U_2 & U_3 \end{pmatrix} \begin{pmatrix} \lambda I & (I-GH)^T \\ I-GH & I \end{pmatrix} \right\} \\ &= \quad L^2\lambda + Tr(GCG^T) - \lambda Tr(U_1) - 2Tr(U_2(I-GH)) \\ &- \quad Tr(U_3)\,, \end{aligned}$$

$$U \quad := \quad \begin{pmatrix} U_1 & U_2^T \\ U_2 & U_3 \end{pmatrix} \succeq 0$$

Differentiating the Lagrangian with respect to G:

$$\frac{\partial L}{\partial G} = 0 \quad \Leftrightarrow \quad \boxed{G = U_2 H^T C^{-1}}$$

$$\Leftrightarrow \quad GH = U_2 (H^T C^{-1} H)$$

$$\Rightarrow \quad U_2 = (GH)(H^T C^{-1} H)^{-1}$$

change of variables: $D = V^T (GH) V \qquad (V^T V = I)$

$$\Leftrightarrow V D V^T = GH$$

$$\Rightarrow \boxed{U_2 = V D V^T (H^T C^{-1} H)^{-1}}$$

$$\boxed{G = V D V^T (H^T C^{-1} H)^{-1} H^T C^{-1}}$$

In particular, if the orthogonal matrix V is chosen as the matrix which diagonalizes $H^T C^{-1} H$, i.e.

$$(H^T C^{-1} H) = V \underbrace{\operatorname{diag}(\sigma_1, \ldots, \sigma_n)}_{\Sigma} V^T$$

then our problem (A), after substituting G becomes

(B) $$\boxed{\begin{array}{c}\min\limits_{D,\lambda}\ L^2\lambda + Tr(D^T D\Sigma^{-1}) \\ (I-D)^T(I-D) \preccurlyeq \lambda I\end{array}}$$

Second part of the proof ("optimal D can be chosen diagonal").

Let $\mathcal{J}_n$ be the set of 2^n matrices which are $n \times n$, diagonal, with the entries in the diagonal being $+1$ or -1.

Claim If D^* is an optimal solution of (B), then so is

$$JD^*J, \quad \forall\ J \in \mathcal{J}_n$$

Proof

$$\begin{array}{rcl} Tr[(JDJ)^T(JDJ)\Sigma^{-1} & = & Tr(D^T D\Sigma^{-1})] \\ (I-JDJ)^T(I-JDJ) \preceq \lambda I & \Leftrightarrow & (I-D)^T(I-D) \end{array}$$

Conclusion Since (B) is a convex problem $\Rightarrow$ its optimal solution set is convex, so if D^* is an optimal solution, so is

$$\frac{1}{2^n} \sum_{J\in\mathcal{J}_n} (JD^*J)$$

$$D = \begin{bmatrix} a, & b \\ c, & d \end{bmatrix}$$

$$\mathcal{J}_2 = \left\{ \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{J_1}, \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{J_2}, \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}}_{J_3} \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}}_{J_4} \right.$$

$$J_1 D J_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad J_2 D J_2 = \begin{pmatrix} a & -b \\ -c & d \end{pmatrix}$$

$$J_3 D J_3 = \begin{pmatrix} a & -b \\ -c & d \end{pmatrix} \qquad J_4 D J_4 = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\frac{1}{4} \sum_{i=1}^{4} J_i D J_i = \frac{1}{4} \begin{bmatrix} 4a & 0 \\ 0 & 4d \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$$

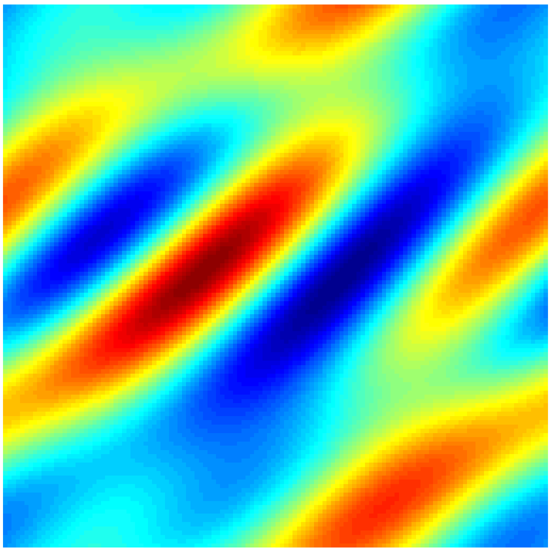
General result

$$\frac{1}{2^n} \sum_{J \in \mathcal{J}_n} JDJ = \text{diag } D$$

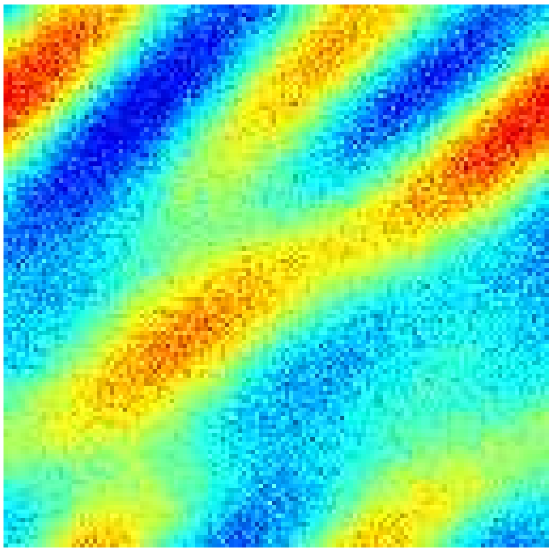
Part 3 of the proof with $D = \text{diag}(d_1, \ldots, d_n)$ problem (A) $\Leftrightarrow$ (B) reduces to

$$\boxed{\begin{array}{ll} \min\limits_{d_i, \lambda} & L^2 \lambda + \sum (d_i^2 / \sigma_i) \\ \text{s.t.} & (1 - d_i)^2 \leq \lambda, \quad \forall\, i \end{array}}$$

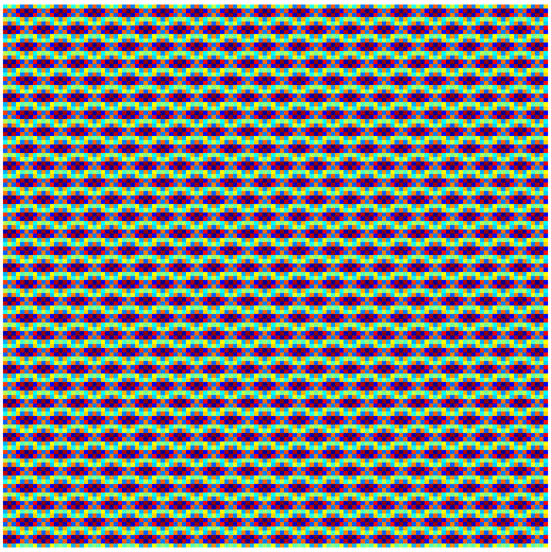
This problem can be solved analytically, which gives the final result claimed in Theorem II.



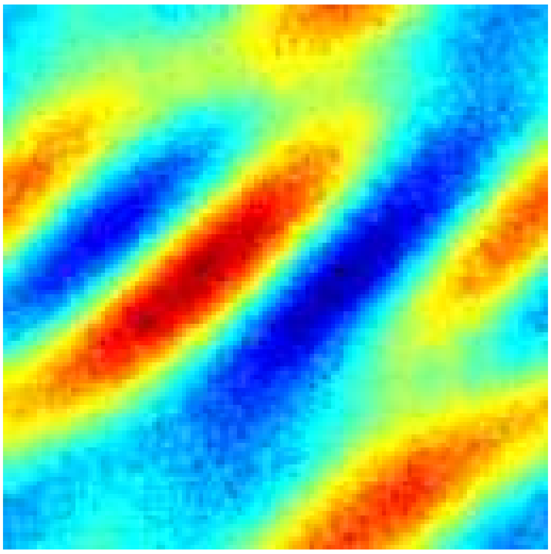
True signal



Observations



LS



Minmax Use

Support Vector Machine (SVM) Learning

A powerful tool for solving large-scale classification problems.

Given to sets of point I^+ and I^- we are looking for a hyperplane $H = \{x | w^T x + b = 0\}$ which separates the point in I^+ from those in I^-.

Minimum distance of x from $H = \left| \frac{w^T x + b}{\|w\|} \right|$; hence we want to solve:

$$\max_{w,b} \left\{ \min_{i \in I^+} \left| \frac{w^T x_i + b}{\|w\|} \right| + \min_{i \in I^-} \left| \frac{w^T x_i + b}{\|w\|} \right| \right\} \tag{1}$$

Let $y_i = 1$ if $i \in I^+$, $y_i = -1$ if $i \in I^-$

problem (1) can be reduced to

$$\max_{w,b} \left\{ \frac{2}{\|w\|} \middle| y_i \left(w^T x_i + b \right) \geq 1, \ \forall \, i \right\}$$

or, finally to the so-called Linear SVM problem

$$\min \left\{ \frac{1}{2} \|w\|^2 \middle| y_i \left(w^T x_i + b \right) \geq 1, \ \forall \, i \right\} \tag{2}$$

When the two sets cannot be separated by H, we relax the constraints by penalizing points which are on the wrong side of H. The final formulation is

$$\text{(P)} \qquad \min_{w,b,\xi\geq 0} \left\{ \frac{\|w\|^2}{2} + c\Sigma\xi_i \,\middle|\, y_i\left(w^T x_i + b\right) \geq 1 - \epsilon_i,\ \forall\, i \right\}$$

The dual problem of (P) is

$$\text{(D)} \qquad \max_{\alpha} \left\{ \Sigma\alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i x_j \,\middle|\, \begin{array}{c} 0 \leq \alpha_i \leq c \\ \alpha^T y = 0 \end{array} \right\}$$

When x_i is mapped by a function ϕ to a (high-dimensional) “feature space”

$x_i^T x_j$ is replaced by $\phi(x_i)^T\phi(x_j) = k(x_i, x_j)$

$0 \preccurlyeq K$ is the **kernel function** associated with ϕ.

The dual problem (D) becomes

(NONLIN-D-SVM)

$$\boxed{\begin{array}{l} \displaystyle\max_{\alpha\in S_n} \left\{ \Sigma\alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j K(x_i, x_j) \right\} \\ S_n = \left\{ \alpha \,\middle|\, \begin{array}{l} 0 \leq \alpha_i \leq c \quad \forall\, i \\ \alpha^T y = 0 \end{array} \right\} \end{array}}$$

Multiple Kernel Learning (MKL)

In multi-modal learning applications, like object categorization, where multiple feature representors are present (e.g. in flower categorization: shape, color, texture, etc.), we should employ multiple kernels to achieve good visual discrimination as well as within-class variation.

Variable Sparsity Kernel Learning (VSKL) for Binary Classification

Training data set:

$$D = \{x_i, , y_i,\ i = 1, \ldots, m \mid x_i \in \mathcal{X},\, y_i \in \{-1, 1\}\}$$

The kernels are divided into n groups (feature representors), j-th group has n_j kernels.

$p_{jk}(\cdot)$ = feature space mapping corresponding to $K_{jk}(\cdot,\cdot)$, the k-th kernel of the j-th component.

Consider the problem of learning a linear discriminant function

$$f(x) = \sum_{j=1}^{n} \sum_{k=1}^{n_j} w_{jk} \phi_{jk}(x) + b\,.$$

VSKL formulation aims at classifiers that will generate a *non-sparse* combination of kernels representing different groups, while having a *sparse* selection of nonredundant kernels with given feature representors.

The optimization problem (NLP-DUAL VSKL) modeling the above requirements:

(P-VSKL)

$$\boxed{\begin{aligned} &\min_{w_{jk}, b, \xi_i \geq 0} \frac{1}{2} \left[\sum_j \left(\sum_k \|w_{jk}\|_2 \right)^{2q} \right]^{1/q} + c \sum_i \xi_i \\ &\text{s.t.} \\ &\qquad y_i \left(\sum_j \sum_k w_{jk} \phi_{jk}(x_i) - b \right) \geq 1 - \xi_i, \ \forall i \end{aligned}} \tag{3}$$

$$9 \geq 1$$

DUAL OF P-VSKL

Let $a_j = \left(\sum_k \|w_{jk}\|_2\right)^2$, using

Lemma 1

$$\begin{cases} (\Sigma a_j^q)^{1/q} = \max\left\{ \Sigma\gamma_j a_j \mid \underbrace{\Sigma\gamma_j^p = 1,\ \gamma_j \geq 0}_{\Delta_p} \right\} \\ p = \dfrac{q}{q-1} \end{cases}$$

the objective function in (3)

$$\boxed{\begin{aligned} &\min_{w_{jk}, b, \xi_i \geq 0} \frac{1}{2}\left[\sum_j \left(\sum_k \|w_{jk}\|_2\right)^{2q}\right]^{1/q} + c\sum_i \xi_i \\ &s.t. \\ &\qquad y_i\left(\sum_j \sum_k w_{jk}\phi_{jk}(x_i) - b\right) \geq 1 - \xi_i,\ \forall i \end{aligned}} \tag{3}$$

becomes

$$\frac{1}{2}\max_{\gamma \in \Delta_p}\left\{ \sum_{j=1}^{n} \gamma_j \left(\sum_{k=1}^{n_j} \|w_{jk}\|_2\right)^2 + c\sum_i \xi_i \right\}.$$

Further, using

Lemma 2

$$\left(\sum_k \|w_{jk}\|_2\right)^2 = \min_{\lambda_{jk}} \left\{ \sum_k \frac{\|w_{jk}\|^2}{\lambda_{jk}} \,\middle|\, \underbrace{\sum_k \lambda_{jk} = 1,\ \lambda_{jk} \geq 0}_{\Delta^j} \right\}$$

Problem (3)*, in variables* w_{jk}*,* λ_{jk}*,* γ_j, b *becomes*

$$\boxed{\begin{array}{l} \min\limits_{w_{jk},b,\xi_i} \max\limits_{\gamma \in \Delta_p} \min\limits_{\lambda_j \in \Delta^j} \underbrace{\left\{ \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n_j} \gamma_j \frac{\|w_{jk}\|^2}{\lambda_{jk}} + c \sum_{i=1}^{m} \xi_i \right\}}_{f(w,\lambda,\gamma,\xi)} \\ s.t. \\ (4a) \quad y_i \left(\sum_j \sum_k w_{jk}^T \phi_{jk}(x_i) - b \right) \geq 1 - \xi_i \ \forall\, i = 1, \ldots, m \\ (4b) \quad \xi_i \geq 0 \end{array}} \tag{4}$$

f is concave (linear) in γ on Δ_p = compact convex set and convex in λ on $\otimes \Delta^j$ = compact convex set.

By **Sion-Kabutani: minmax theorem**

$$\max_{\gamma \in \Delta_p} \min_{\lambda \in \otimes \Delta^j} f(w, \lambda, \gamma, \xi) = \min_{\lambda} \max_{\gamma} f(w, \lambda, \gamma, \xi)$$

so, a (partial) dual of (4) is

$$\min_{\lambda \in \otimes \Delta^j} \min_{w,b,\xi} \max_{\gamma \in \Delta_p} f(w, \lambda, \gamma, \xi)\,, \quad \text{s.t. } (4a), (4b) \tag{5}$$

Now, for fixed λ, f is convex in w, b, ξ on the feasible set (4a) (4b), which is a closed convex set. Moreover, f is concave (linear) in γ on the compact convex set Δ_p; hence by **Rockafellar minmax theorem**, (5) is converted to

$$\min_{\lambda \in \otimes \Delta^j} \max_{\gamma \in \Delta_k} \left\{ \min_{w,b,\xi} f(w, \lambda, \gamma, \xi) \quad \text{s.t. } (4a), (4b) \right\} \tag{6}$$

We now dualize the problem in curly brackets $\{\ \}$ to obtain

(D-VSKL)

$$\min_{\lambda \in \otimes \Delta^j} \max_{\gamma \in S_m, \gamma \in \Delta_p} \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \alpha^T \left(\sum_j \sum_k \frac{\lambda_{jk} Q_{jk}}{\gamma_j} \right) \alpha \right\}$$

$$S_m = \{\alpha \in \mathbb{R}^m \mid \Sigma \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq c,\ i = 1, \ldots, m\}$$

$$\Delta_p = \left\{\gamma \in \mathbb{R}^n \mid \Sigma \gamma_j^p = 1,\ \gamma \geq 0\right\} \ p = \frac{q}{q-1}$$

$$\Delta^j = \left\{ \lambda_j \in \mathbb{R}^{n_j} \mid \sum_{k=1}^{n_j} \lambda_{jk} = 1,\ \lambda_j \geq 0 \right\} \ j = 1, \ldots, n$$

$$Q_{jk} \in \mathbb{R}^{m \times m} (Q_{jk})_{ih} = y_i y_j K_{jk}(x_i, x_h)$$

$$K_{jk}(x_i, x_n) = \phi_j(x_i) \phi_k(x_n)$$

Algorithm for solving (D-VSKL)

Essentially we solve

$$\min_{\lambda_j \in \Delta^j} \left\{ \begin{array}{l} G(\lambda_1, \lambda_2, \ldots, \lambda_n) = \\ = \max_{\alpha \in S_n \gamma \in \Delta_p} \underbrace{\left\{ \Sigma \alpha_i - \alpha^T \sum_j \left(\frac{\sum_k \lambda_{jk} Q_{jk}}{\gamma_j} \right) \alpha \right\}}_{f_\lambda(\alpha,\lambda)} \end{array} \right\} \tag{7}$$

G is a convex (nonsmooth) function.

Proposition 1 *If* $\exists\ \mu > 0,\ 0 < \tau < 1$ *s.t.* $\forall\, j, k$*:*

$$\lambda_{mm}(Q_{jk}) \geq \tau\mu\,, \quad \lambda_{\max}(Q_{jk}) \leq \mu$$

then, for all $p \geq 1$*,* $G(\cdot)$ *is Lipschitz continuous in the* ℓ_1 *norm.*

Functions with the above properties are amenable for solution by the Mirror Descent (MD) algorithm. For a feasible set which is a product of simplices $\otimes\Delta^j$, MD based on a proximal term given by relative-entropy is theoretically optimal, and practically efficient: at iteration t:

$$\min_{1 \leq s \leq t} G(\lambda^s) - G^* \leq 0(1)\sqrt{\log n}\frac{1}{\sqrt{t}}$$

Notice the very weak dependence on the design-dimension n.

The MD algorithm needs an *oracle*, a procedure that at every iteration given a current point λ, will generate the **function value** $G(\lambda)$ and a **subgradient of** G at λ.

The Oracle for problem (7) is obtained by the solution of the problem

$$\max_{\alpha\in S_n,\gamma\in\Delta_p} f_\lambda(\alpha,\lambda) \tag{8}$$

Two methods for solving problem (8):

Method I Conic Quadratic Programming

$$\max_{\alpha\in S_n,\gamma\in\Delta_p}\left\{ f_\lambda(\gamma,\alpha) = \Sigma\alpha_i - \frac{1}{2}\alpha^T\left(\sum_j \frac{\sum_k \lambda_{jk}Q_{jk}}{\gamma_j}\right)\alpha\right\} = \tag{9}$$

$$\max_{\alpha\in S_m,\gamma\in\Delta_p}\left\{\Sigma\alpha_i - \Sigma v_j \mid 2\gamma_j v_j \ge \alpha^T\left(\sum_k \lambda_{jk}Q_{jk}\right)\alpha,\ \forall j\right\} \tag{10}$$

Using the identity $2\gamma_j v_j = \frac{1}{2}(\gamma_j+v_j)^2 - \frac{1}{2}(\gamma_j - v_j)^2$ the j-th constraint in (9) becomes

$$\alpha^T\left(\sum_k \lambda_{jk}Q_{jk}\right)\alpha + \frac{1}{2}(\gamma_j - v_j)^2 \le \frac{1}{2}(\gamma_j+v_j)^2$$

which is a CQ constraint. The other constraint and the objective function in (10) are linear, so (10) is indeed a CQ problem that can be solved by IP methods (MOSEK, CPLEX,...)

Method II Block Coordinate Descent (BCD)

This method is based on the fact that in problem (9):

$$\max_{\alpha \in S_n, \gamma \in \Delta_p} \left\{ f_\lambda(\gamma, \alpha) = \Sigma \alpha_i - \frac{1}{2} \alpha^T \left(\sum_j \frac{\sum_k \lambda_{jk} Q_{jk}}{\gamma_j} \right) \alpha \right\} = \tag{9}$$

For fixed γ, the problem in α is the classical SVM problem, but with matrix

$$Q = \sum_j \frac{\sum_k \lambda_{jk} \lambda_{jk} Q_{jk}}{\gamma_j} .$$

For SVM extremely efficient algorithms are available.

For fixed α, the problem in γ has an explicit solution ($q > 1$, $p < \infty$)

$$\gamma_i = \frac{D_i^{\frac{1}{p+1}}}{\left(\sum_{j=1}^n D_j^{\frac{p}{p+1}} \right)^{1/p}} \quad i = 1, \ldots, n$$

where

$$D_j = \sum_{k=1}^{n_j} \lambda_{jk} \alpha^T Q_{jk} \alpha$$

If $q = 1 \quad p = \infty$, then $\gamma_i = 1$.

The above fact suggests an algorithm that sequentially solves the α-problem with fixed γ, and then compute γ from the above expression (A BCD method).

Generally, a BCD method may not converge, but it can be proved, that for our particular problem (8),

$$\max_{\alpha \in S_n, \gamma \in \Delta_p} f_\lambda(\alpha, \lambda) \tag{8}$$

the BCD method does converge.

CONCLUSION

For very large-scale ML problems, like VSKL, method II combined with the MD algorithm offer a very efficient algorithm (the main step in MD is projection on $\otimes\Delta^p$ w.r.t. an entropy distance, which has an explicit solution as well!)

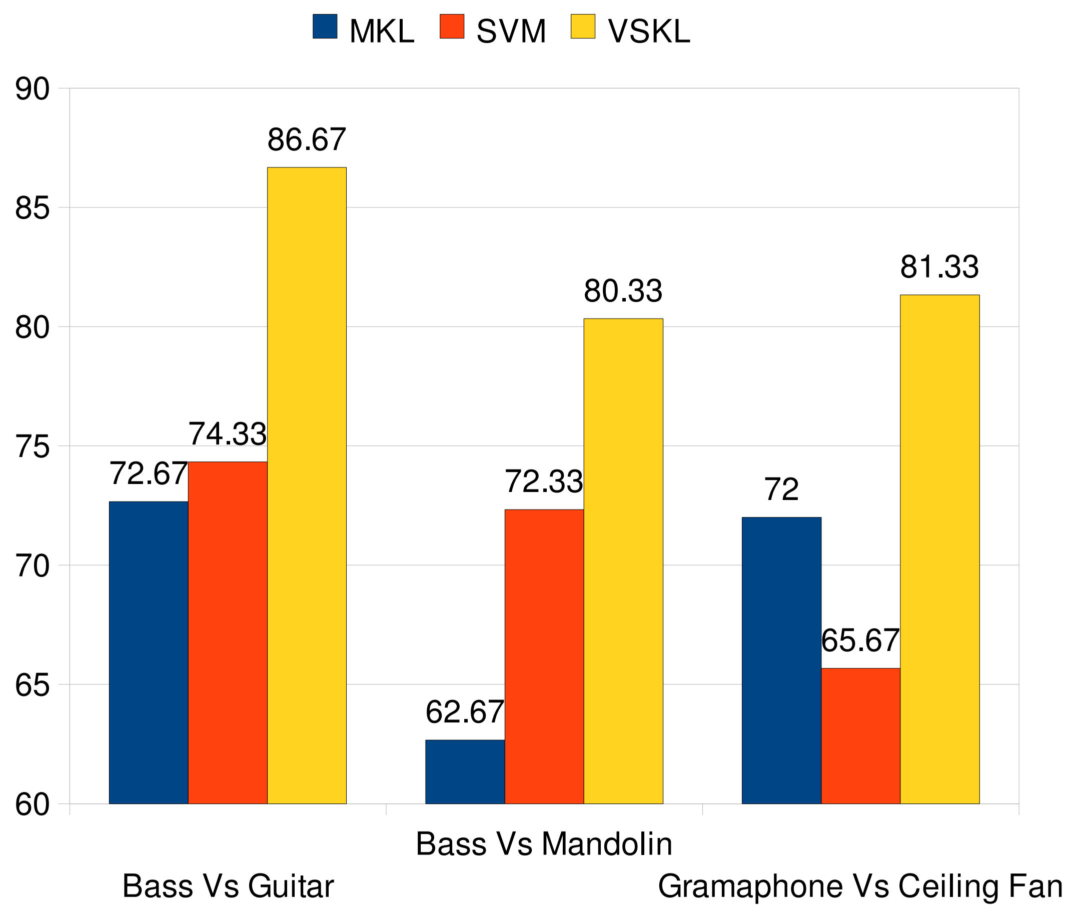
Numerical Experiment

Dataset: Caltech-101 collection of 101 categories of objects like faces, watches, ants, etc. Number of images per category 40–800. Each image is represented using five feature descriptors (SIFTs)

Specific example: 3 binary classification problems

groups: $n = 5$, # of kernels in group j, $n_j = 9 \; \forall \, j$

Comparing average testset accuracy for 3 methods MKL, SVM and our VSKL.


MKL
SVM
VSKL
90
85
80
75
70
65
60
86.67
74.33
72.67
80.33
72.33
62.67
81.33
72
65.67
Bass Vs Guitar
Bass Vs Mandolin
Gramaphone Vs Ceiling Fan